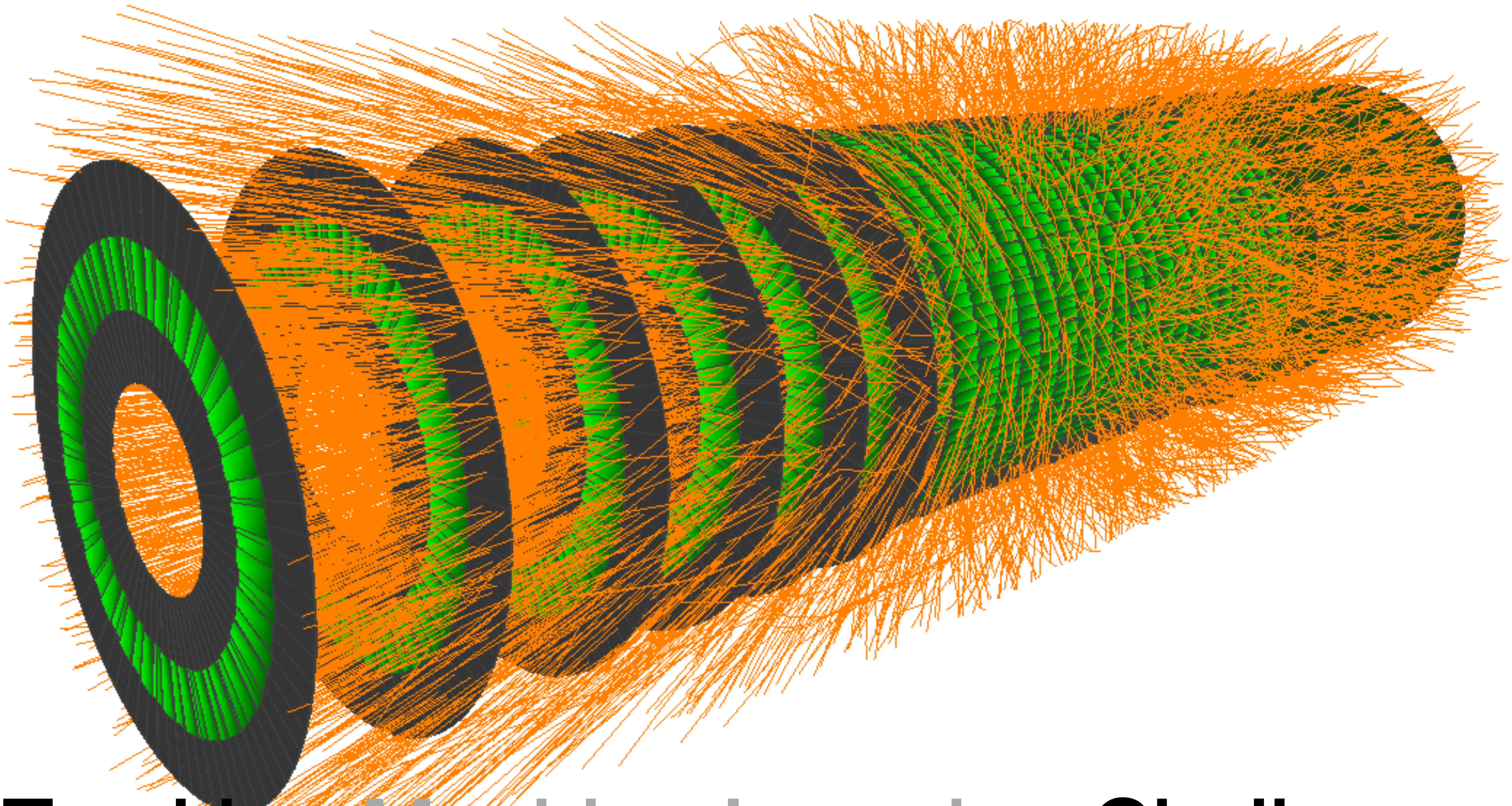


Tracking Machine Learning Challenge

Summary from Phase 1 & Phase 2

A. Salzburger (CERN)
@SaltyBurger

How it all began CTD2015 in Berkeley

David Rousseau liked

trackml @trackmllhc · 2h

It all started with some « what if » slides at the end of a talk on #HiggsML #kaggle challenge in the first occurrence of @ctdwit in Berkeley in 2015.

Andreas Salzburger @SaltyBurger

Nice sunrise in Valencia for the first day of @ctdwit - we will have a summary of both phases of @trackmllhc including a discussion session on Thursday

1 4

A HEP tracking pattern recognition challenge ?

Conclusion

- The Higgs Machine Learning Challenge successful in having ML experts tackle one specific HEP problem
  - re-import to HEP of ML techniques exposed on-going (and will take long)
- We (HEP) expect that breakthrough in pattern recognition would be invaluable to efficiently reconstruct future HL-LHC data
- ➔A Challenge on HEP pattern recognition could allow to make such breakthrough happen
- A personal note : I'm still quite busy with HiggsML, so I try to promote this idea, but I don't own it and cannot have a leading role in making it happen.

David Rousseau HiggsML and tracking challenges CTD 2015 Berkeley 39

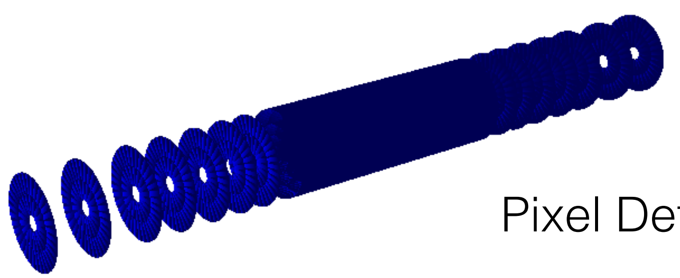
TrackML Who and How

Organisation team:

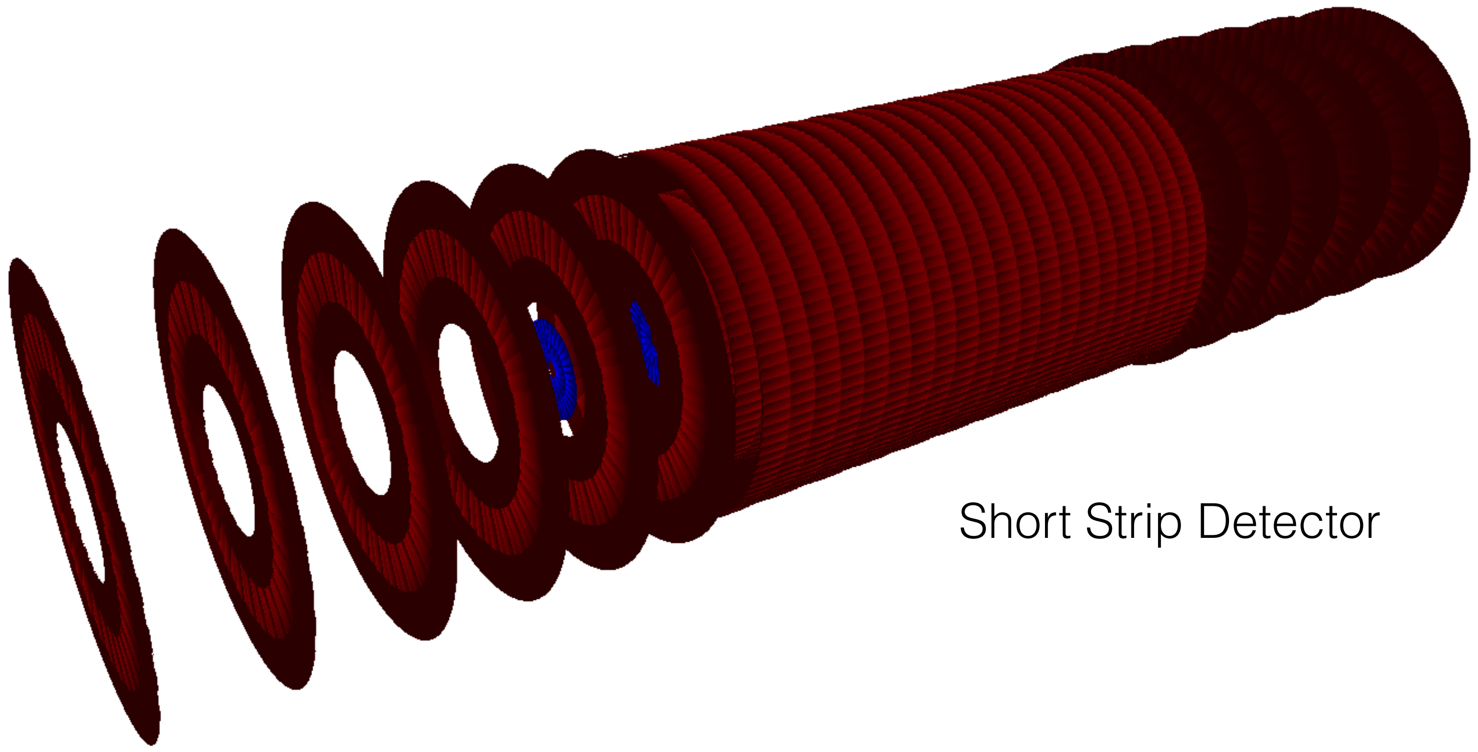
Jean-Roch Vlimant (Caltech), Vincenzo Innocente, Andreas Salzburger (CERN), Isabelle Guyon (ChaLearn), Sabrina Amrouche, Tobias Golling, Moritz Kiehn (Geneva University), David Rousseau, Yetkin Yilmaz (LAL-Orsay), Paolo Calafiura, Steven Farrell, Heather Gray (LBNL), Vladimir Vava Gligorov (LPNHE-Paris),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

