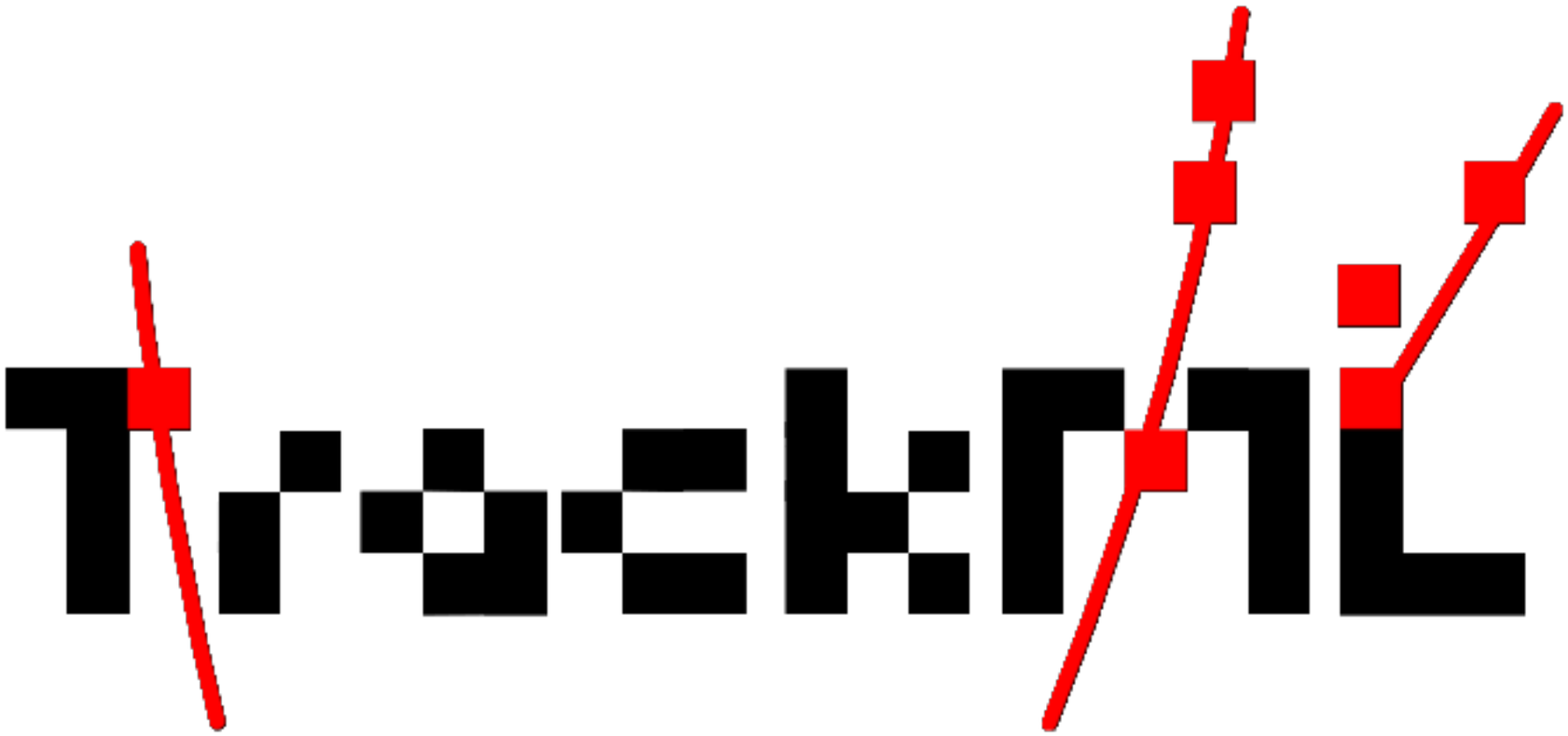


UNIVERSITÉ DE GENÈVE

a tracking machine learning challenge

Tobias Golling, University of Geneva

ICHEP2018, Seoul, July 5 2018

TrackML Organisation Team

Paolo Calafiura, Steven Farrell, Heather Gray (LBNL-Berkeley),

Jean-Roch Vlimant (CalTech),

Edward Moyse (University of Massachusetts),

Cécile Germain (LAL/LRI U Paris Saclay),

Isabelle Guyon (ChaLearn, U Paris Saclay),

David Rousseau, Yetkin Yilnaz (LAL Orsay U Paris Saclay),

Vincenzo Innocente, Andreas Salzburger (CERN),

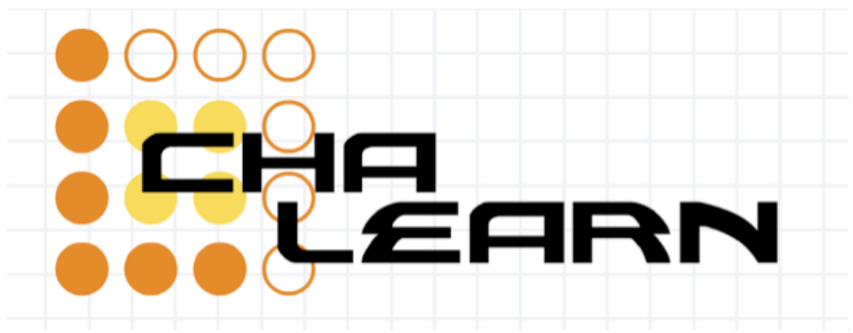
Tobias Golling, Moritz Kiehn, Sabrina Amrouche (U Geneva),

Vava Gligorov (LPNHE-Paris),

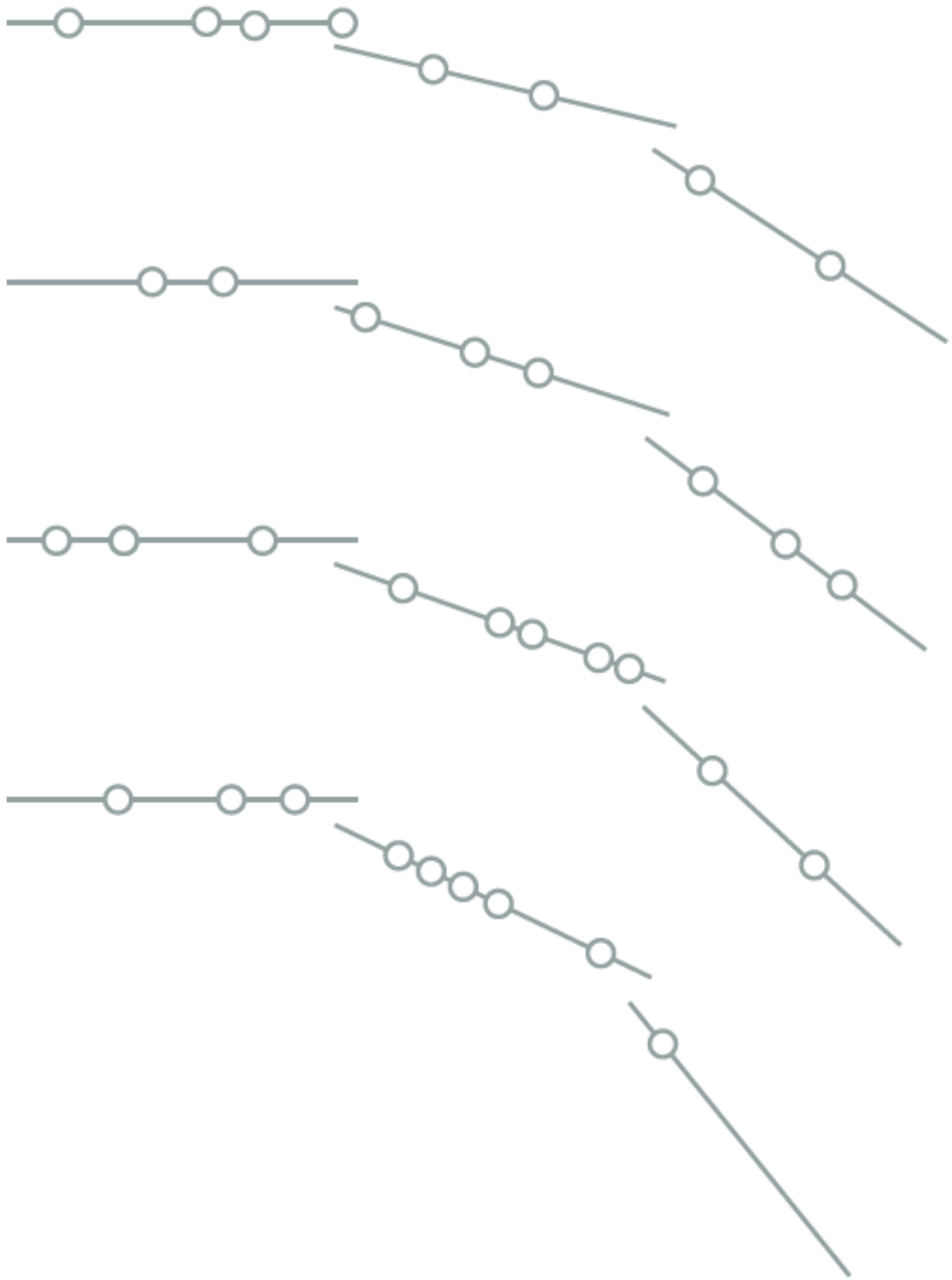
Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex),

Ilija Vukotic (U of Chicago)

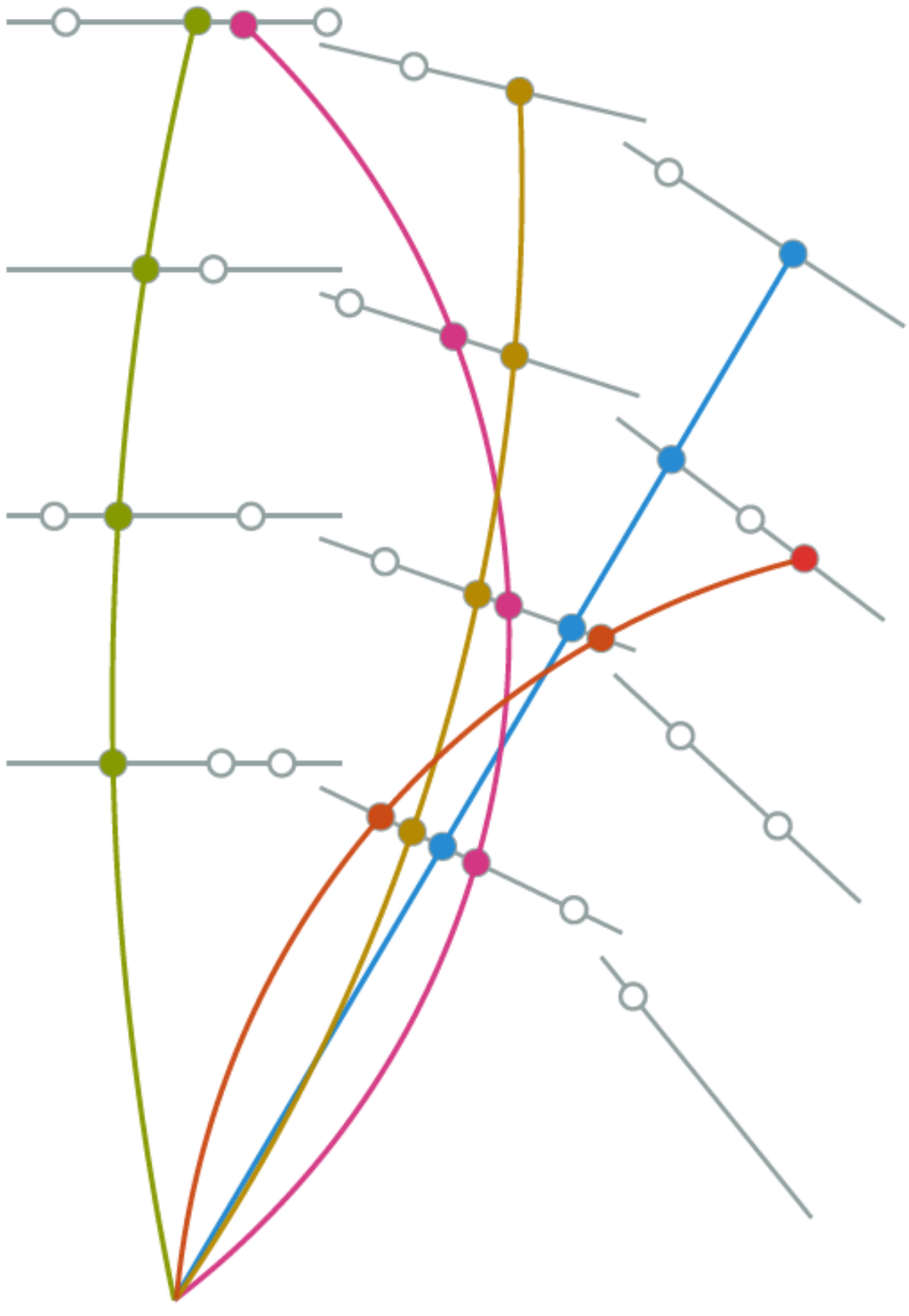
TrackML Sponsors



The challenge: connecting the dots

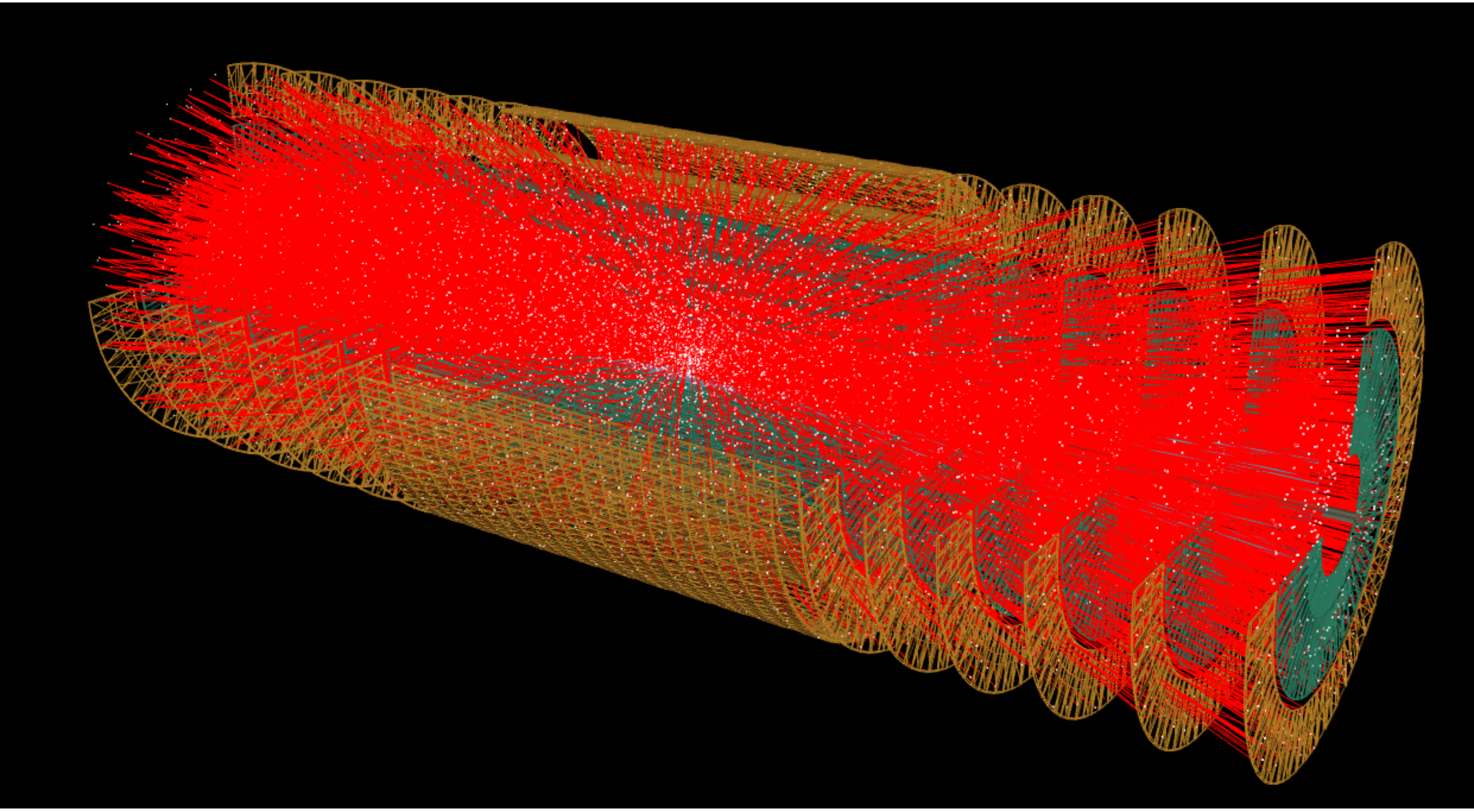


The challenge: connecting the dots…



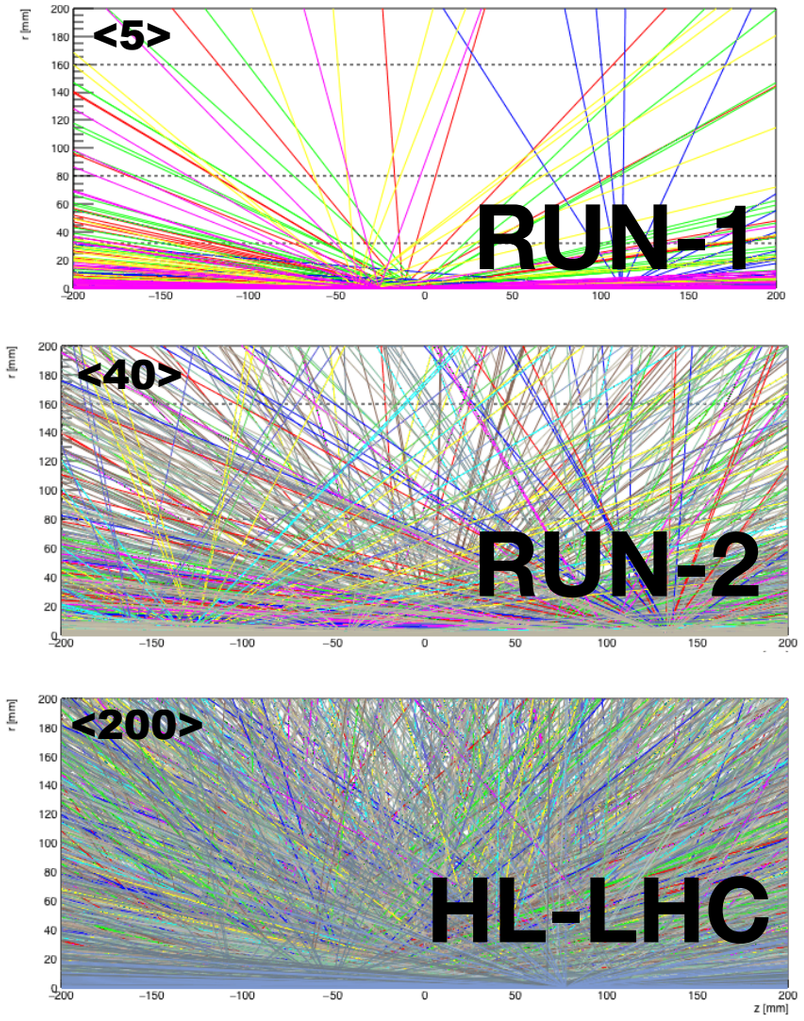
…but how?