Partners
& Sponsors

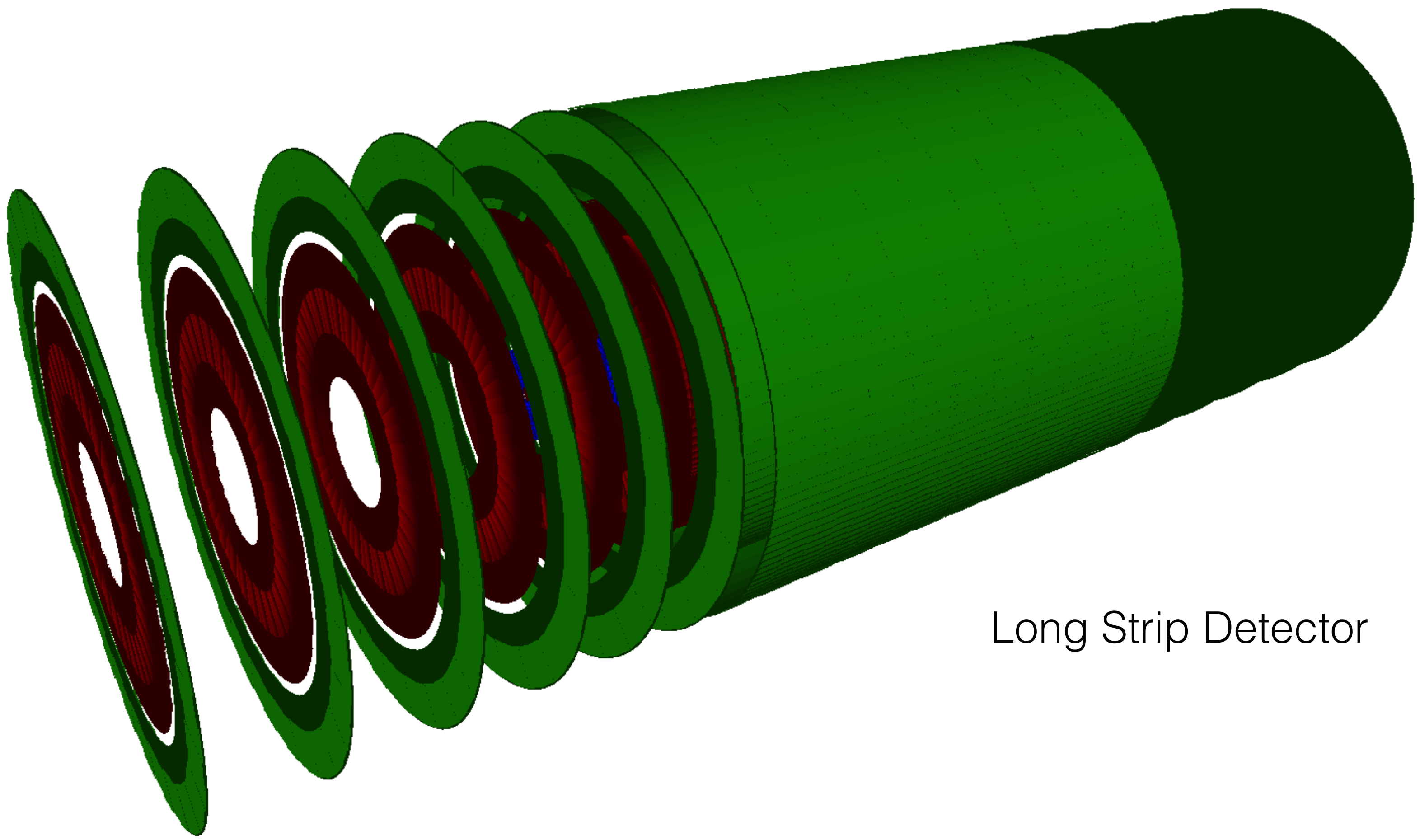
The challenge



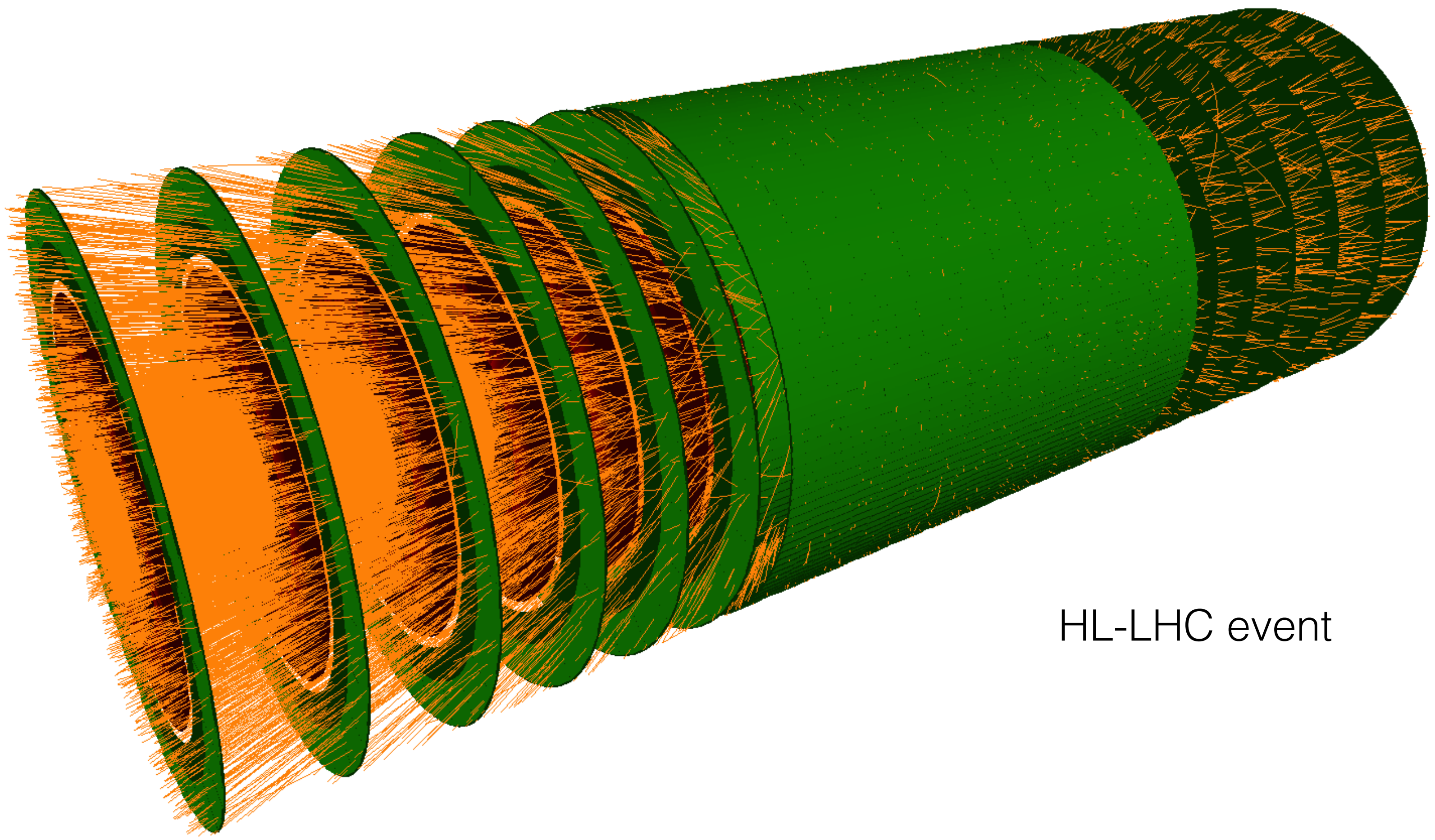
The challenge



The challenge

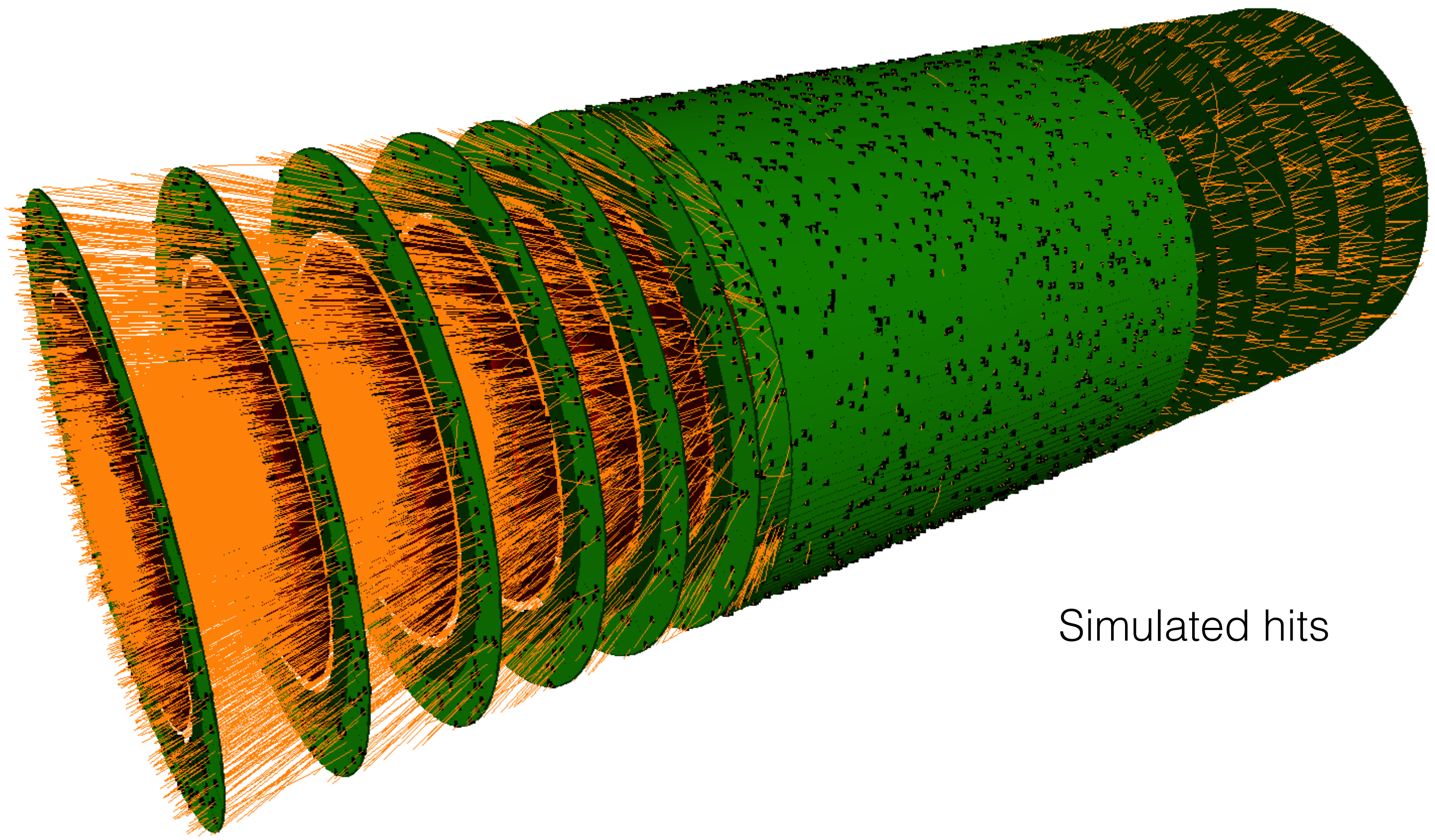


The challenge