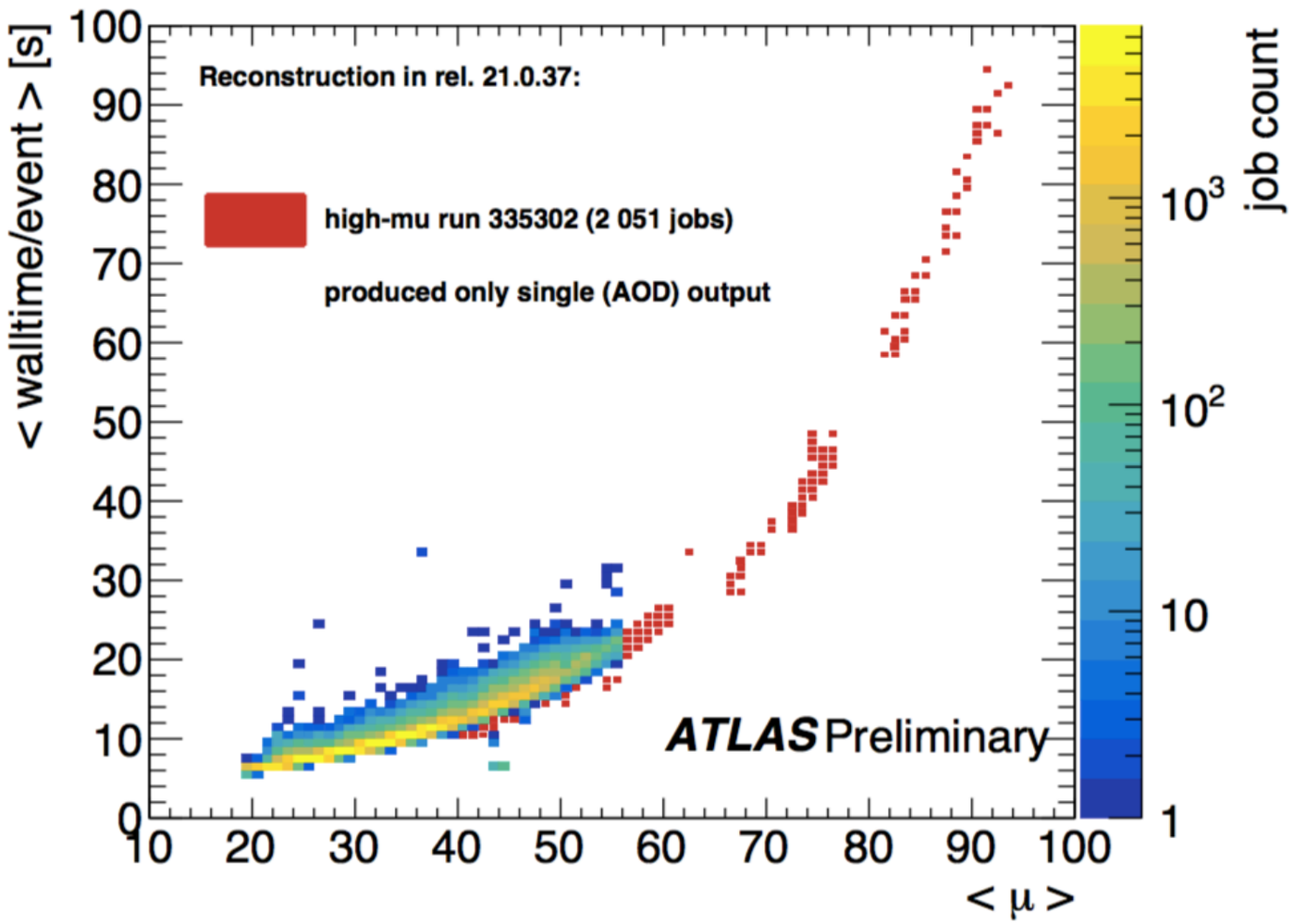
High density of charged particles in solenoid B-field



The HL-LHC tracking challenge:
clustering 100k hits into 10k tracks per event

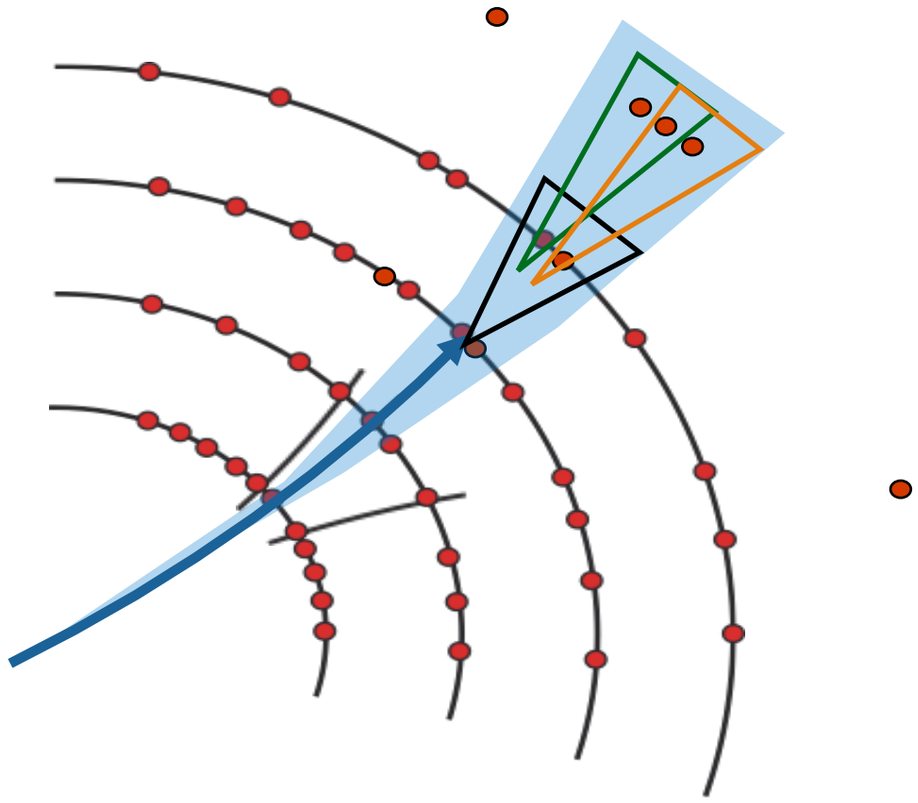
Tracking at LHC, HL-LHC and FCC-hh

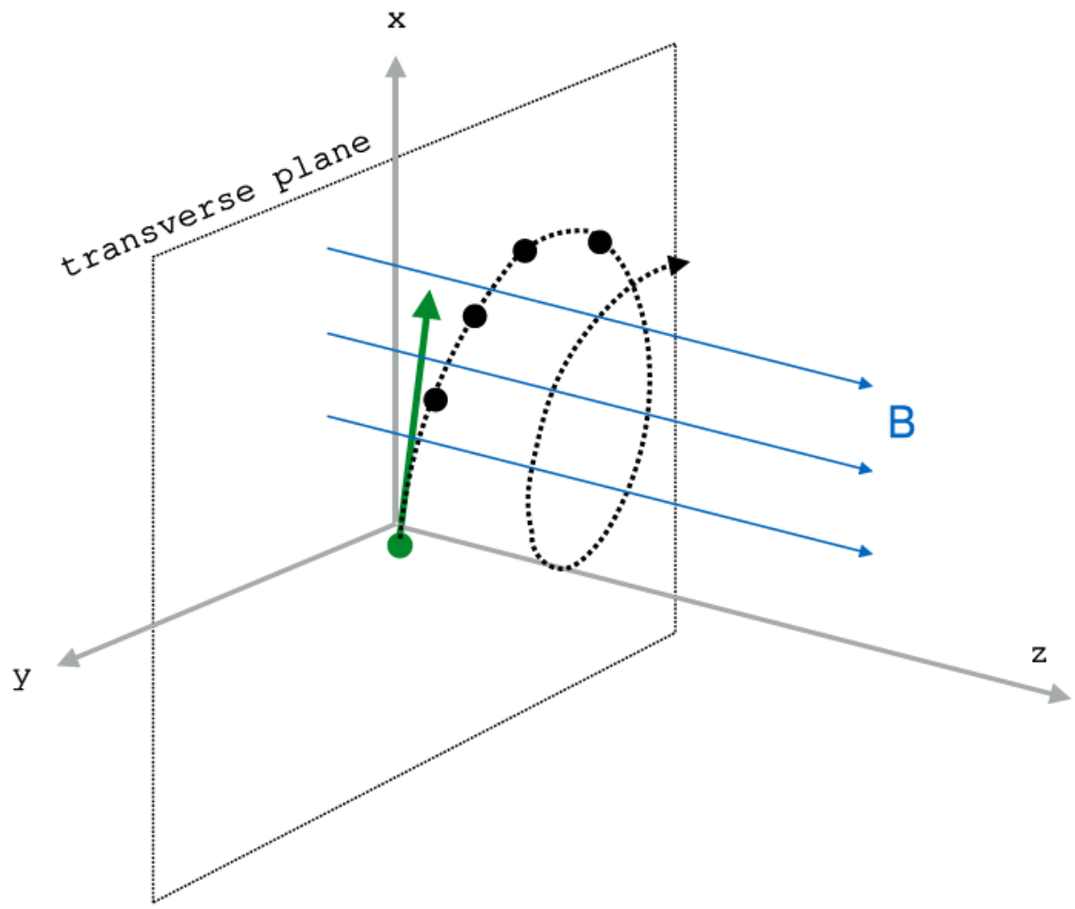




Pattern recognition is main CPU driver due to its combinatorial behavior

State of the art tracking in HEP





Current algorithm:

- Pattern: connect 3D points into tracks
- Essentially combinatorial approach
- Tracks are (not perfect) helices pointing (approximately) to the origin

Machine Learning for Tracking

Can we use this for particle tracking ?

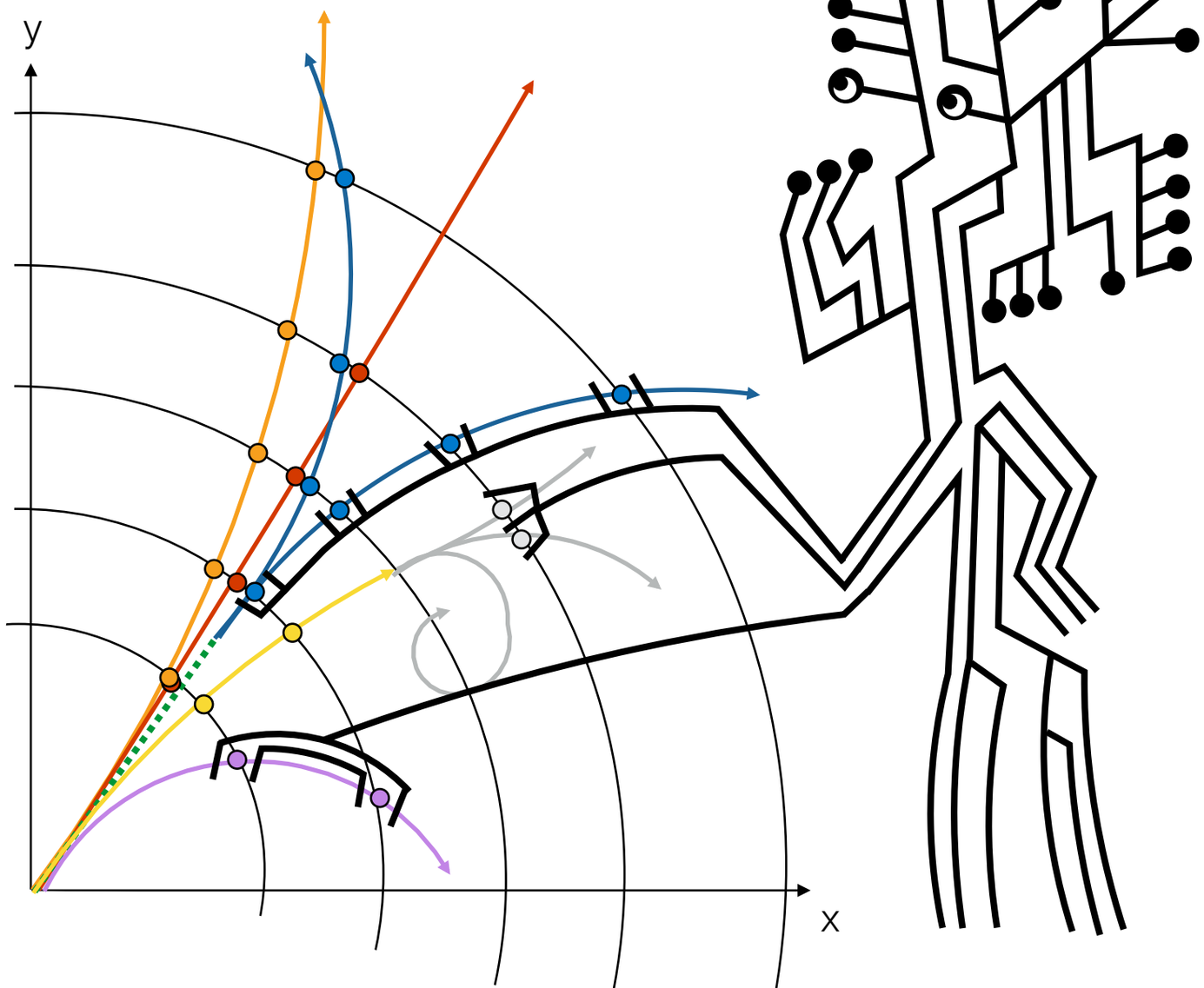
Feature Finding

Unsupervised learning

Metric learning

Classification

Prediction



Particular challenges of this tracking challenge

Unusual dataset

Variable number of inputs, unknown number of outputs

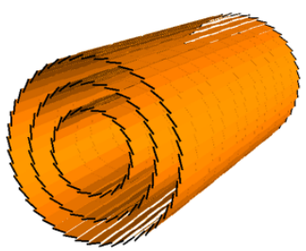
Rich field for applications

Different algorithms and approaches, no obvious optimal algorithm

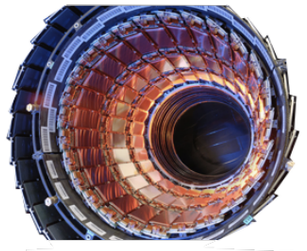
Verifiable

Detailed simulation available

What matters in tracking



ideal

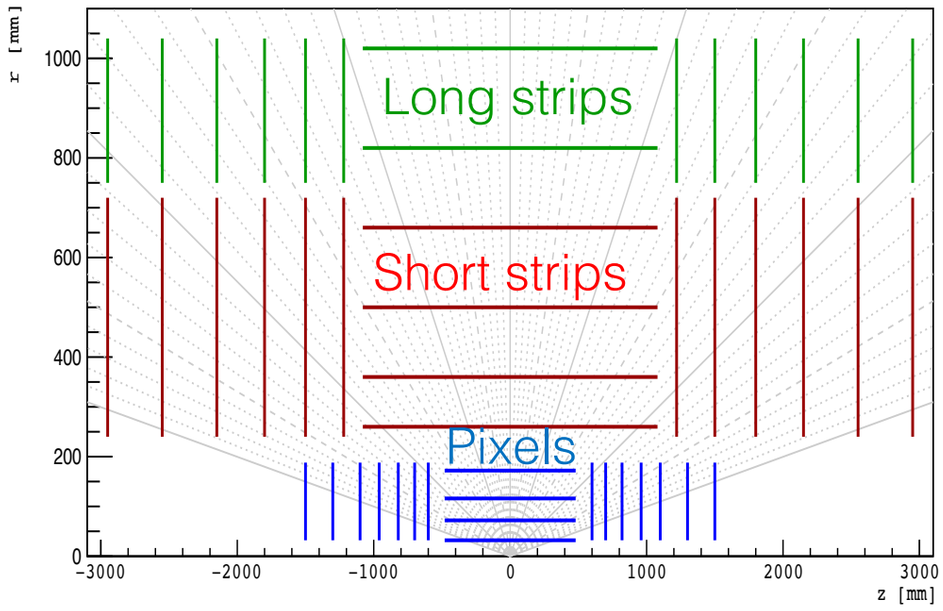


real

An all-silicon tracking detector

- Non-homogenous solenoid field
- Complex material distribution & geometry
- Wide variety of physics events
- **Tracking efficiency**
- **Minimum fake rate**
- **Impact parameter resolution**
- **p_T resolution**

A typical ATLAS-CMS HL-LHC tracker



Simplify but don't oversimplify!

- No hit merging/splitting/sharing
- Approximate material by uniform cylinders/disks
- One event type (ttbar)
- **One single metric**

Event simulation

- ACTS (https://gitlab.cern.ch/acts/acts-core) simulation engine

- Pythia ttbar events
- Overlaid with Poisson(200) Pythia minimum bias (~10.000 tracks/event)

- Luminous region with gaussian width 5.5 mm in z and 15μm transverse
- 15% random hits
- Trajectories are deterministic, except for multiple scattering, energy loss and hadronic interaction

Data format: the dots and how to connect them

One file with list of 3D points

event...-hits.csv

```
hit_id,      x,      y,      z, volume_id, layer_id, module_id
...
      2, -812.8, -631.4,  411.5, ...
      3,  649.4, -785.7, -309.1, ...
...
```

Ground truth for training: one file with point to particle association

event...-truth.csv

```
hit_id,         particle_id,    tx,    ty     tz,   tpx,   tpy,   tpz
...
      2,   4513289073590272, -813., -630.,  378., -0.74, -0.19,  0.27
      3, 117094208786923520,  648., -786., -352.,  0.12, -0.63, -0.21
```