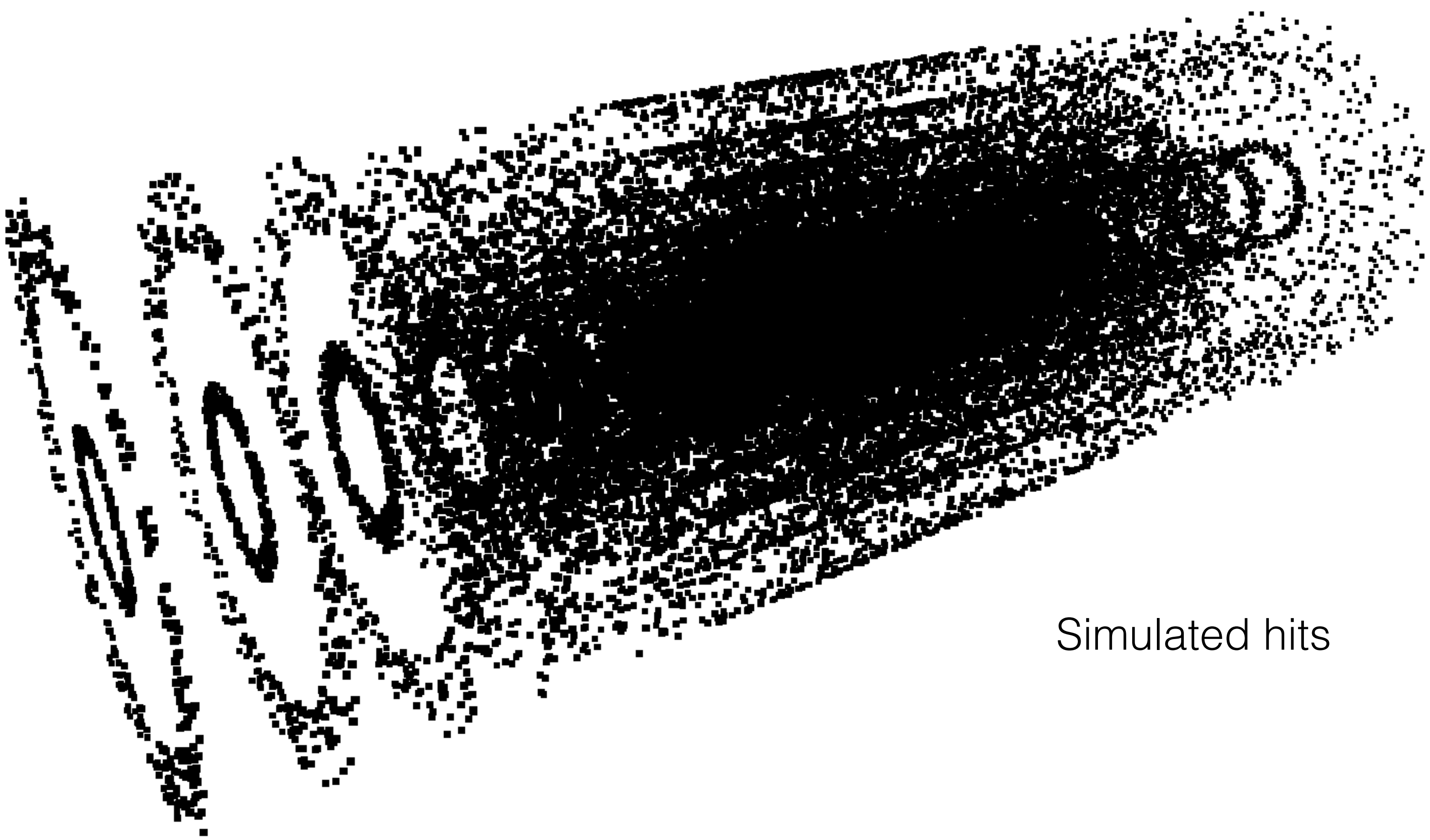
HL-LHC event

The challenge



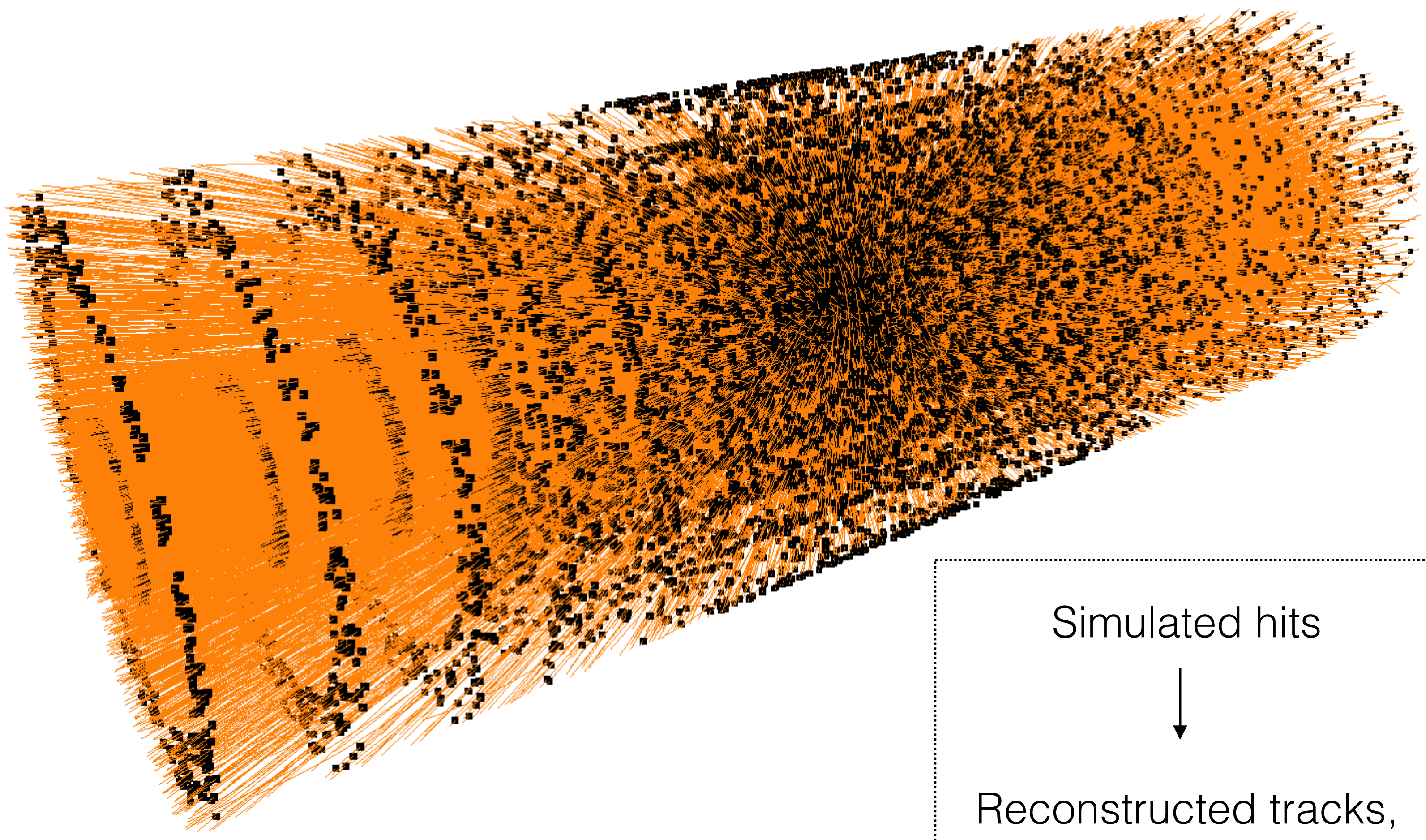
Simulated hits

The challenge



Simulated hits

The challenge

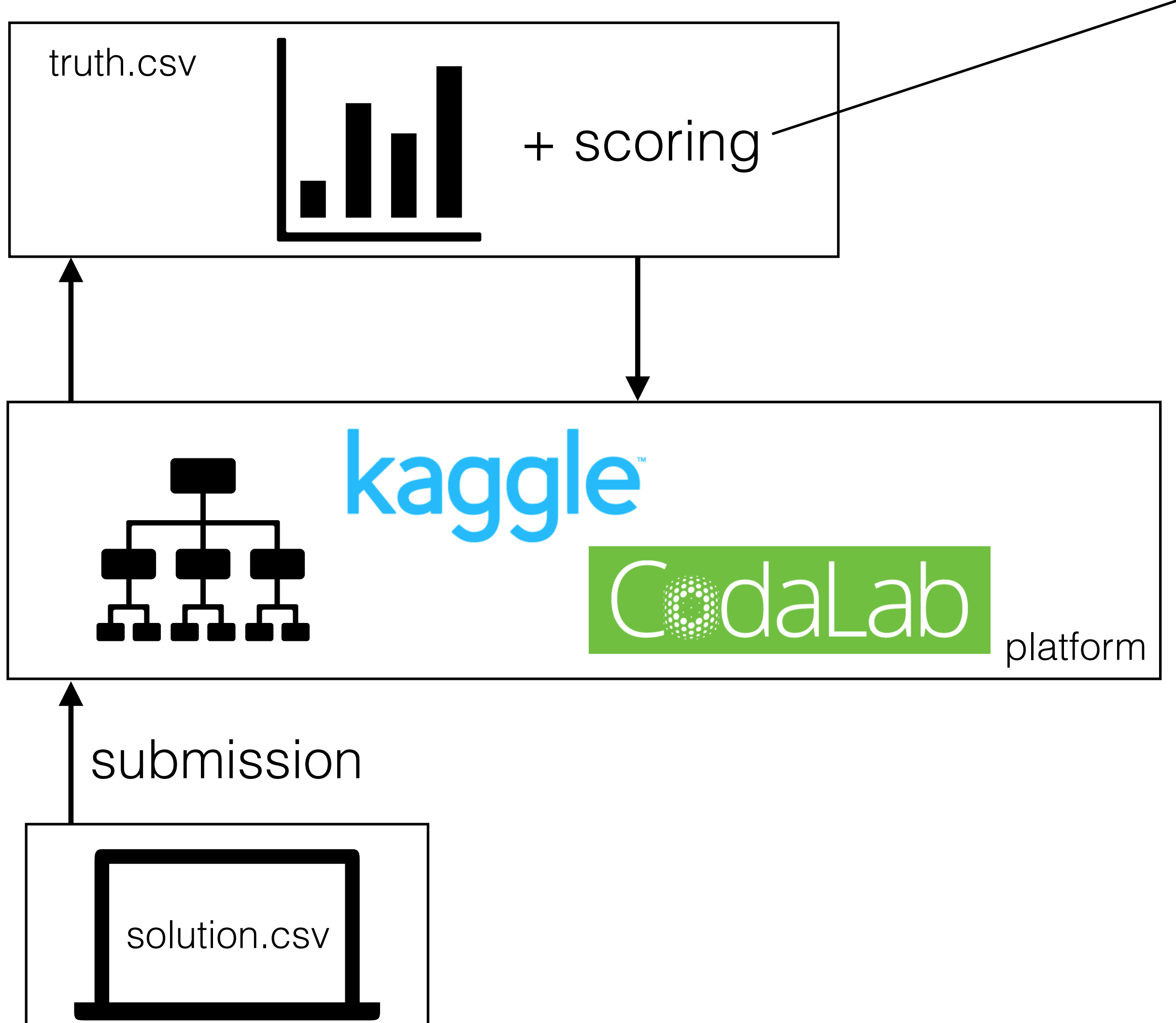


Simulated hits

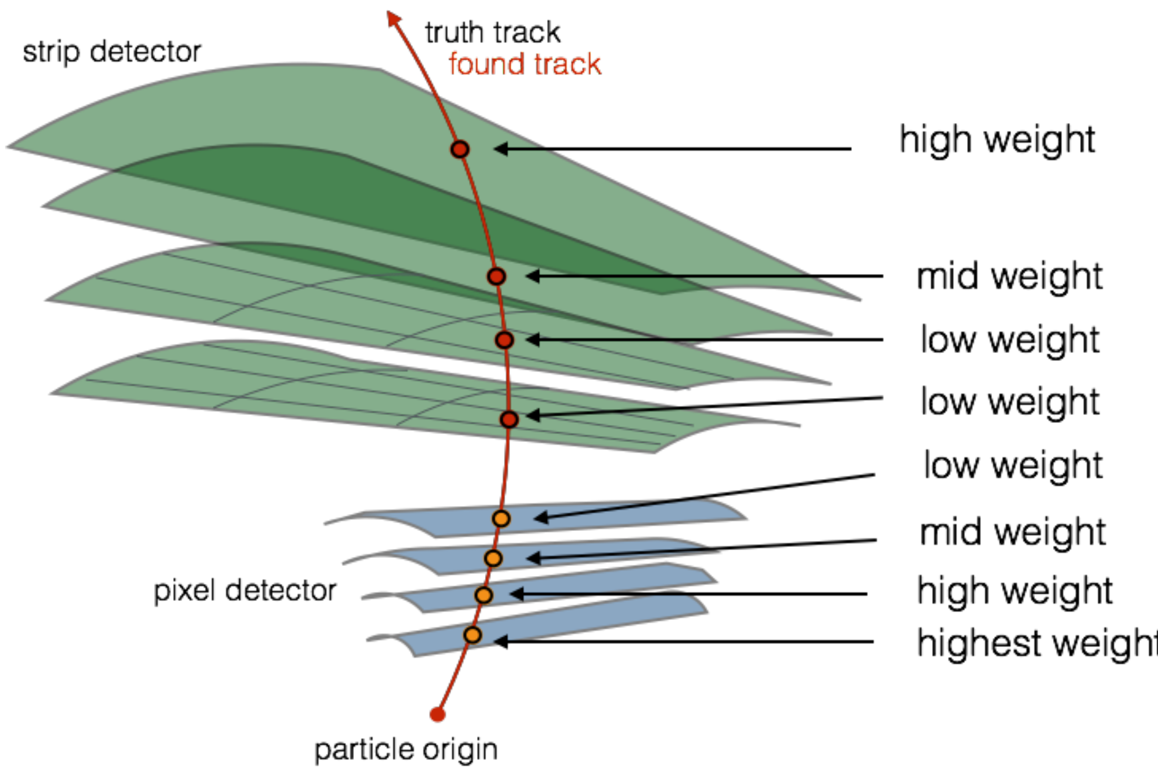
↓

Reconstructed tracks, Sequence of labelled hits

Submission

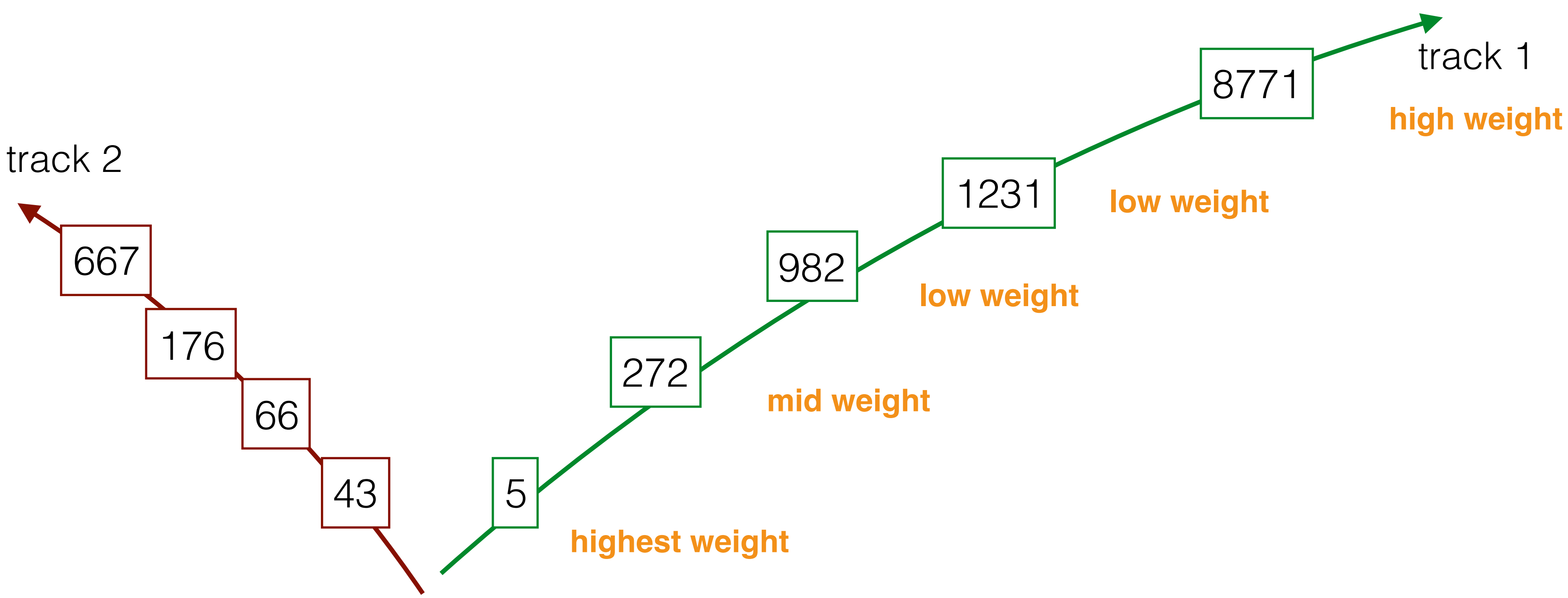


hits on track have **weights**

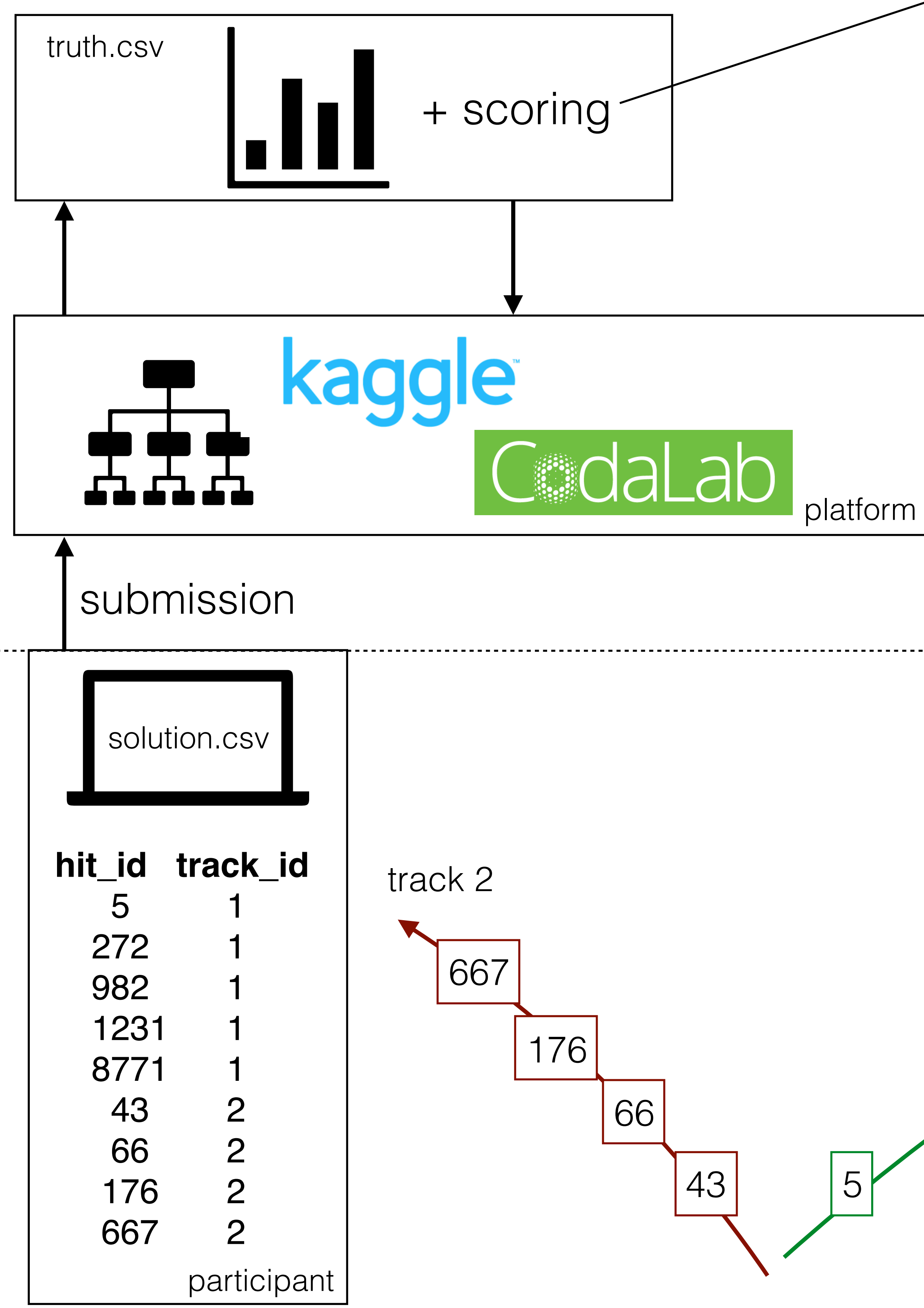


hit_id	track_id
5	1
272	1