Objective

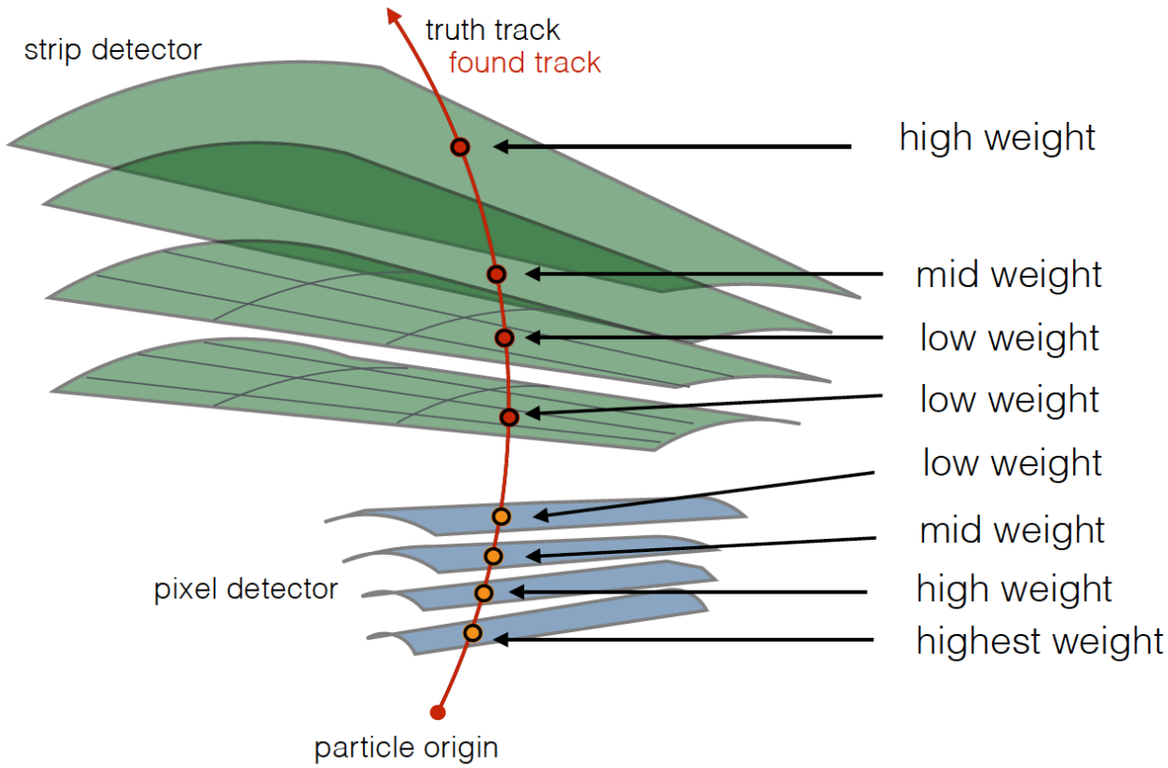
- Participants to upload tracks they have found
- A track is a list of 3D points
- (do not consider estimation of particle parameter)
- Score: fraction of points correctly grouped together
- Evaluation on test sample with per-mille precision on 100 events

submission.csv

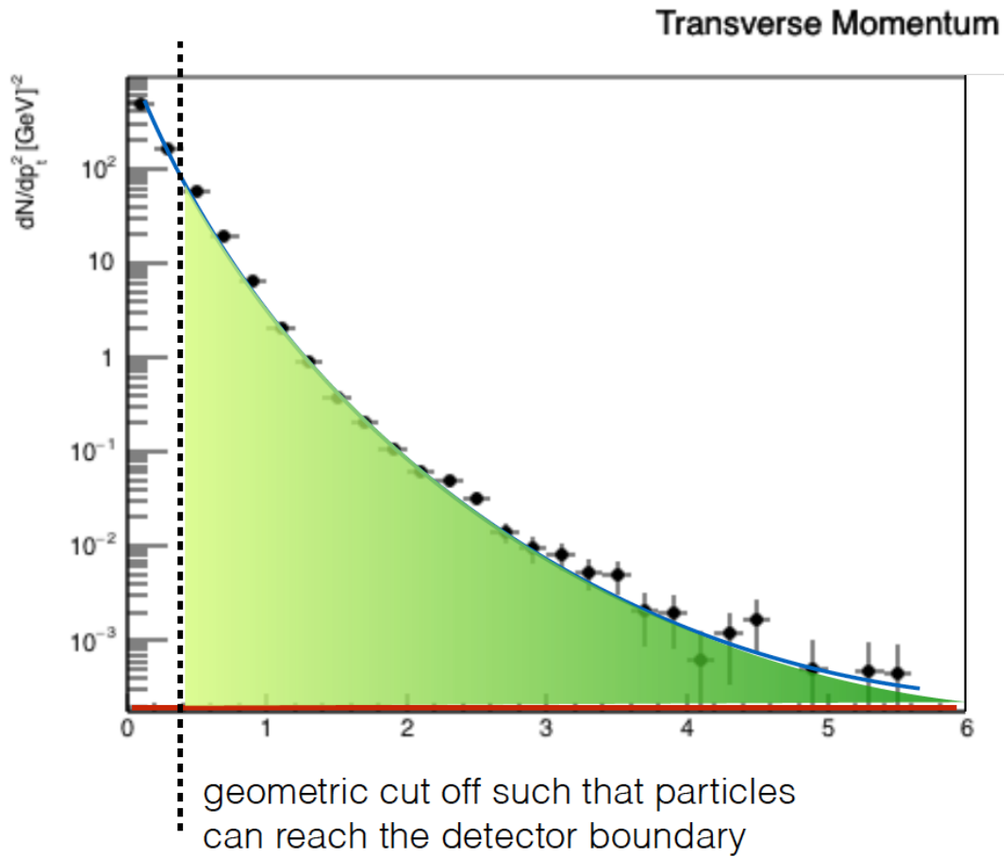
```
1  event_id, hit_id, track_id
2  ...
3          51,      2,        42
4          51,      3,        23
5  ...
```

Weighting scheme

$weight_{hit}$



$weight_{pT}$



Total $weight = weight_{hit}$ x $weight_{pT}$

Track scoring

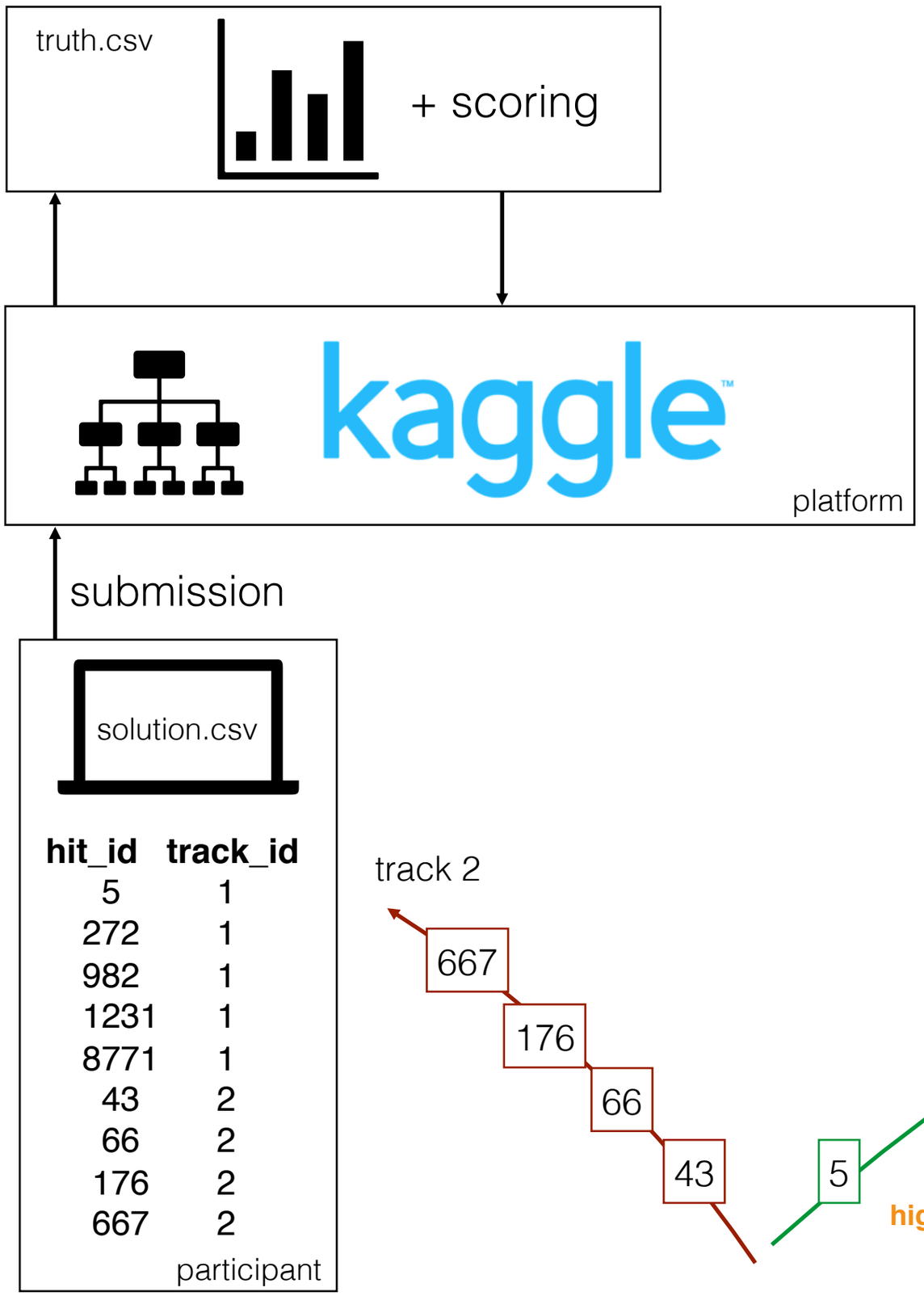
Figure of merit: $\sum$ (weights of reco track) / $\sum$ (weights of true track)

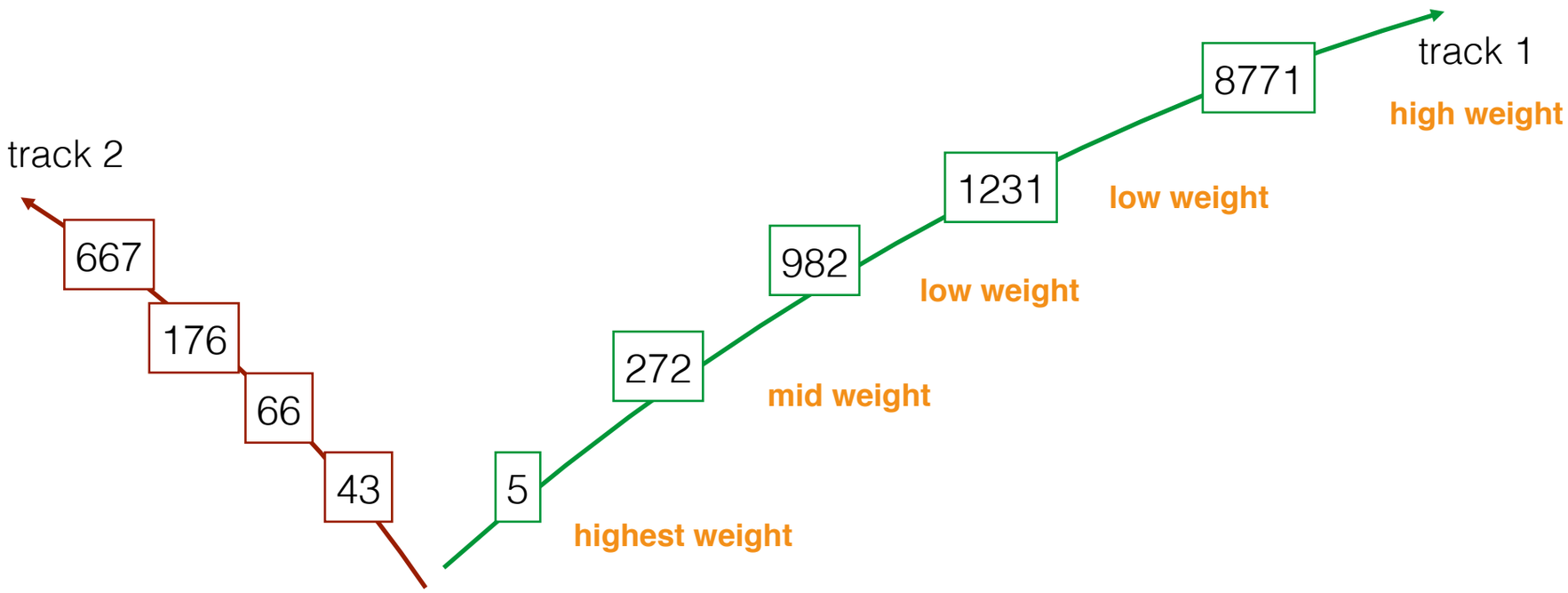
Ideal score is 1

Final score averaged over 100 events: statistical precision ~0.1%

Require >50% of hits from same true particle
Require >50% of hits from this true particle in this reco track

Submission


truth.csv
+ scoring
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


track 2
667
176
66
43
5
272
982
1231
8771
track 1
mid weight
low weight
low weight
high weight
highest weight

The challenge in 2 phases

phase 1: **accuracy**

who solves it best ?

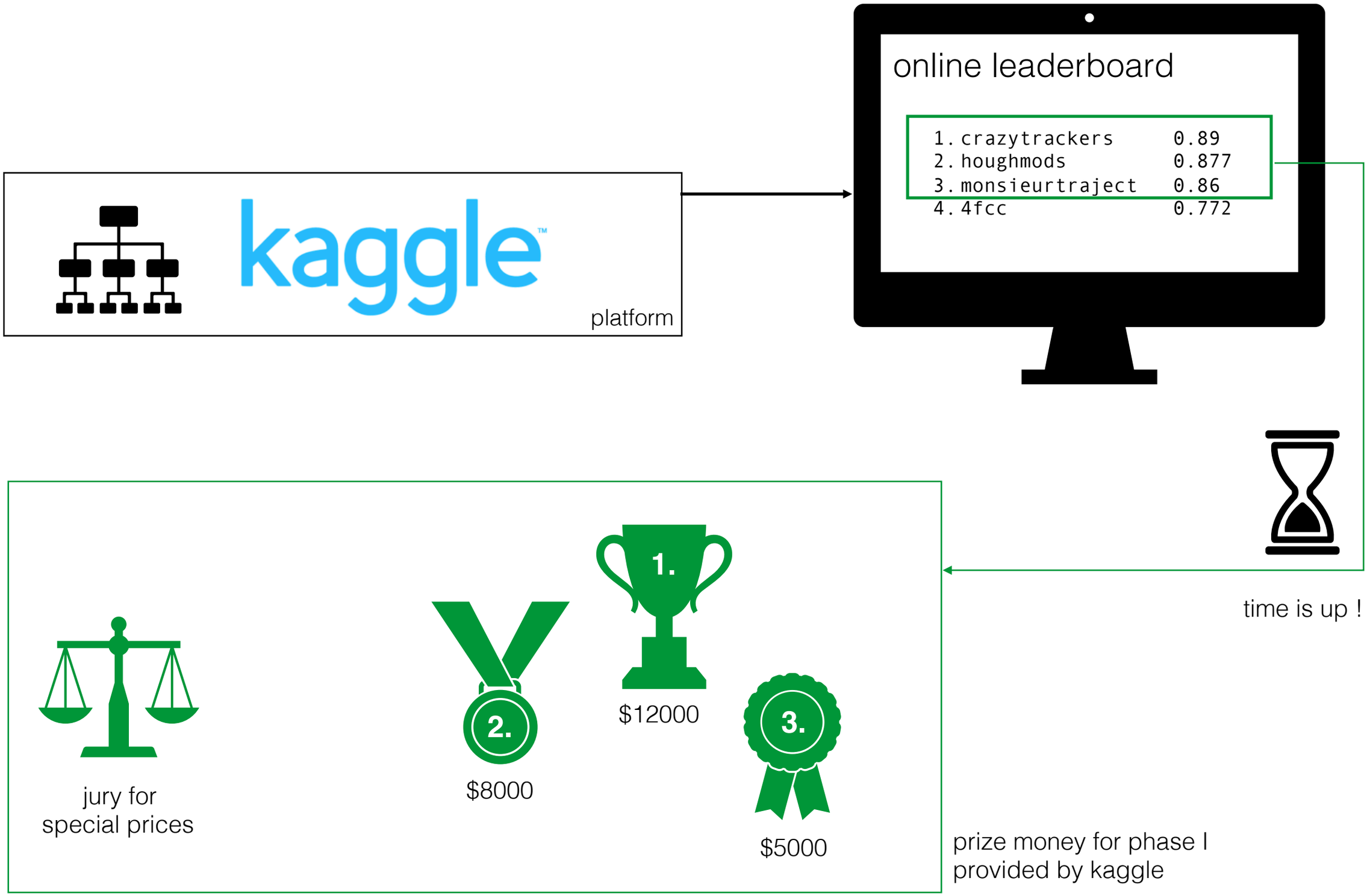
- started on May 1 until Aug 13 2018

phase 2: **throughput**

who solves it best and is fast ?

- phase foreseen to start **end of summer 2018**
- significant computing needed for evaluation
- prototype for controlled throughput being currently developed

Winning


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

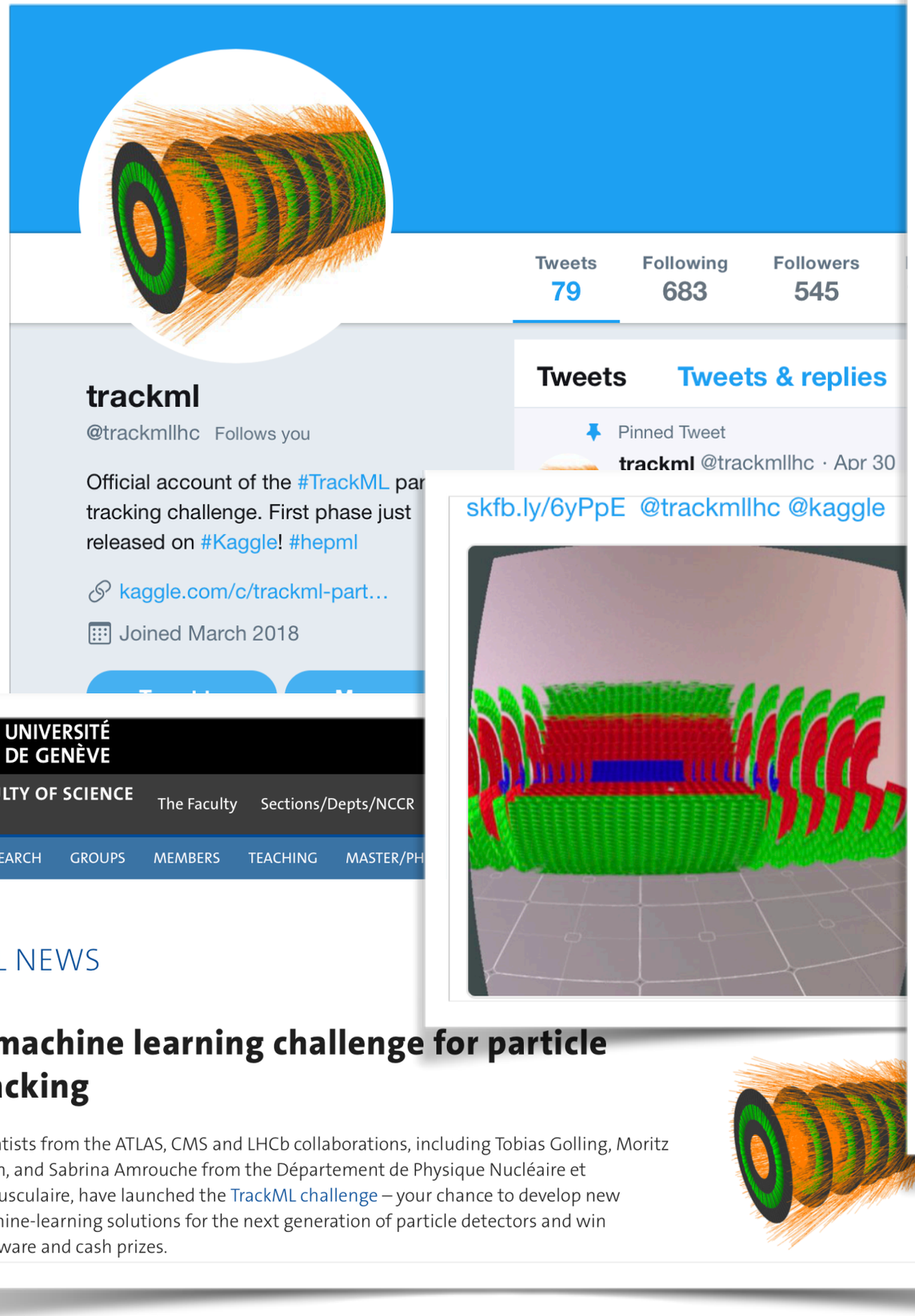
Jury & Prizes

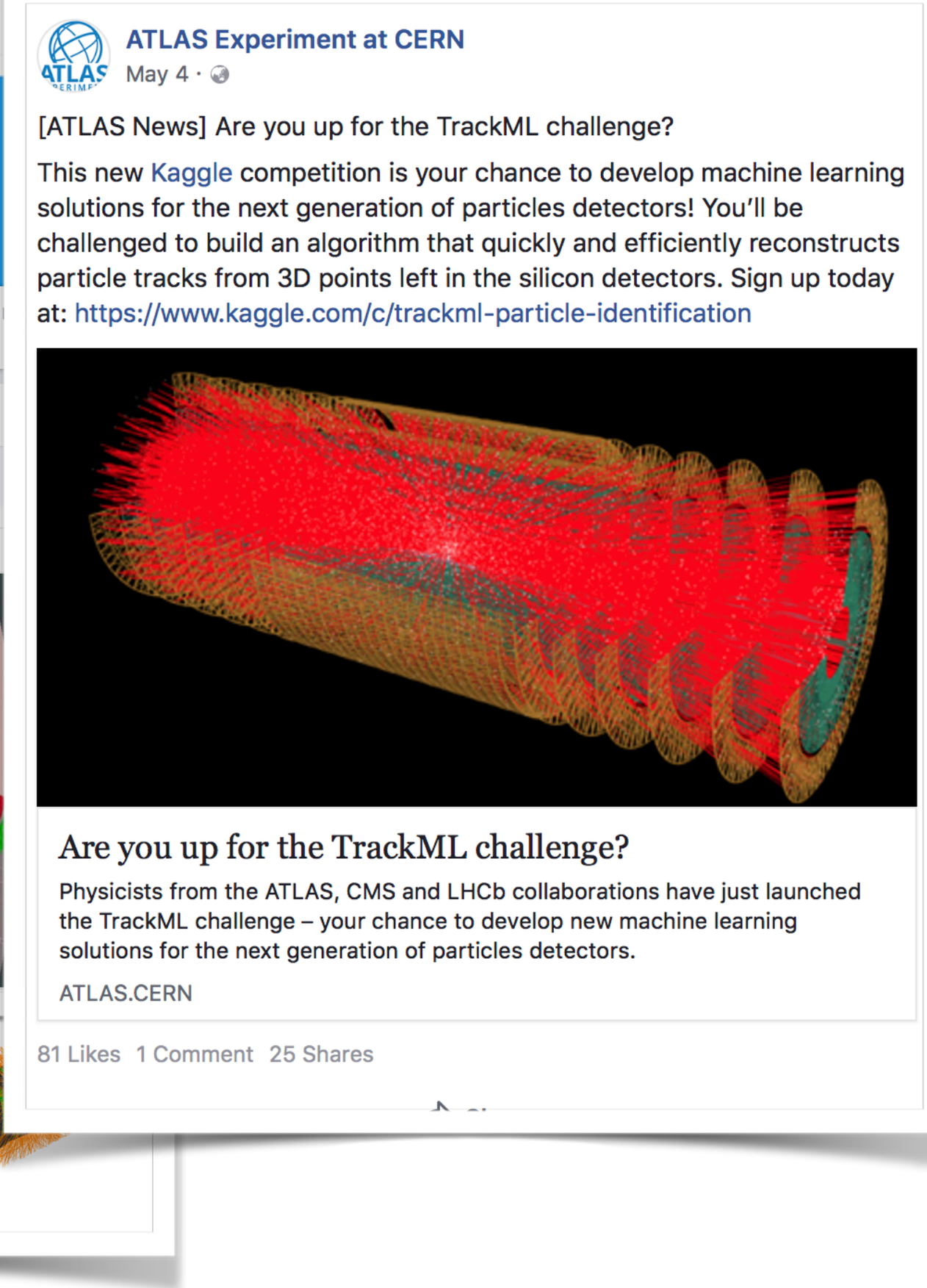
- International Advisory Committee (IAC) jury for prize:
  - ML: Danilo Rezende (DeepMind), Marc Schoenauer (INRIA), Svyatoslav Voloshynovsky (UniGenève), Alison Lowndes (Nvidia)
  - HEP: Frank Gaede (ILC), Markus Elsing (ATLAS), Maurizio Pierini (CMS)