982	1
1231	1
8771	1
43	2
66	2
176	2
667	2

participant



Submission & scoring

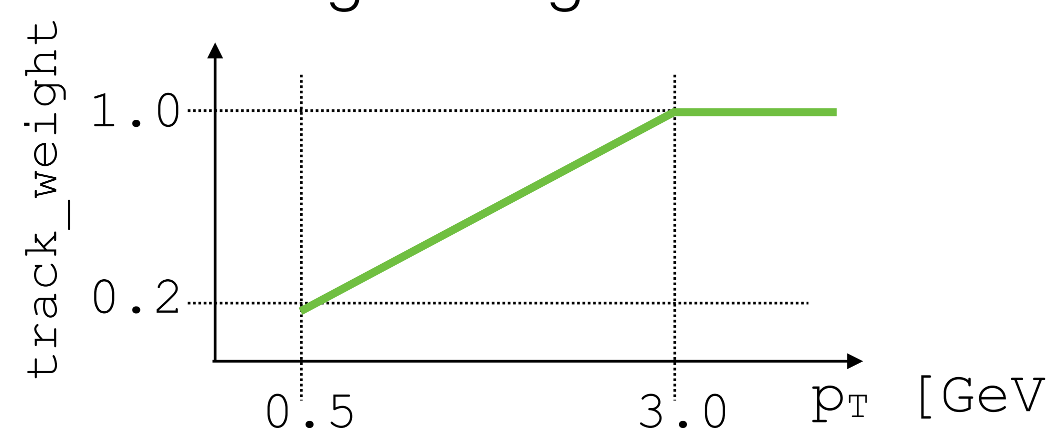


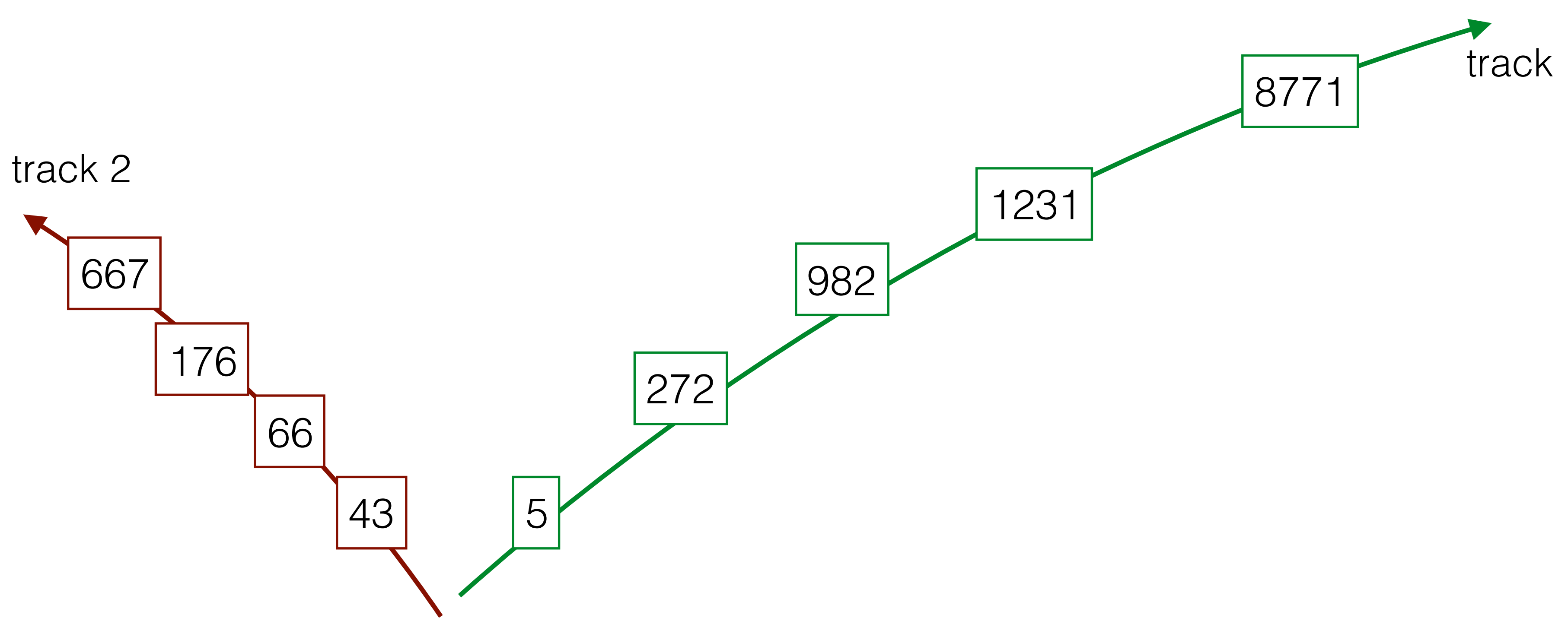
hit_id	track_id
5	1
272	1
982	1
1231	1
8771	1
43	2
66	2
176	2
667	2

overall_score =

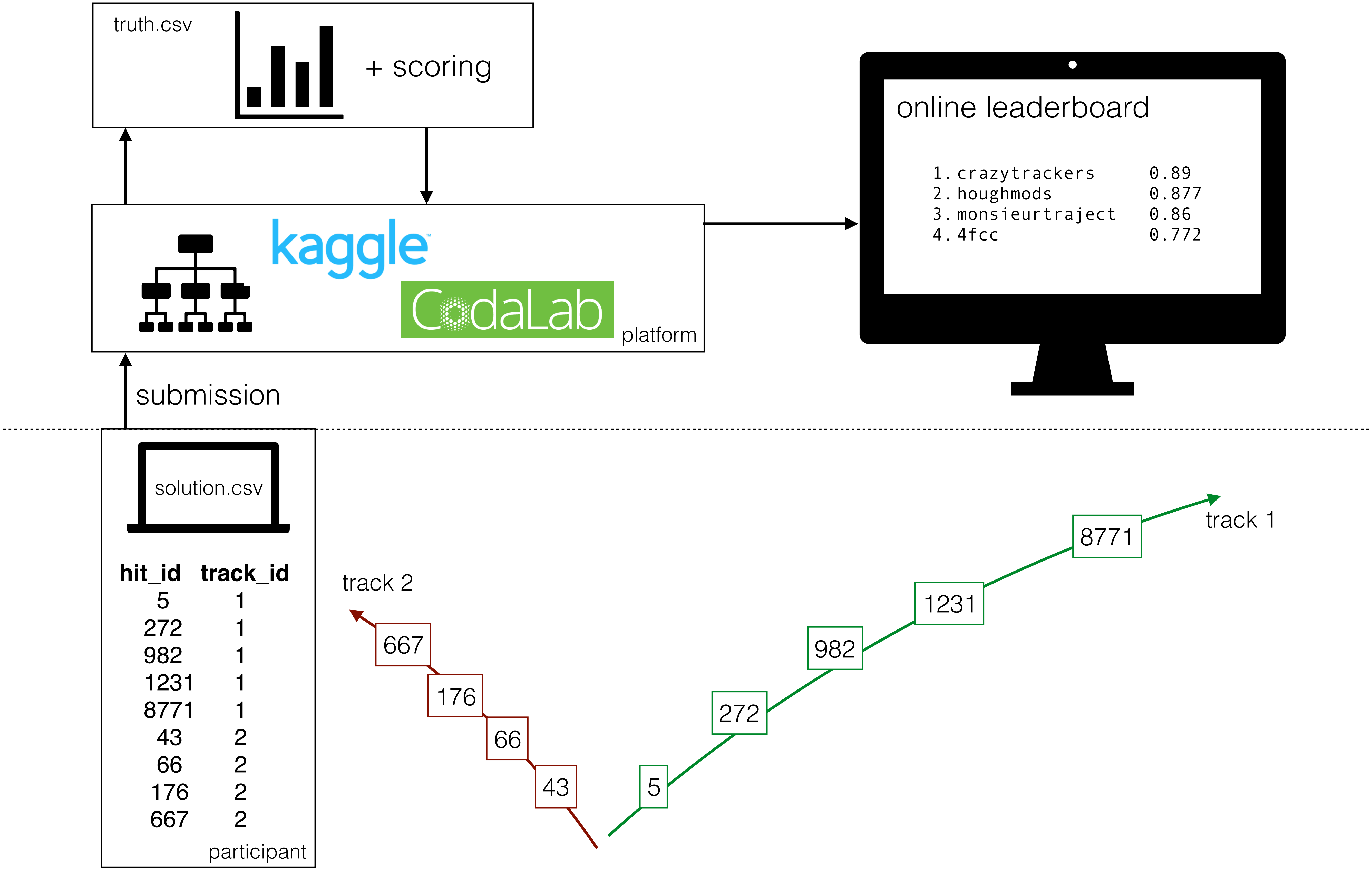
$$\sum_{\text{events}} \sum_{\text{tracks}} \texttt{track\_weight} * \texttt{track\_score}$$

higher momentum gives higher score:

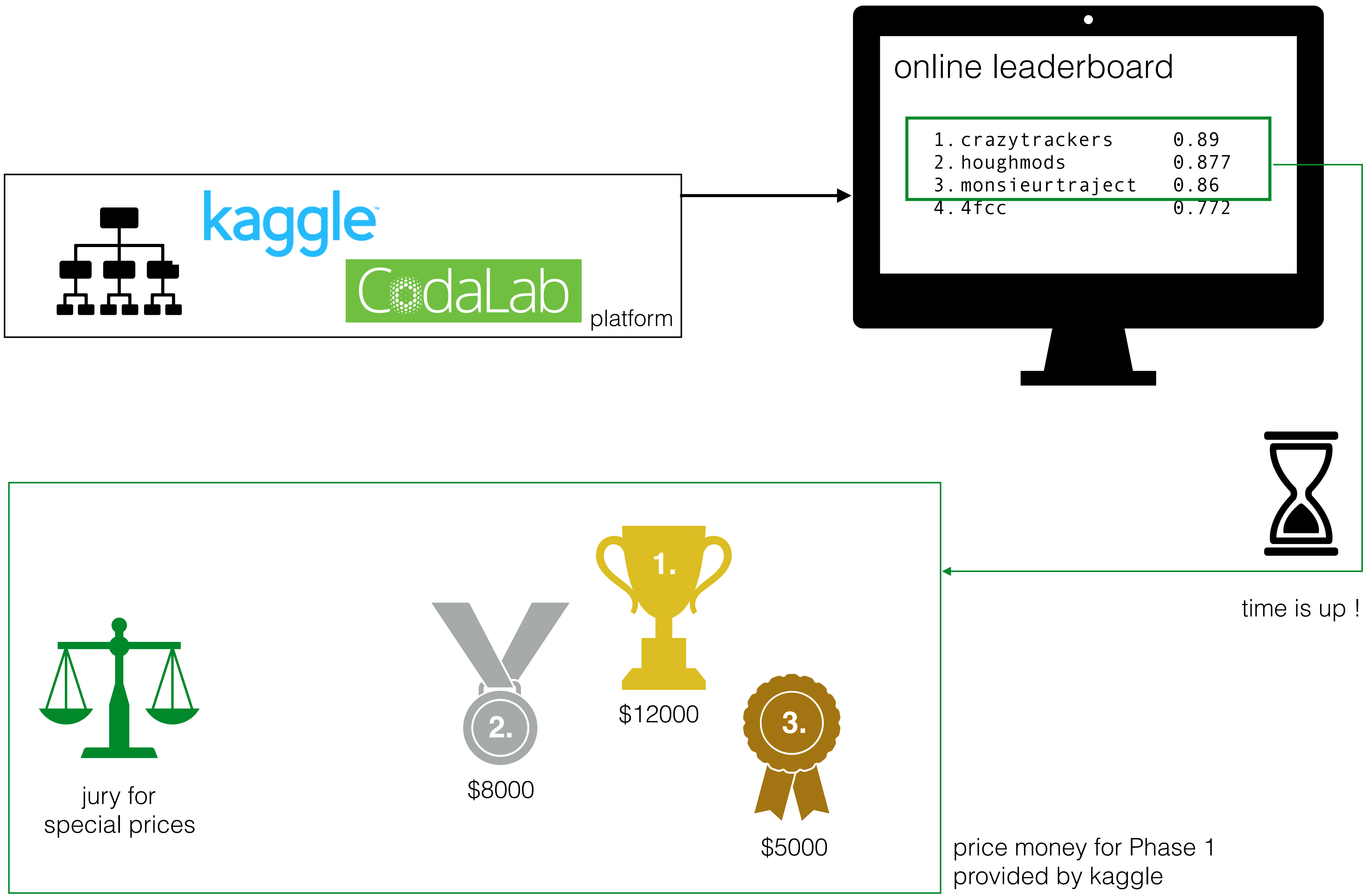




Submission & scoring

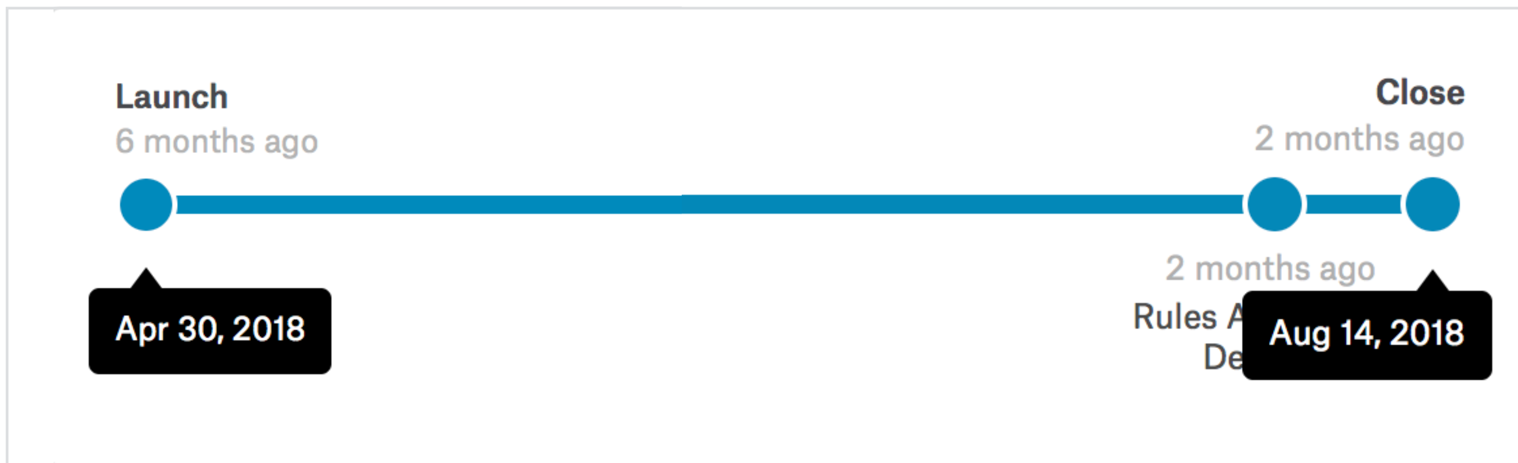


Winning

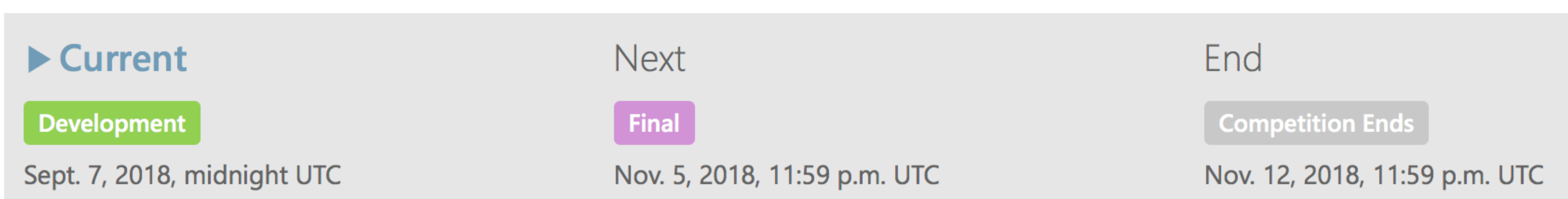

kaggle
CodaLab
platform
online leaderboard
1. crazytrackers 0.89
2. houghmods 0.877
3. monsieurtraject 0.86
4. 4fcc 0.772
time is up !
jury for
special prices
1.
$12000
2.
$8000
3.
$5000
price money for Phase 1
provided by kaggle

The challenge in 2 phases

Phase 1: **accuracy phase**