- *HEP meets ML* Prizes (**releasing documented code mandatory**):
  - 2 NVIDIA V100 GPUs (1 for each phase)
  - Invitations to NIPS (Dec 2018)
  - CERN workshop (Spring 2019)

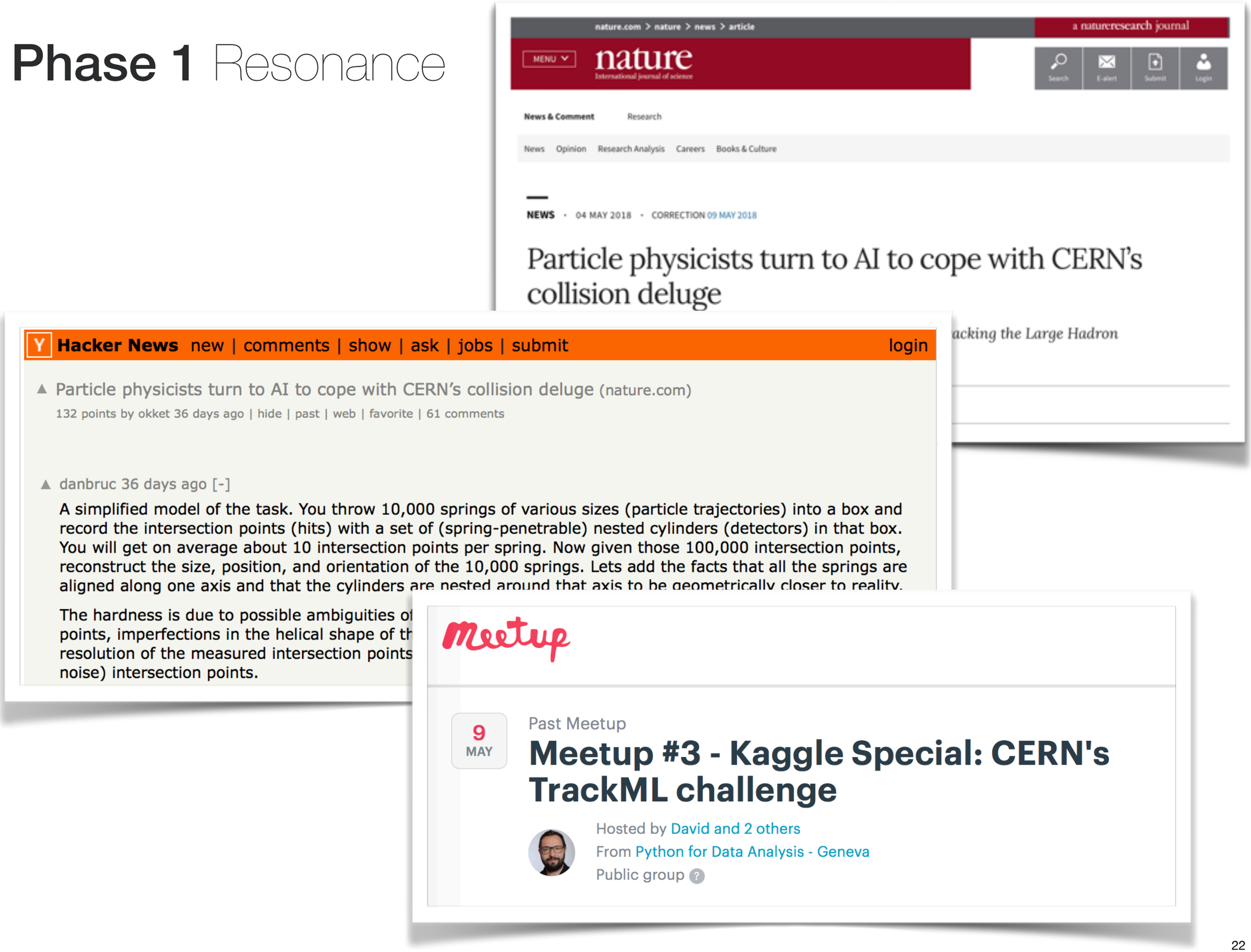
- Criteria:
  - Originality
  - Anticipated usefulness for HEP

Phase 1 Channels

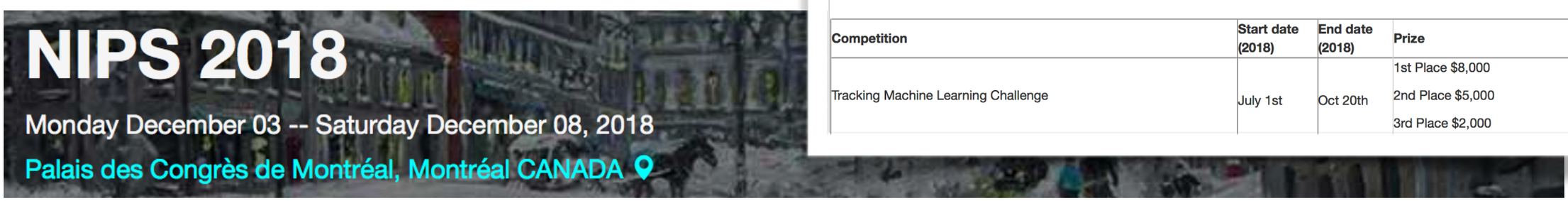




Phase 1 Resonance



Phase 1 Conferences



NIPS 2018 Competition Track

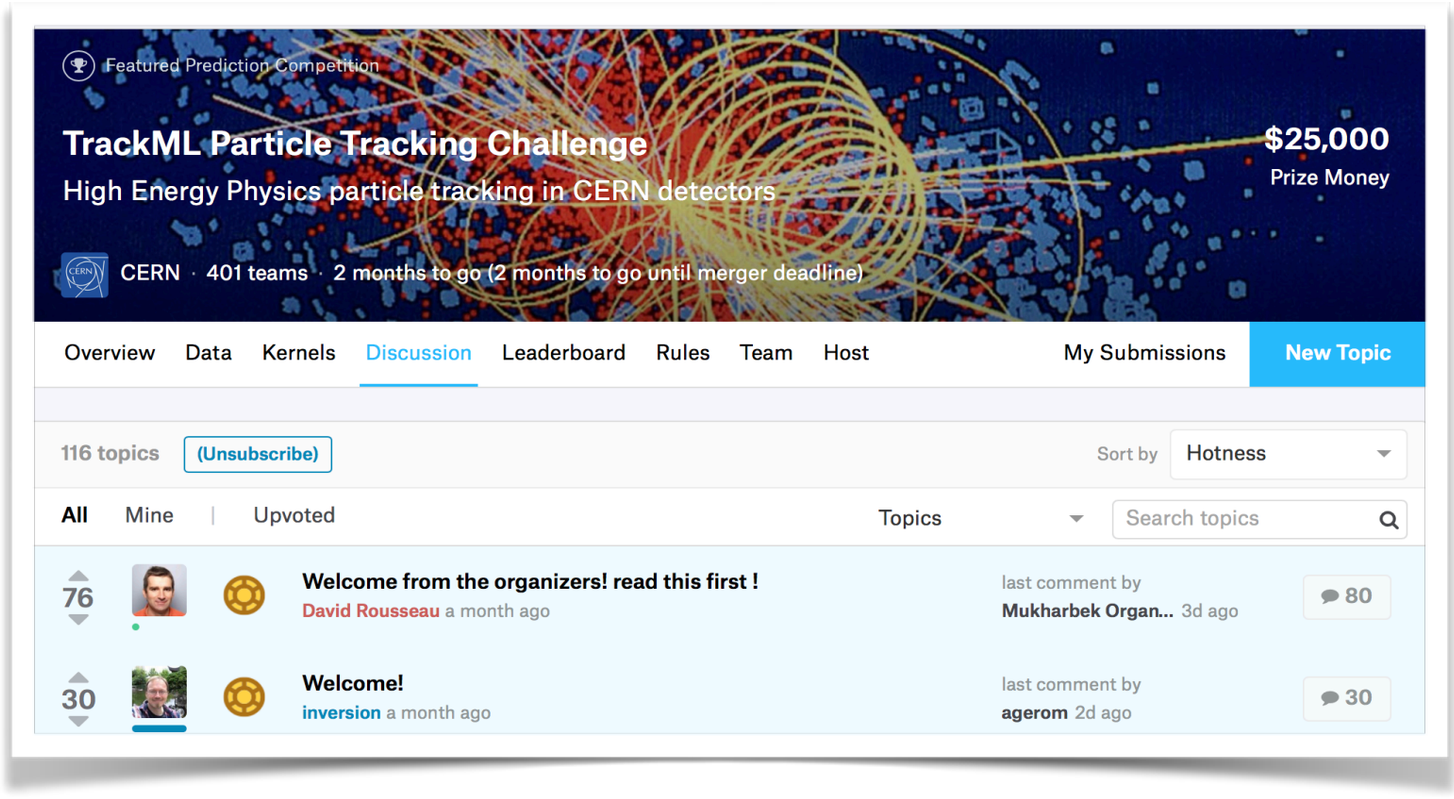
This is the second NIPS edition on "NIPS Competitions". We received 21 competition proposals related to data-driven and live competitions on different aspects of NIPS. Proposals were reviewed by several high qualified researchers and experts in challenges organization. **Eight top-scored competitions were accepted** to be run and present their results during the NIPS 2018 Competition track days. Evaluation was based on the quality of data, problem interest and impact, promoting the design of new models, and a proper schedule and managing procedure. Below, you can find the eight accepted competitions. Please visit each competition webpage to read more about the competition, its schedule, and how to participate. Each competition has its own schedule defined by its organizers. The results of the competitions, including organizers and top ranked participants talks will be presented during the 2 Competition track days at NIPS 2018. Organizers and participants will be invited to submit their contribution as a book chapter to the upcoming NIPS 2018 Competition book, within Springer Series in Challenges in Machine Learning.

Competition	Start date (2018)	End date (2018)	Prize
Tracking Machine Learning Challenge	July 1st	Oct 20th	1st Place $8,000 2nd Place $5,000 3rd Place $2,000

Phase 1 Lessons learned

Contestants find it very complicated – lots of discussion

- We did our best to try to simplify the problem and provide appropriate description
- Data volume (O(100 GB)) for some people difficult to handle



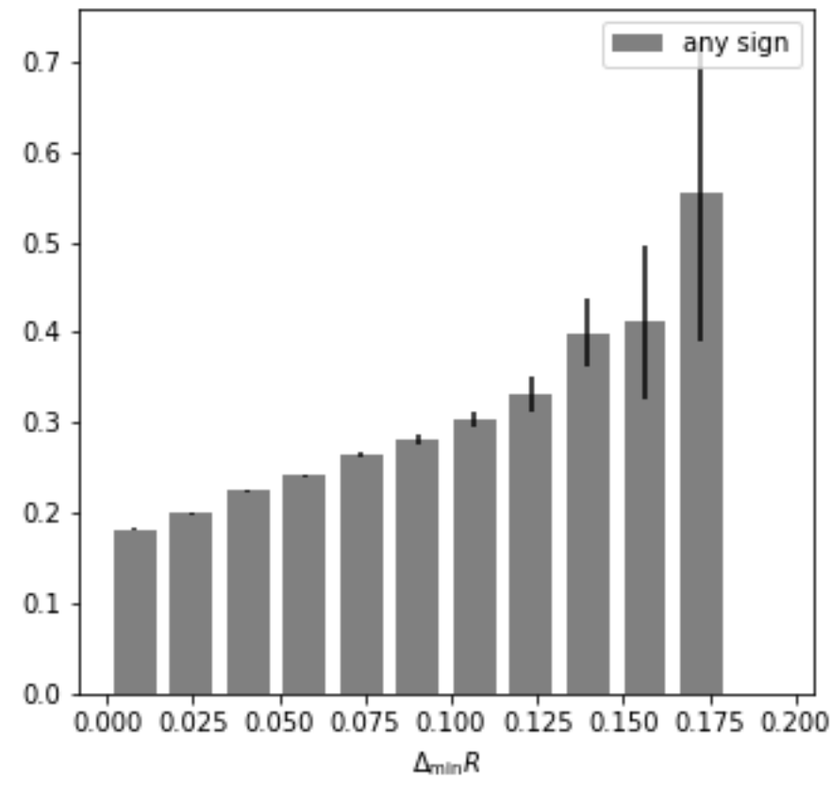
We had a few rather harmless bugs/feature in the dataset

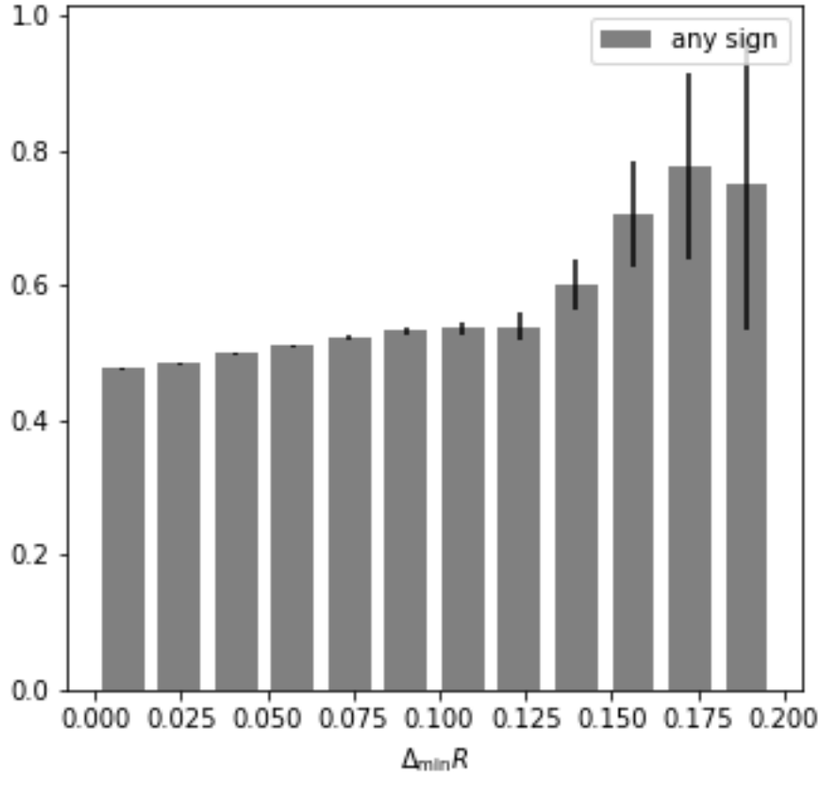
- Beam spot σ_z=5.5 mm instead of 5.5 cm
- Sensor thickness twice too big (but radiation length is OK)
- We will likely fix this before we re-simulate the dataset for archiving …

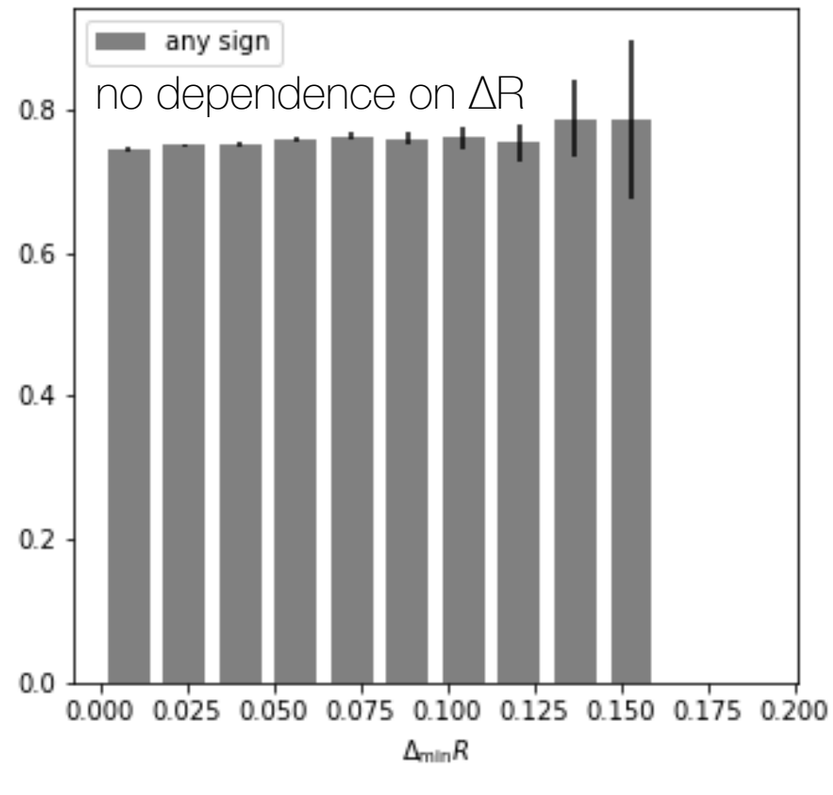
Phase 1 starting from reference notebooks: DBSCAN

- We provided simple reference solutions, in particular DBSCAN(*)
- Few lines of preprocessing and calling sklearn DBSCAN clusterer yielding the **non-trivial score of 20% (left)**
- This was **quickly brought to 50% (middle)** and even higher (hints on the forum)
- Best submission now **70-80% (right)**

efficiency vs. ΔR to next track





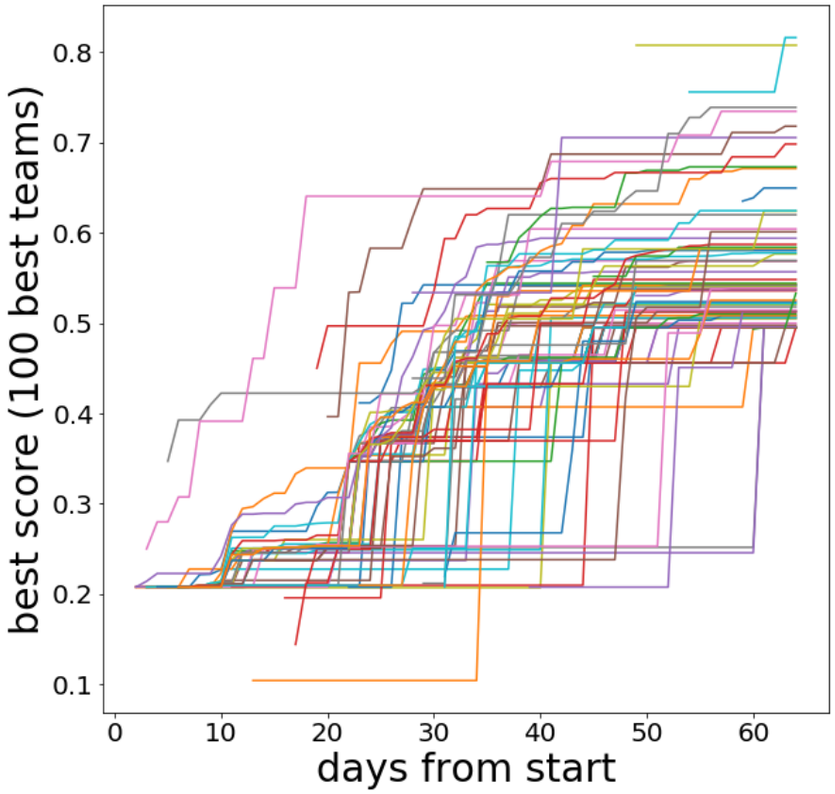


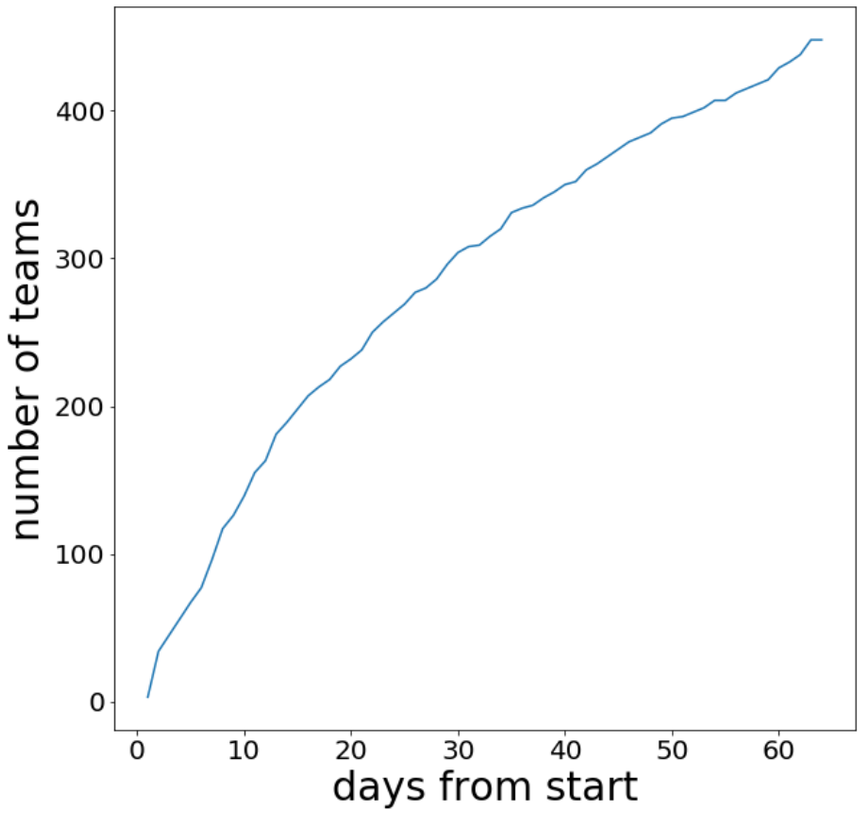
(*) DBSCAN see https://www.kaggle.com/mikhailhushchyn/dbscan-benchmark

Phase 1 Status

The competition is well under way

Steady increase of teams submitting and improving scores on leaderboard

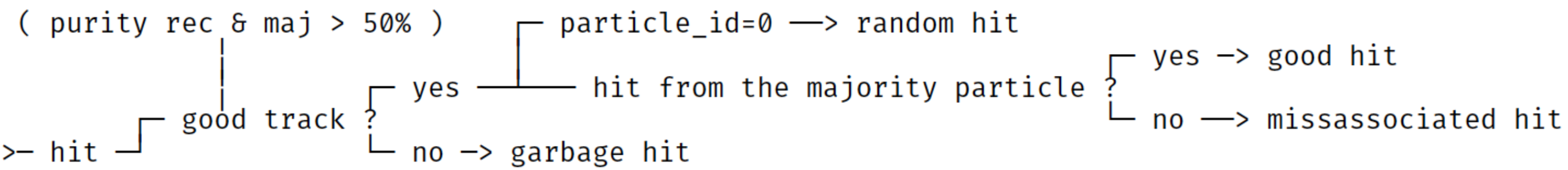


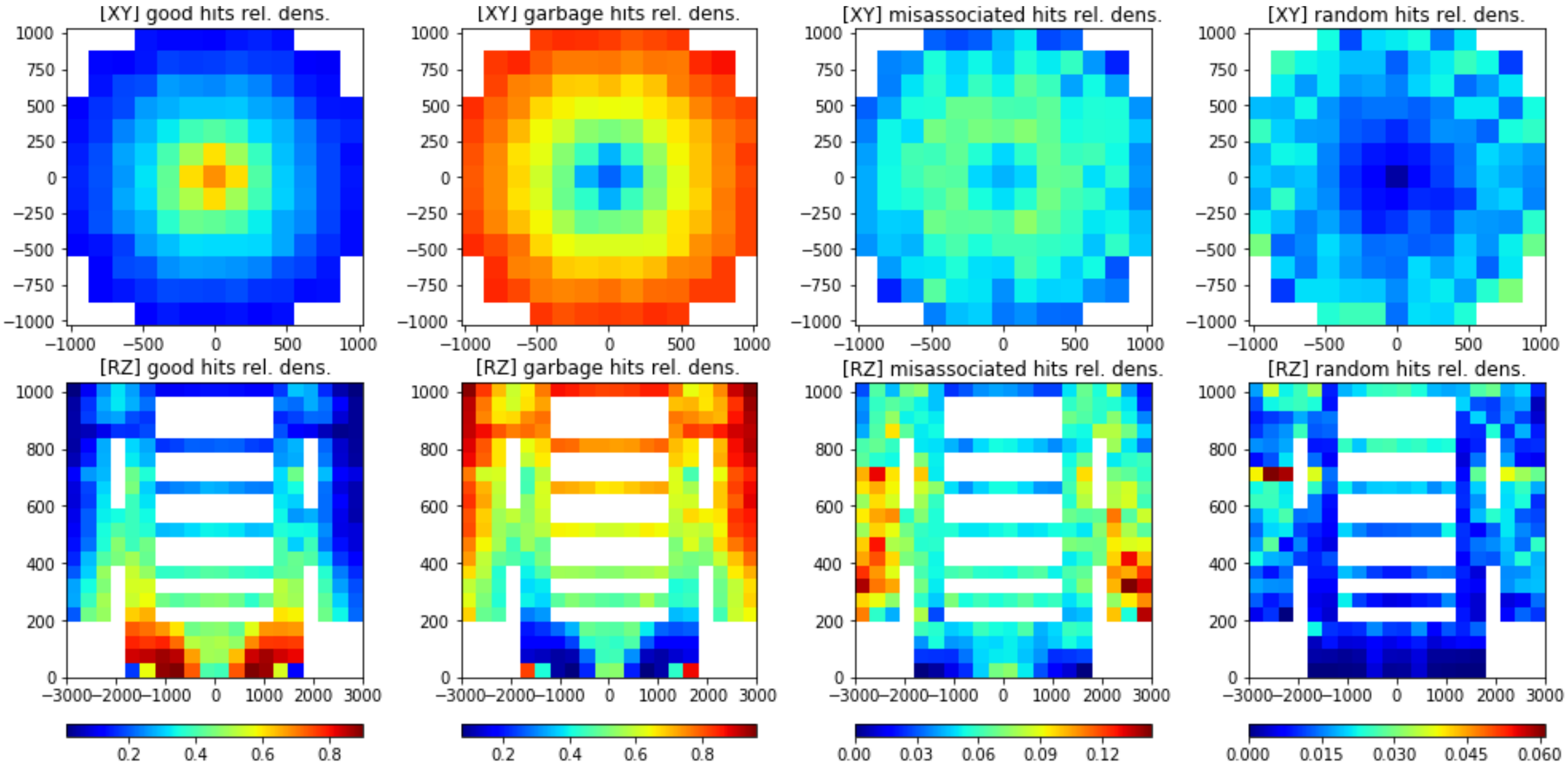


In the money Gold Silver Bronze

#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲1	**Edwin Steiner**			0.8161	2	14h
2	▼1	**outrunner**			0.8070	1	15d
3	—	**yuval r**			0.7389	14	8d
4	—	**Mickey**			0.7345	10	6d
5	▲1	**Zidmie**			0.7180	10	1d

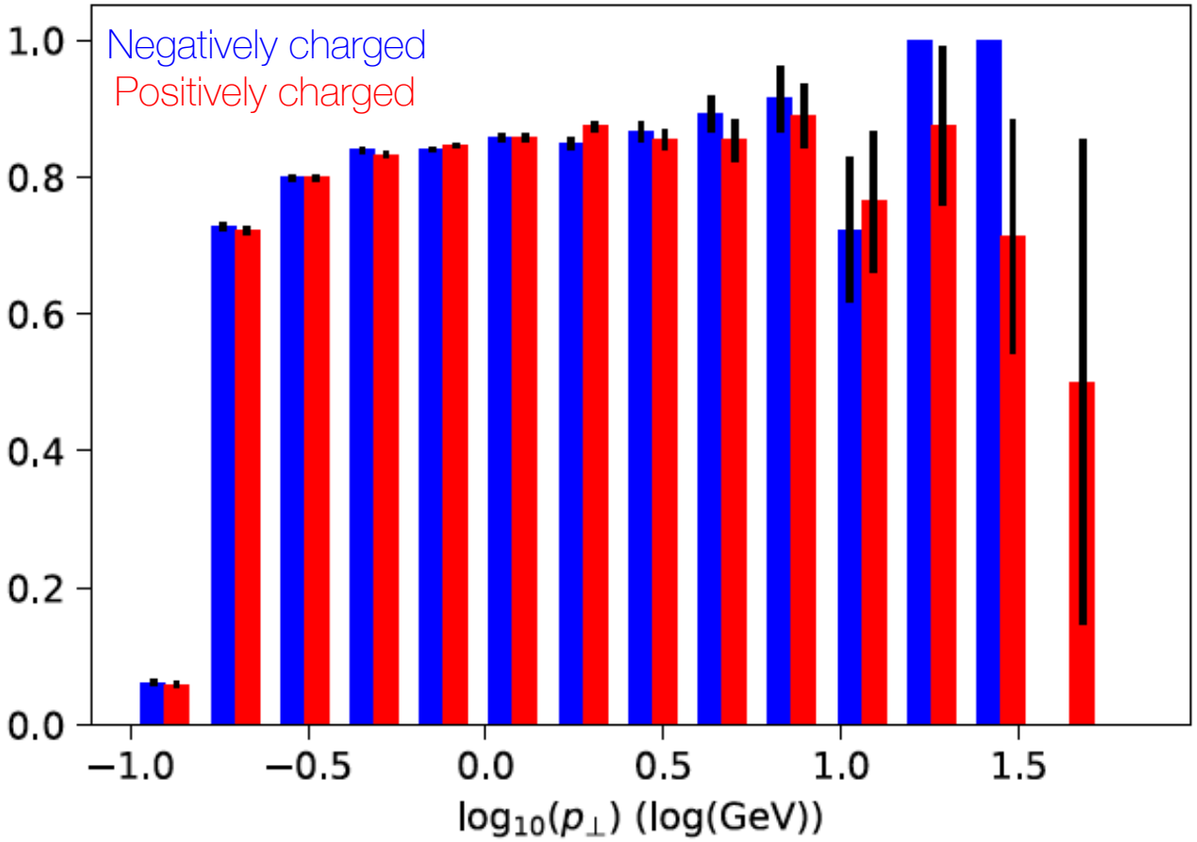
Preliminary analysis of submitted results typical hit map

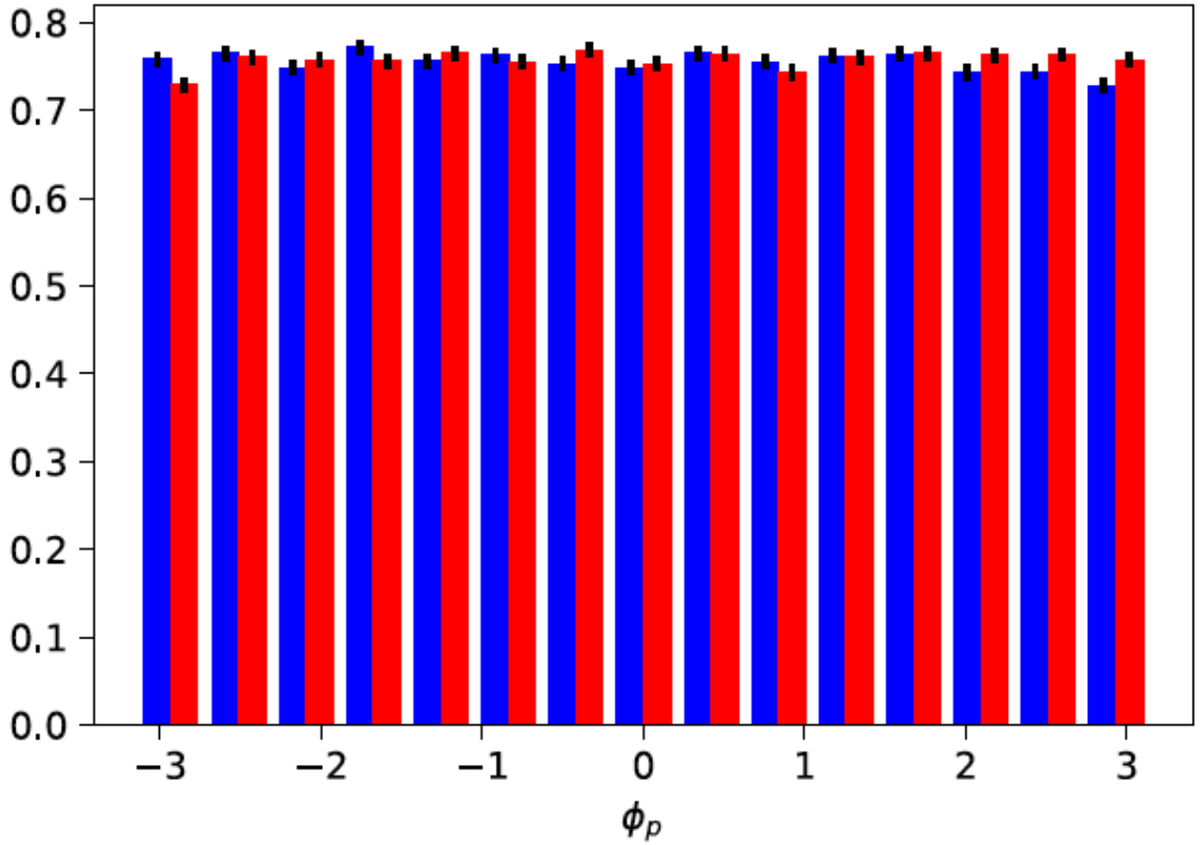




Preliminary results primary track efficiency vs. p_T and Φ

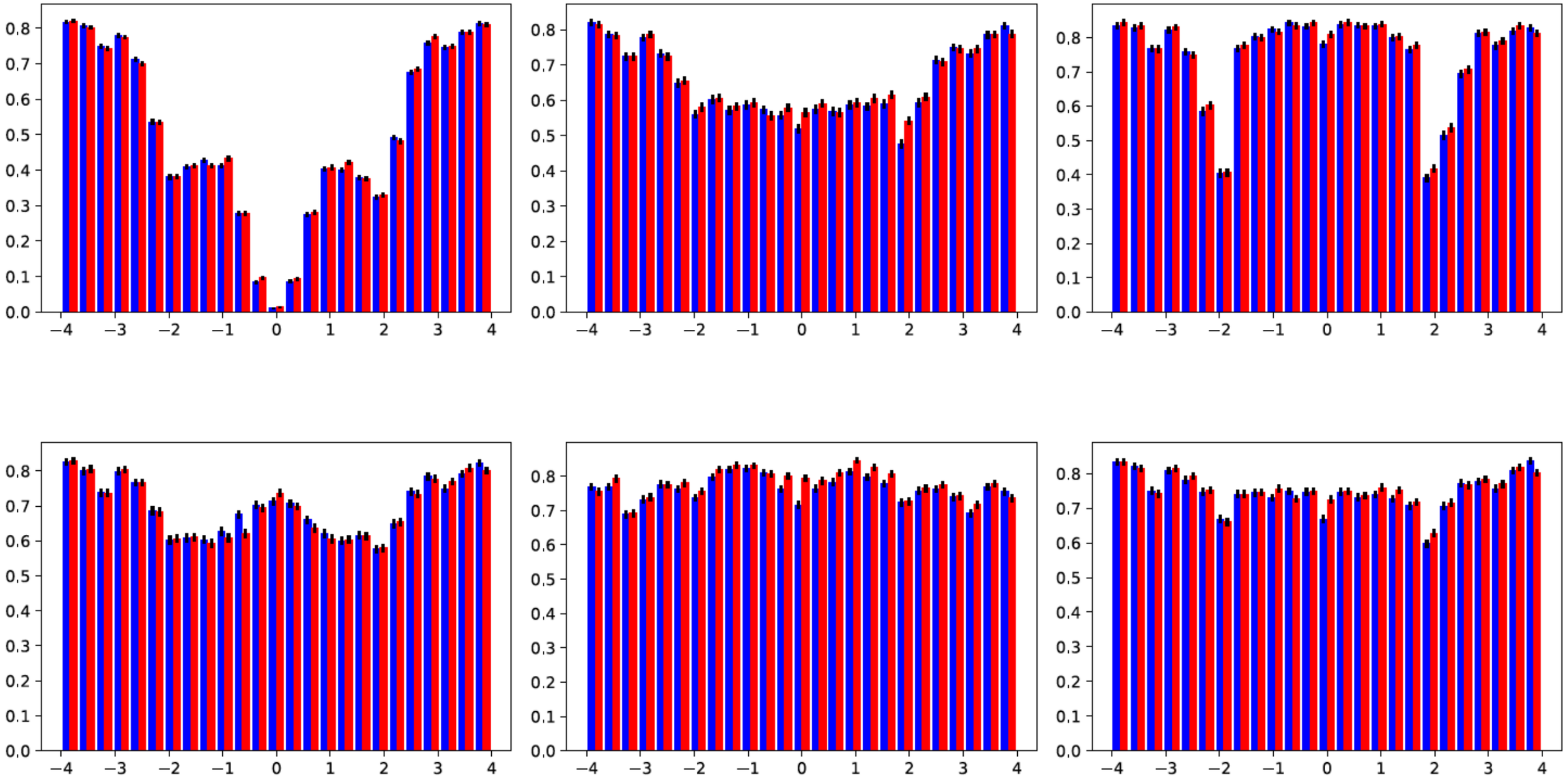
- No dependence on charge of charged particle nor on Φ
- Typically efficiency drops for low p_T





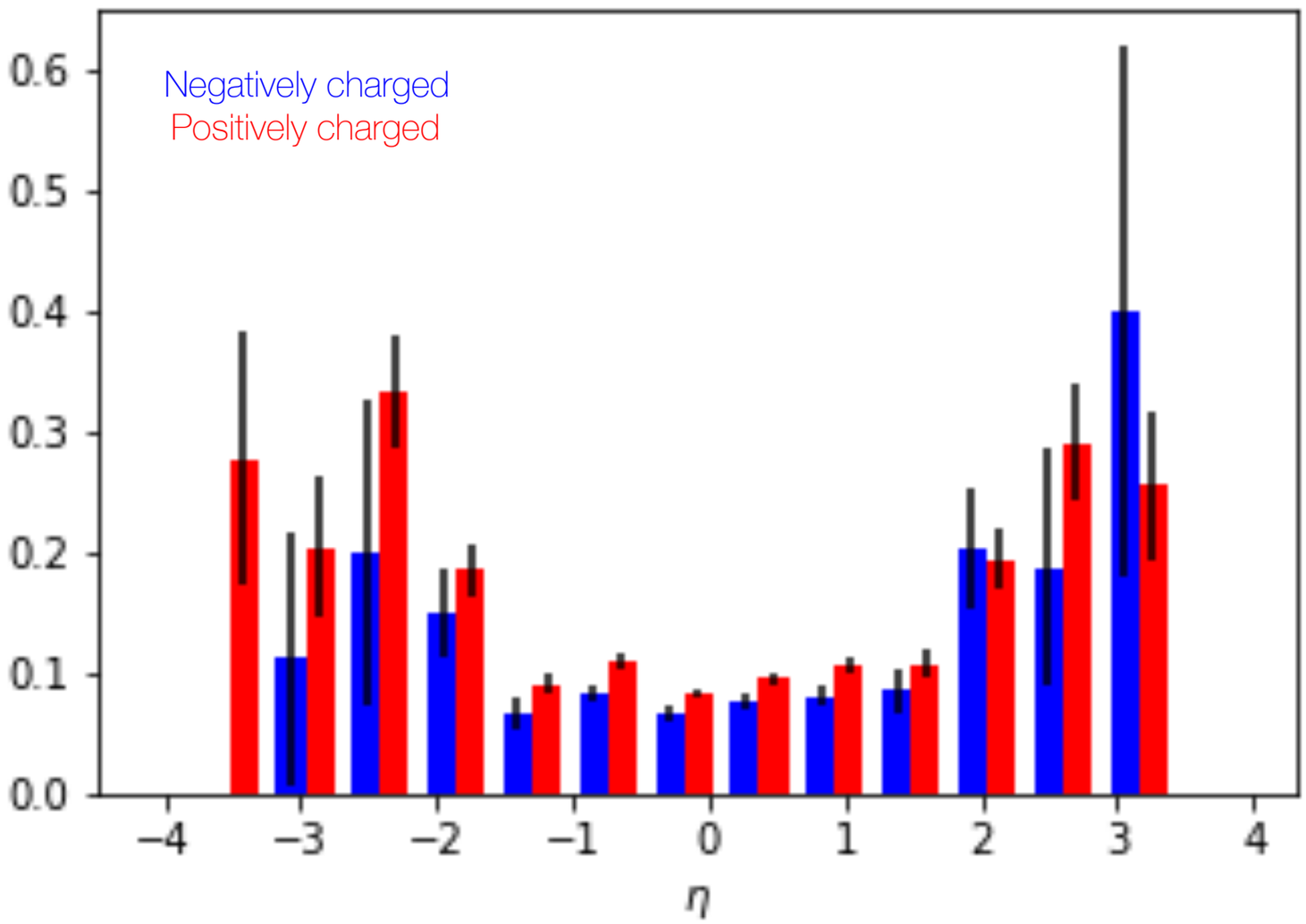
Preliminary results primary track efficiency vs. η: ordered by increasing score

Large variation in primary track efficiency vs. eta



Preliminary results secondary track efficiency vs. η

Low secondary track efficiency (shown vs. eta)



Summary & Outlook

- Initial DBSCAN was provided with starting score of 20%, quickly improved to 50% and now 70-80%
- No clear deficiencies identified with best solutions
- We don’t know yet what algorithms are used for best solutions

- Now: slower pace of progression, and less submissions
- Getting closer to money time: less sharing of information

- Results of accuracy phase will be advertised in September
- Results of Throughput phase (with code submission) in December

More Information

trackml.contact@gmail.com

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

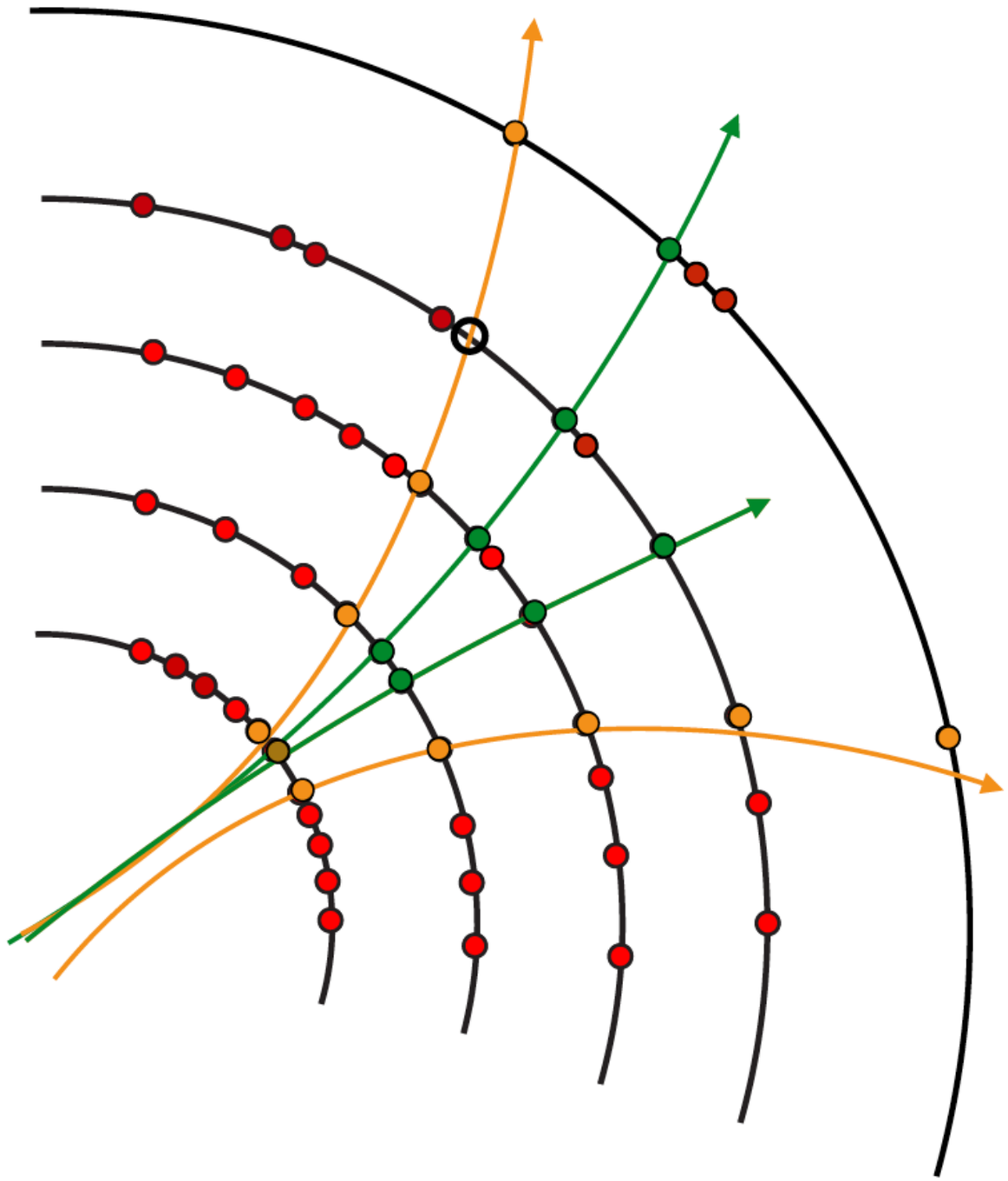
@trackmllhc

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

Backup slides

Track evaluation

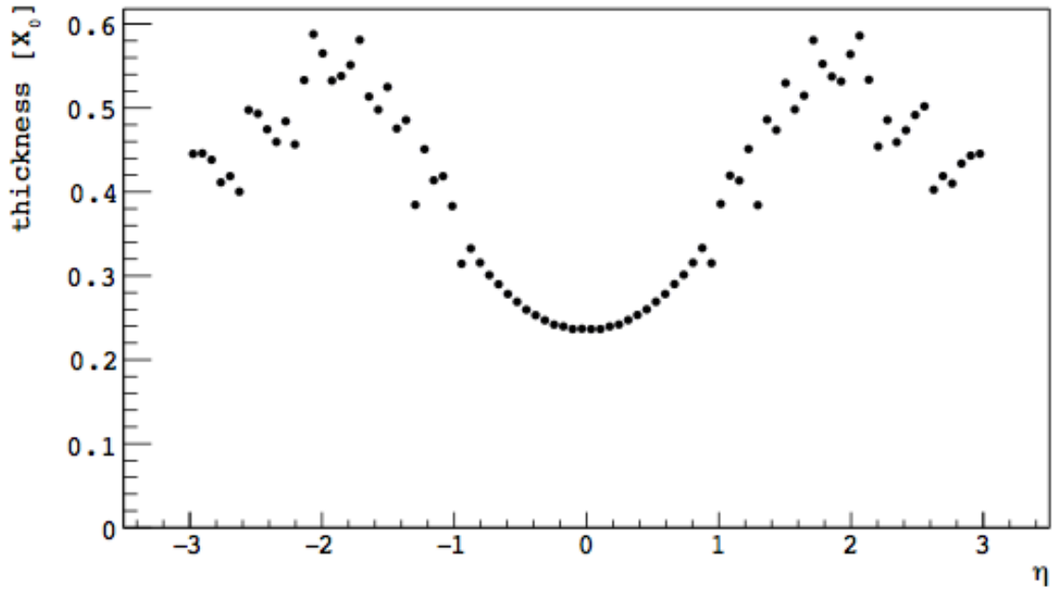
good track	**not so good track**
many compatible hits	short tracks
completeness	holes
uniqueness	shared hits
low χ^2/ndf	bad fit quality, outliers
small impact parameter (for primaries)	

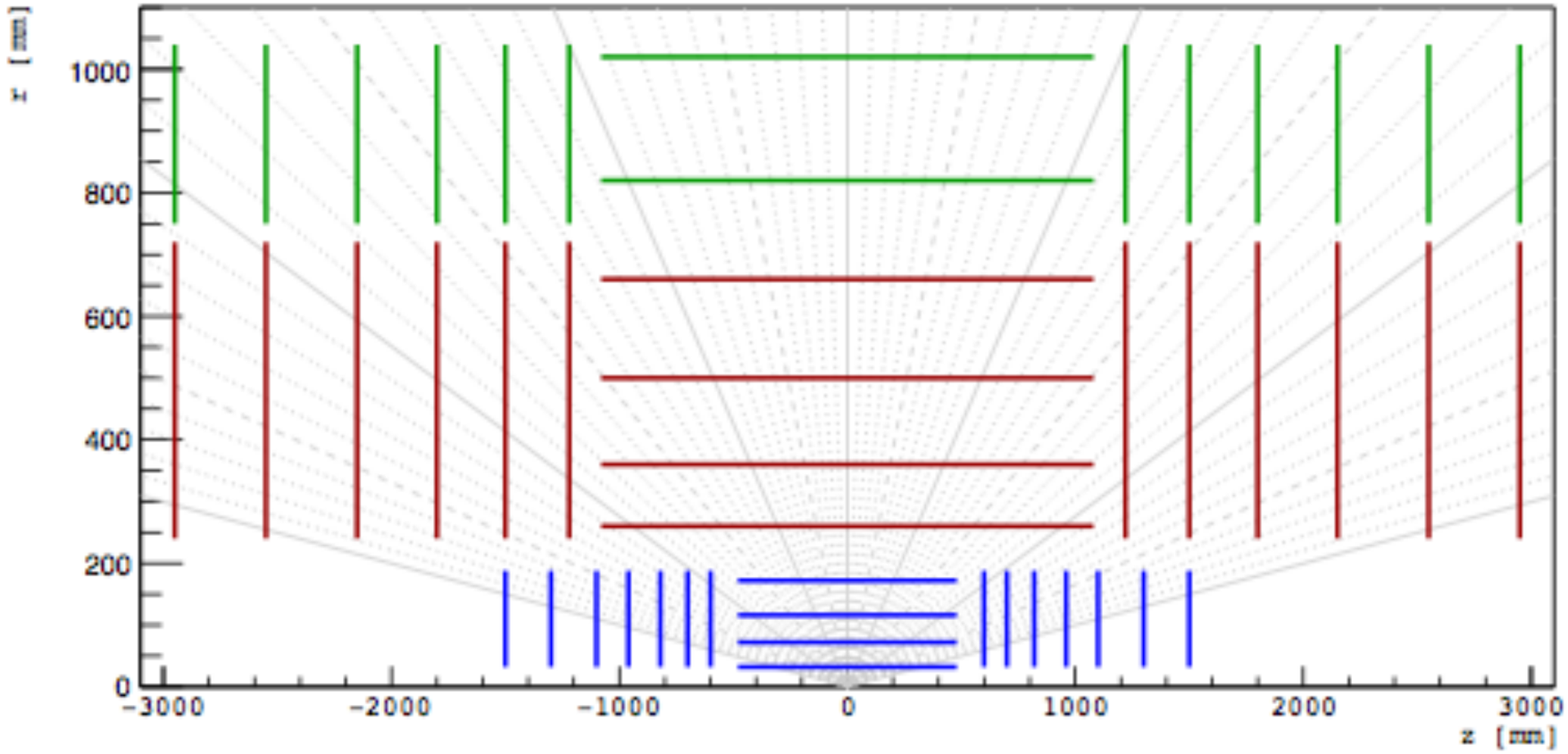


The **detector**

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

TrackML Why

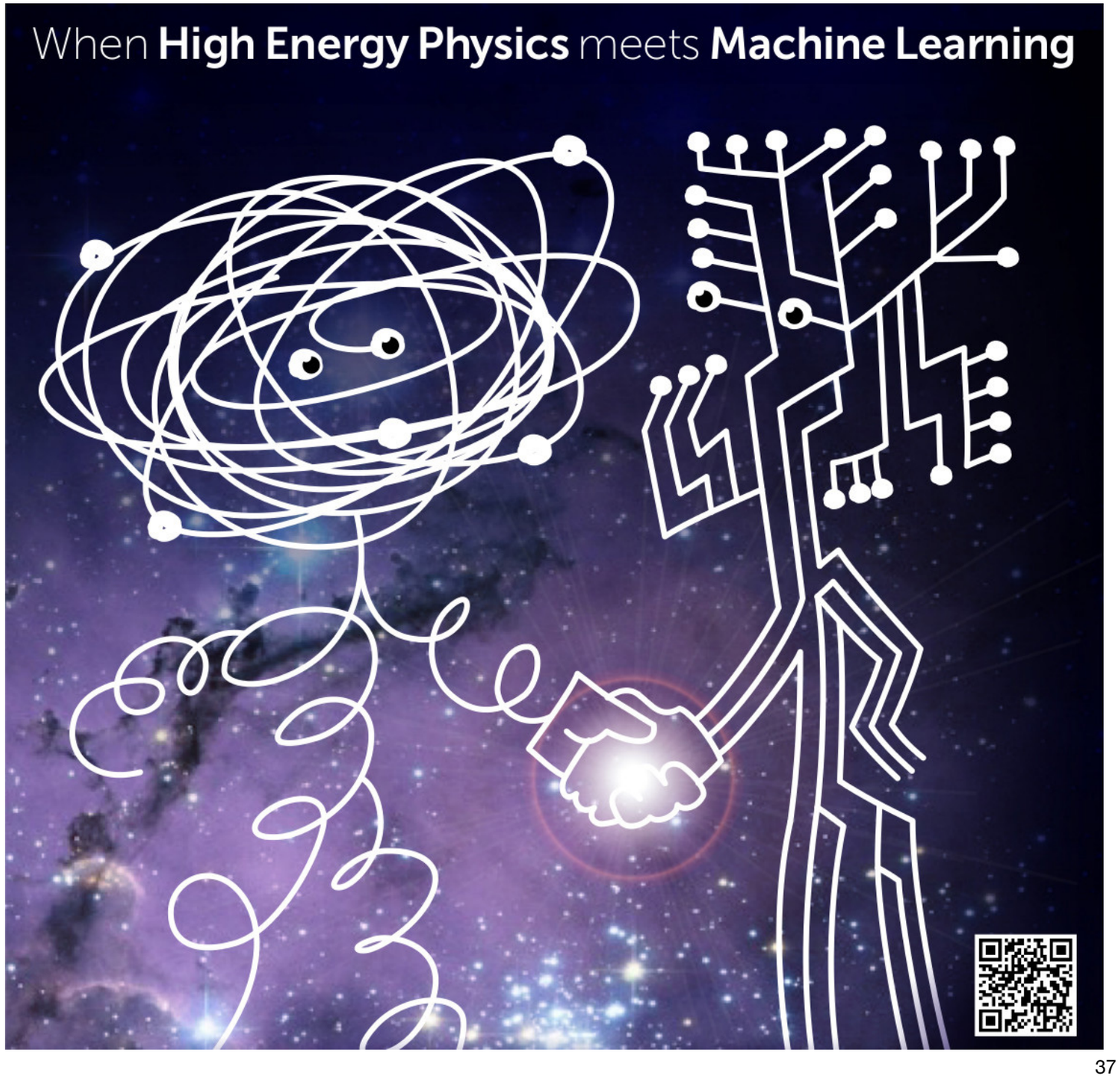
Machine learning is a rapidly growing sector

- strongly supported by industry (including special hardware development)
- a lot expertise **inside & outside** the HEP community (ML is all but new for us)

The Higgs ML challenge was a great success

- even though best solution was not new for us (BDT)
- sparked a lot of interest
- great publicity

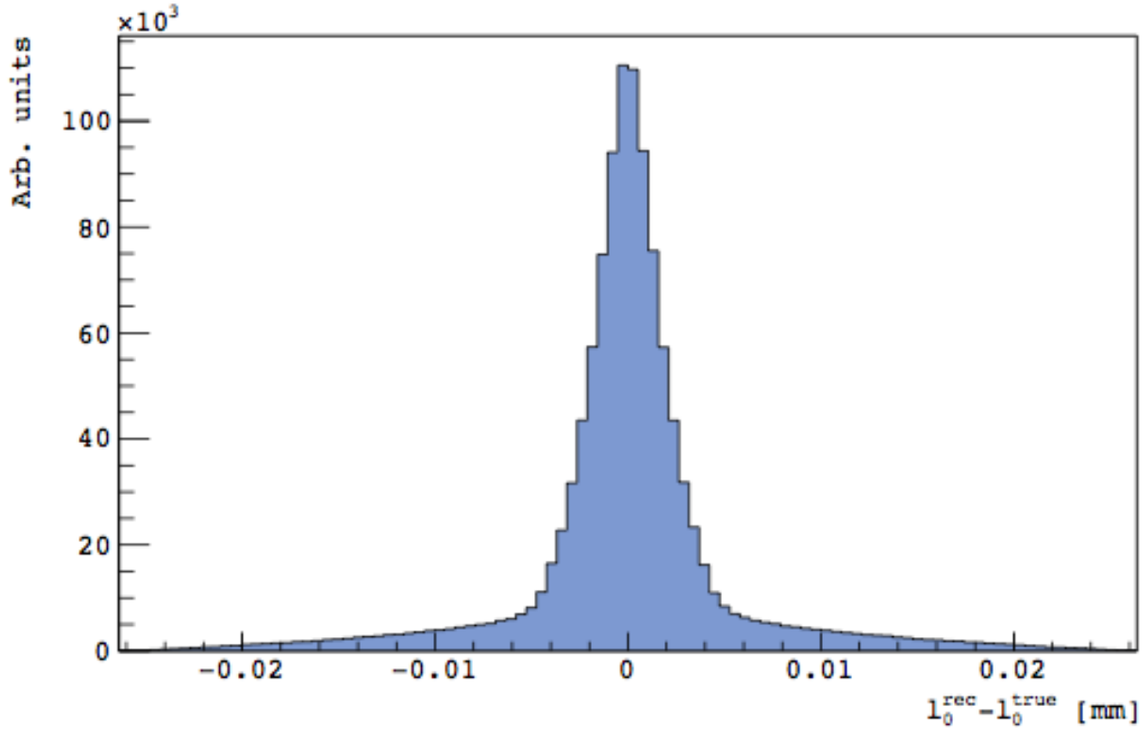
Started discussion in 2015 about a Tracking ML challenge

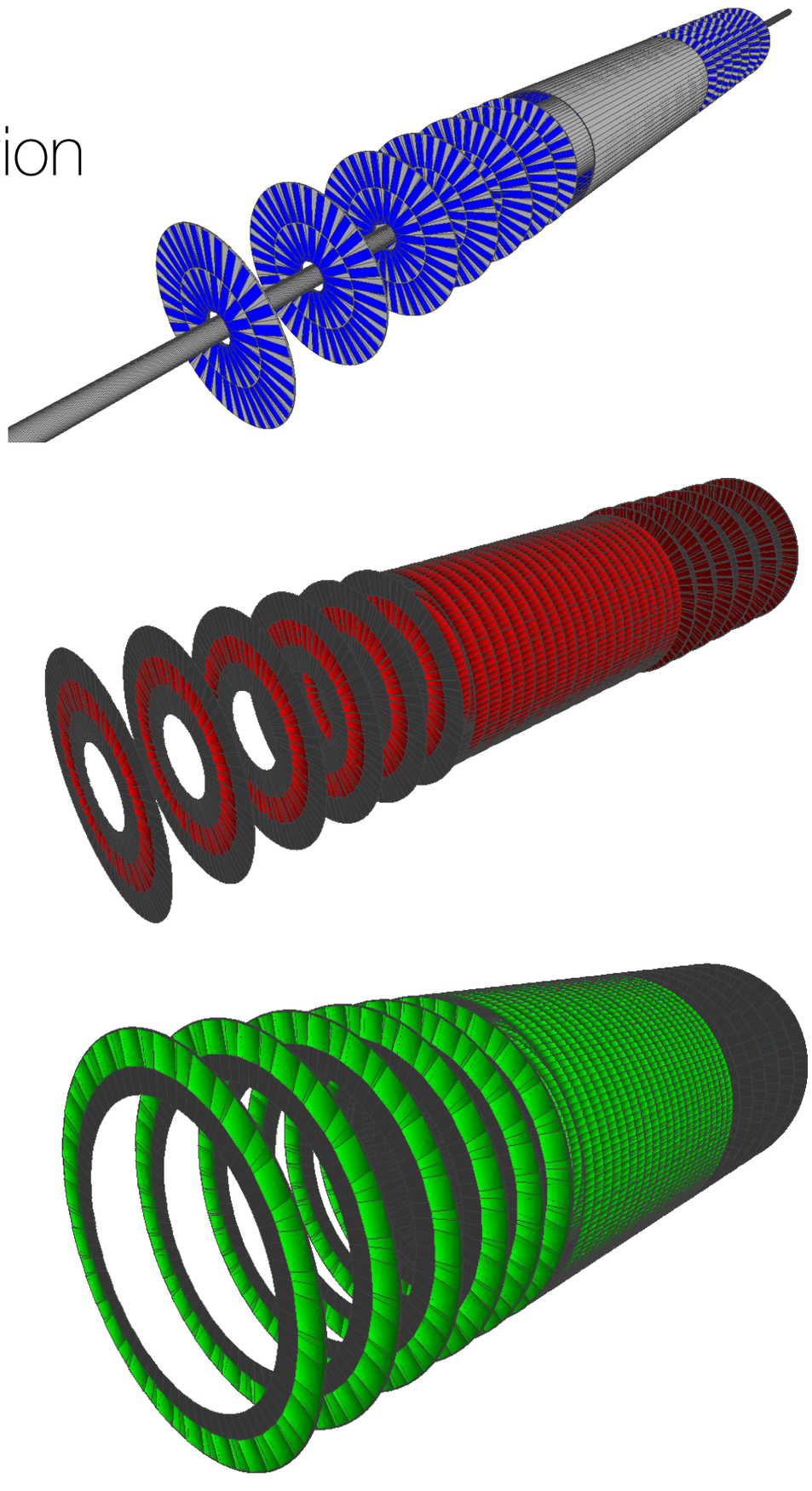


The **detector** (2)

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)



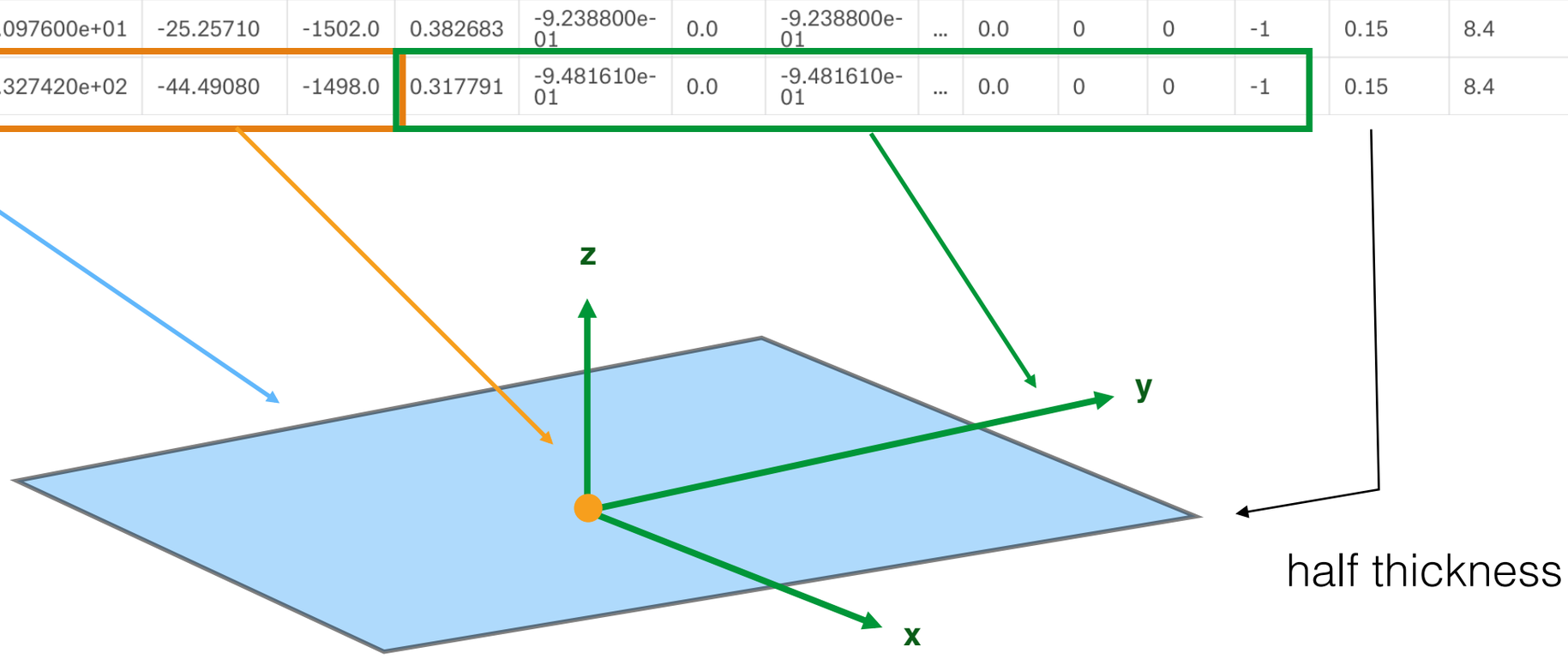


plot & images
(left) estimated pixel resolution distribution
(right) 3D view of pixel, short strip and long strip detector

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



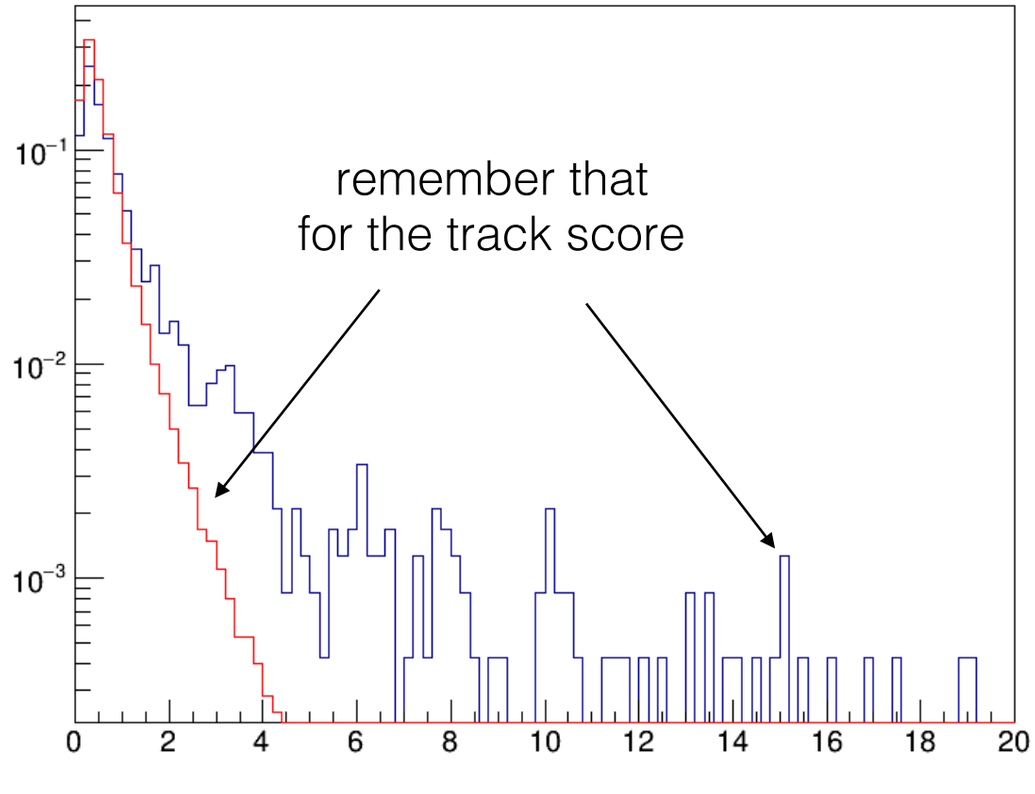
plot & image
(top) csv file format for the detector
(bottom) module center and orientation

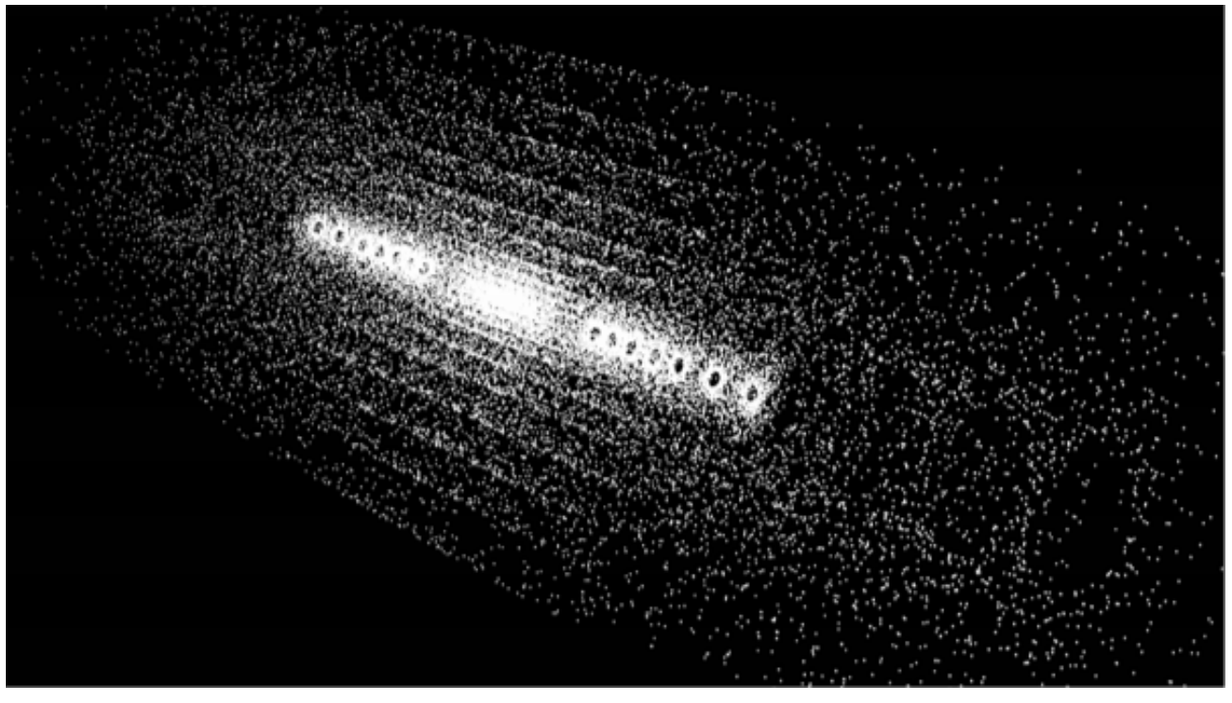
The dataset - physics

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

Smeared beam spot
- $\sigma_Z = 5.5$ mm, $\sigma_T = 15$ 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

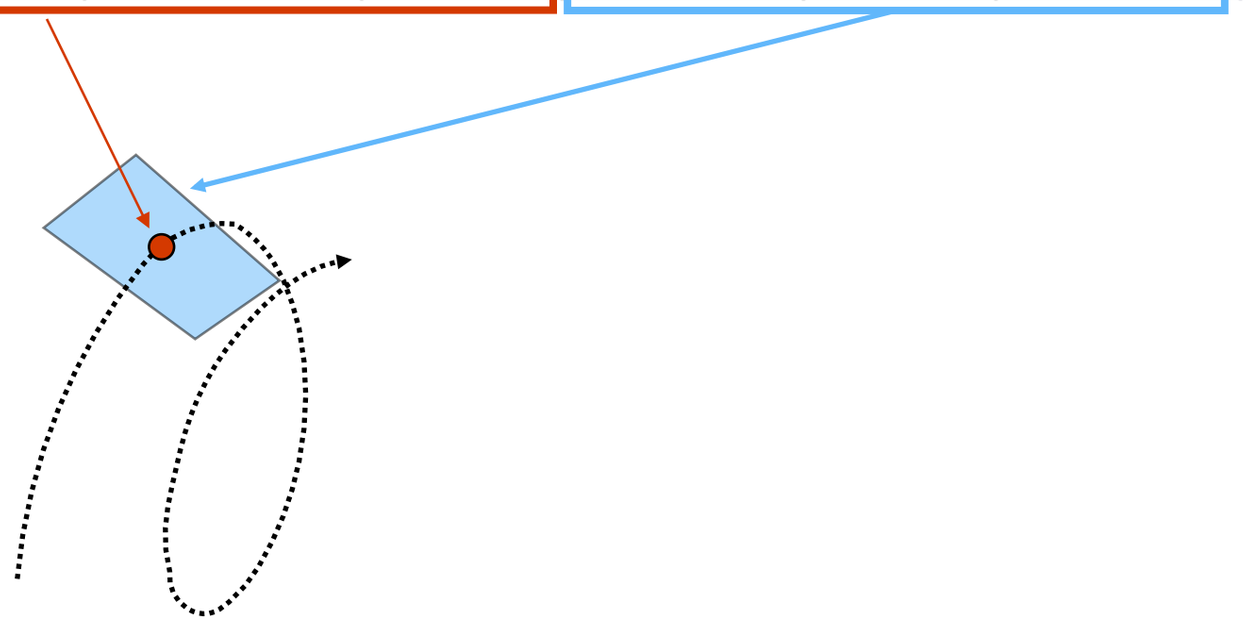


table & images
(top) csv file format for the hit file
(bottom) illustration of the hit information

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

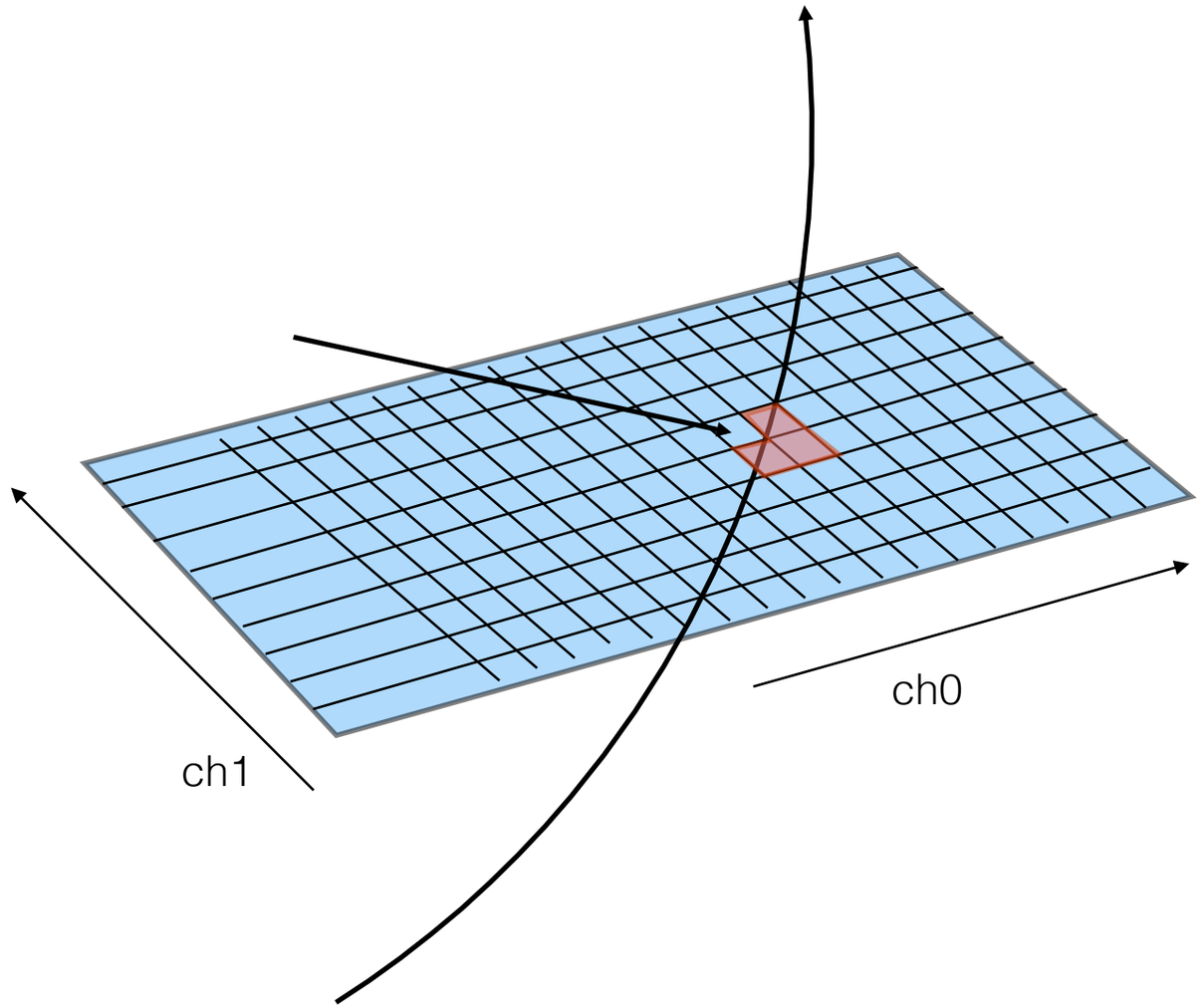


table & images
(top) csv file format for the hit file
(bottom left) csv file format of the cells information
(bottom right) cell information illustration

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

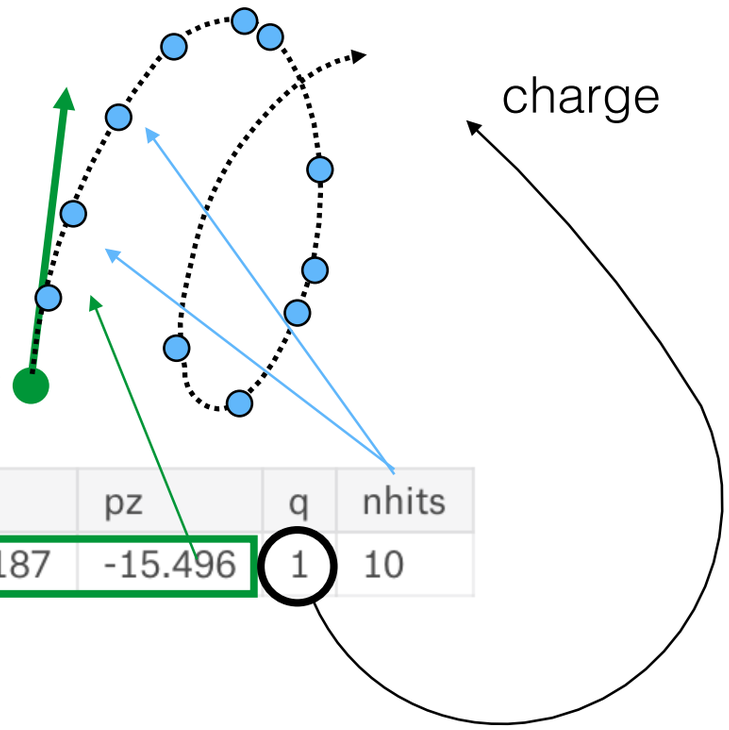
	hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight
0	1	0	-64.411598	-7.164120	-1502.5	250710.000000	-149908.000000	-956385.000000	0.000000
1	2	22525763437723648	-55.338501	0.630805	-1502.5	-0.570605	0.028390	-15.492200	0.000010
2	3	0	-83.828003	-1.145580	-1502.5	626295.000000	-169767.000000	-760877.000000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



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

The validation **dataset & solution**

Independent but structurally identical hit dataset

Public Leaderboard **Private Leaderboard**

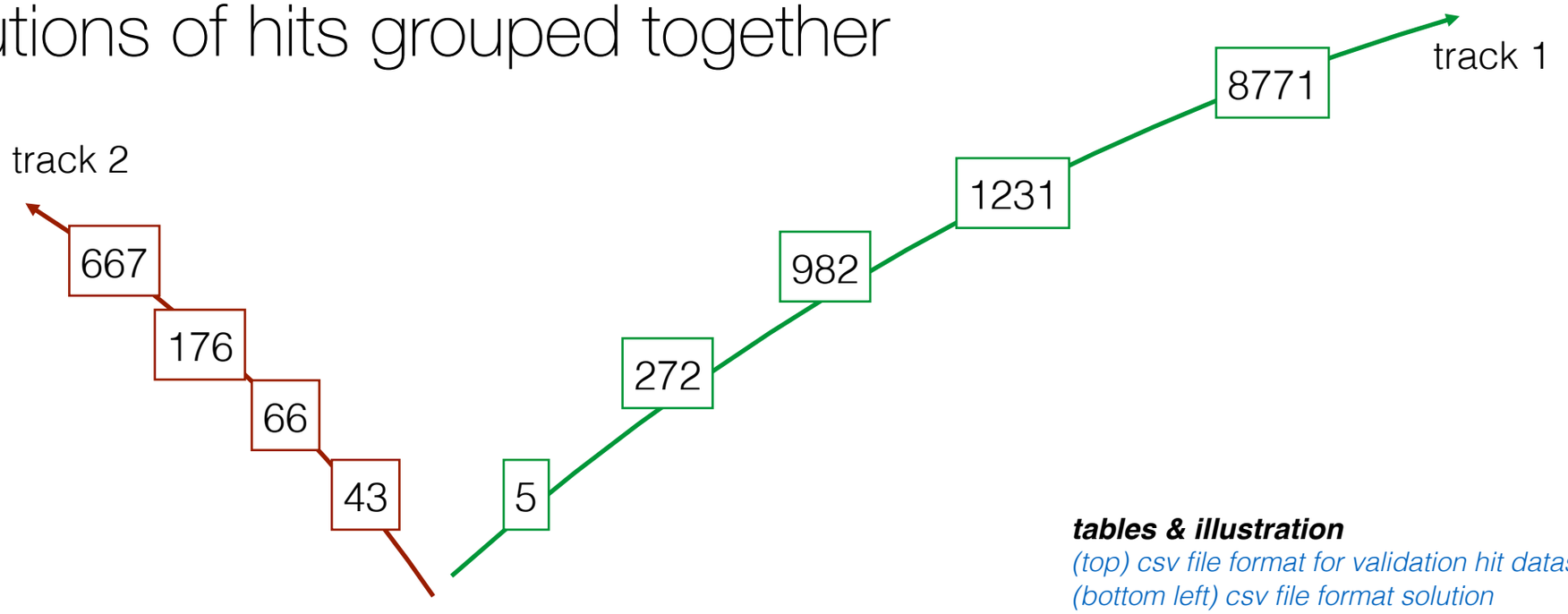
This leaderboard is calculated with approximately 29% of the test data.

The final results will be based on the other 71%, so the final standings may be different.

Raw Data Refresh

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

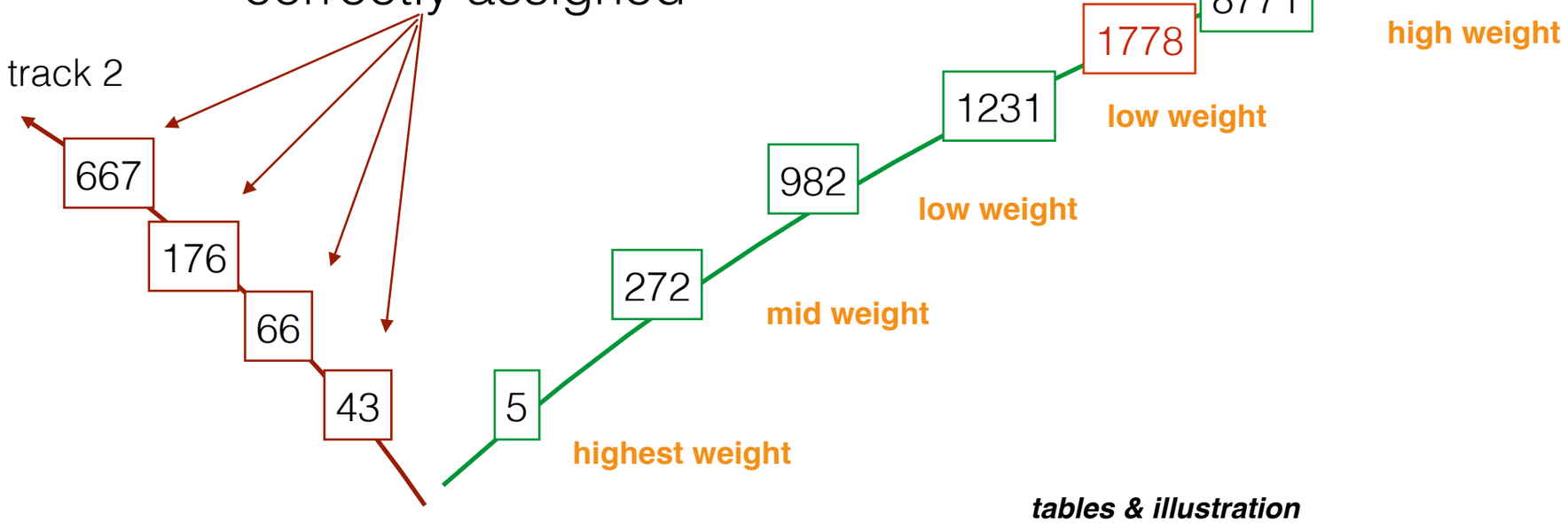
submission

solution.csv

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

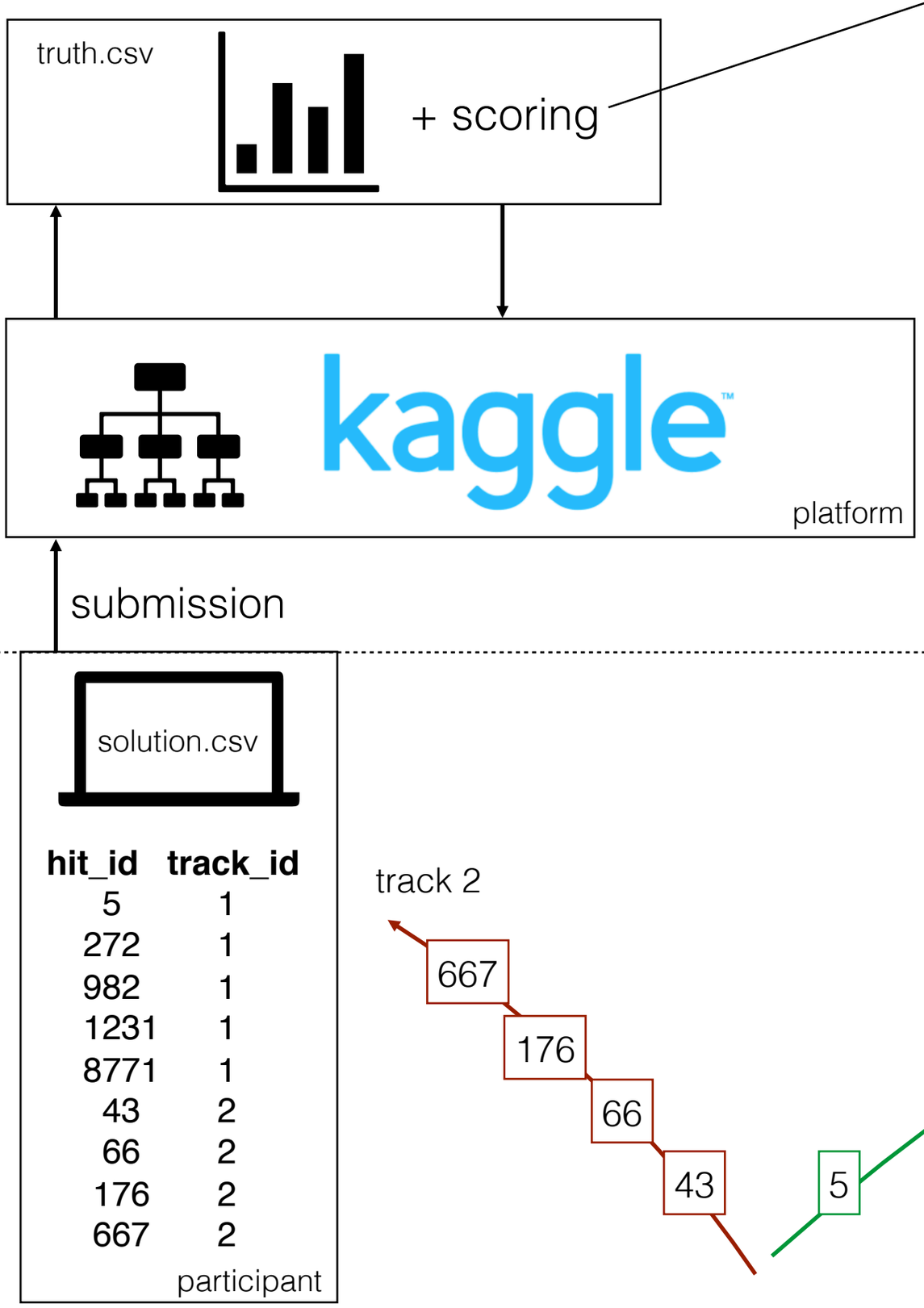
participant

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



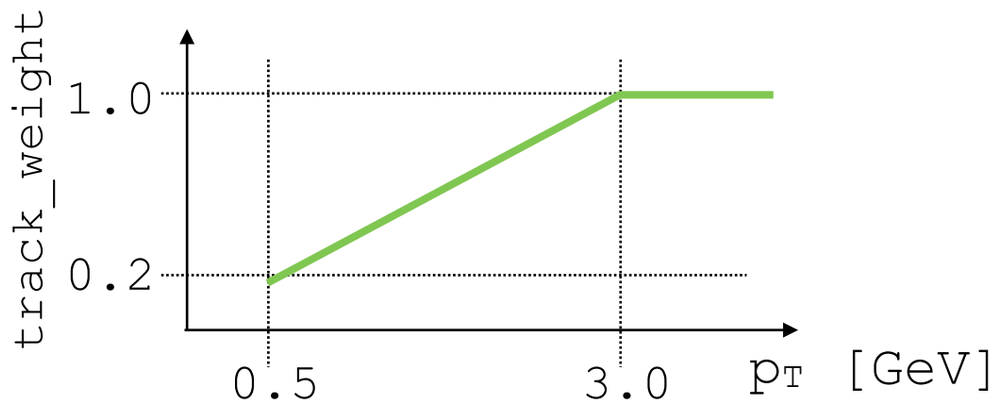
tables & illustration
(top) csv file format for validation hit dataset

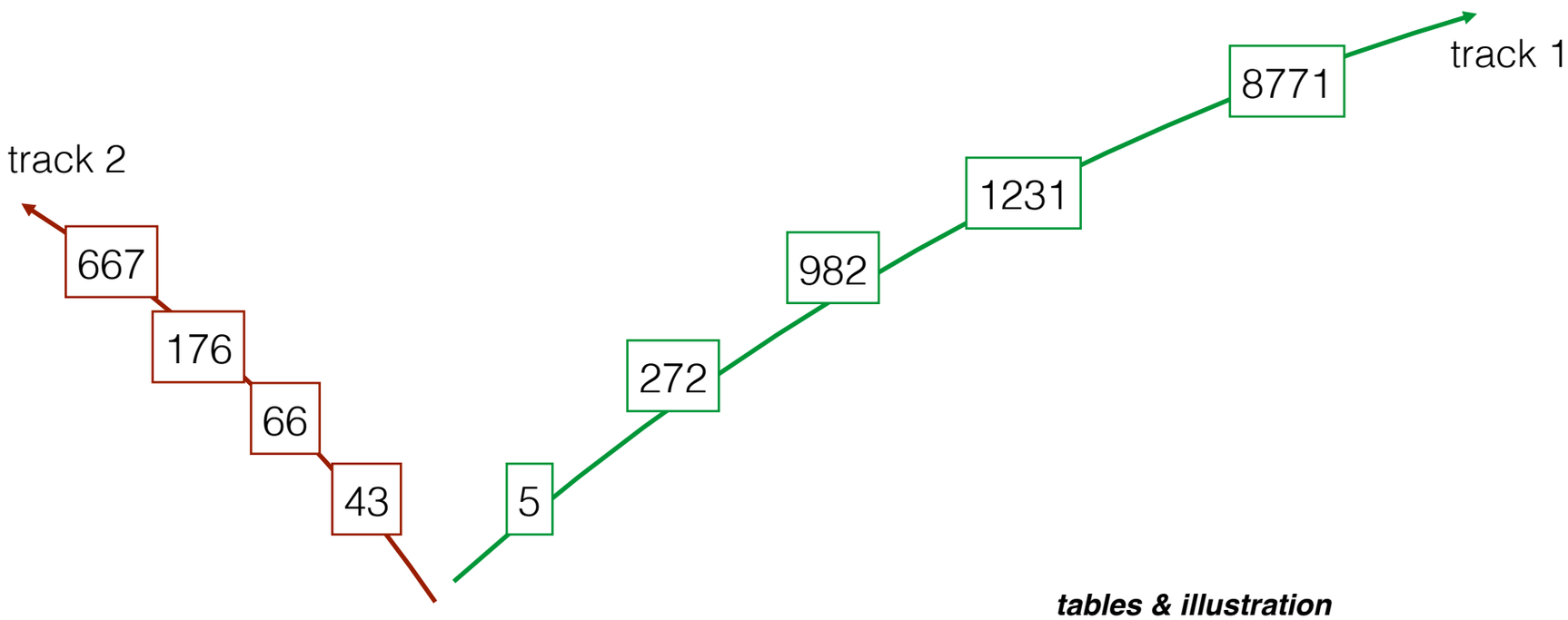
Submission & scoring (3)



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

higher momentum gives higher score:





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

Submission & scoring (4)

truth.csv

+ scoring

online leaderboard

1.	crazytrackers	0.89
2.	houghmods	0.877
3.	monsieurtraject	0.86
4.	4fcc	0.772

kaggle

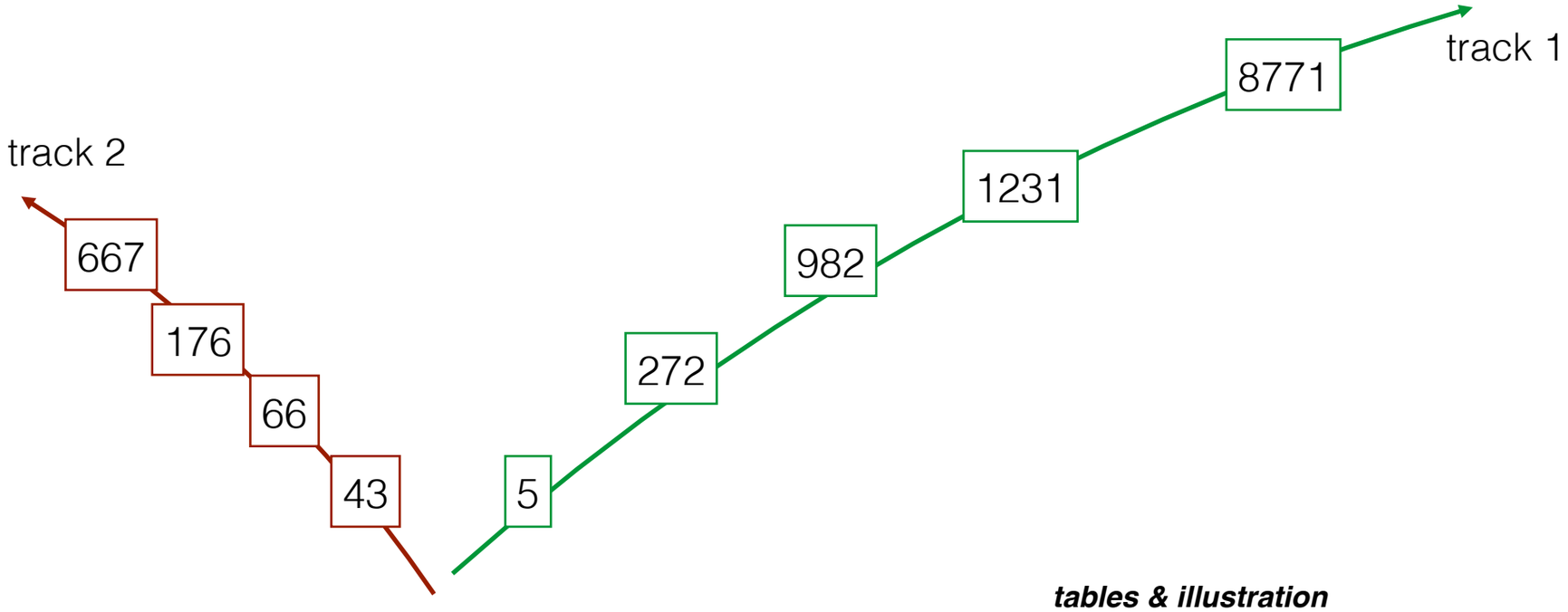
platform

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



tables & illustration

(top) csv file format for validation hit dataset