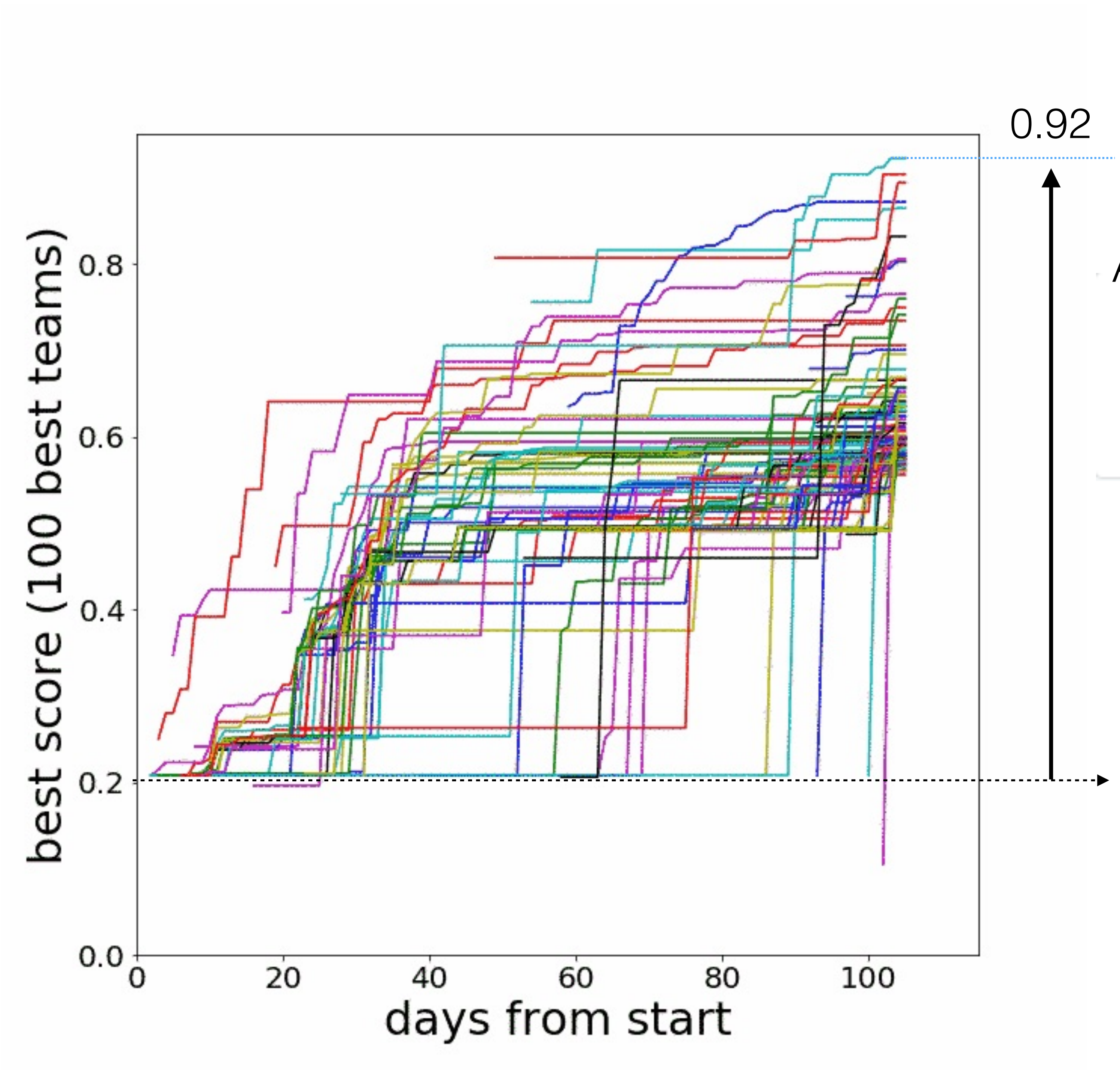
Phase 2: **throughput phase**



Phase 1 Accuracy

kaggle

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 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

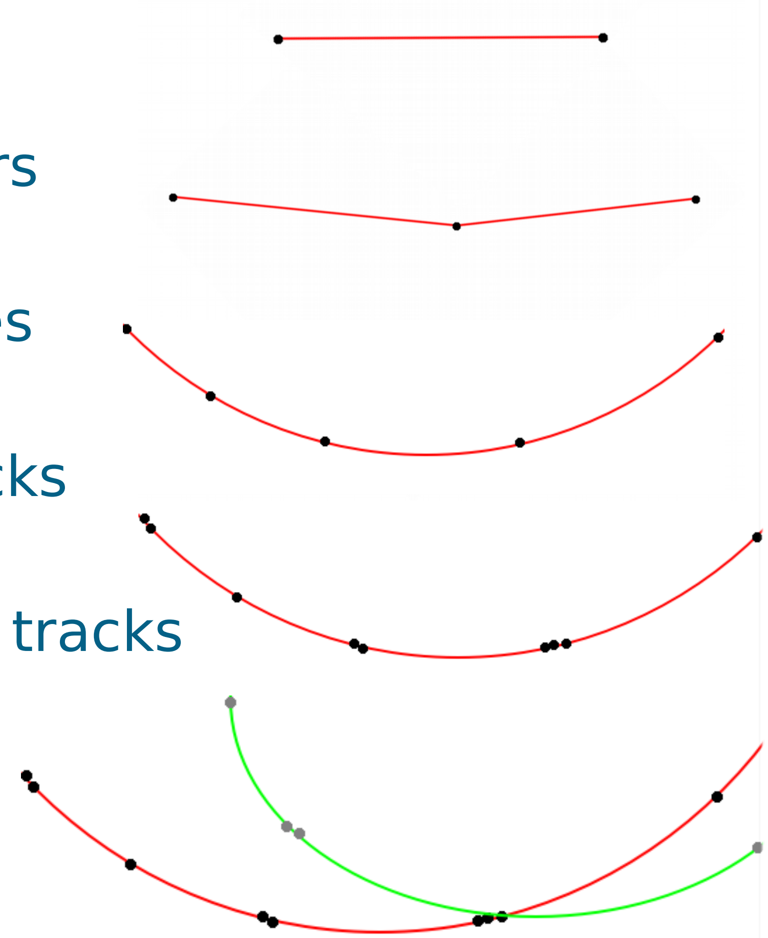
Phase 1 Top Quarks

Author: *J. S. Wind*

	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 outrunner

Author: *Pei-Lien Chou*

"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:

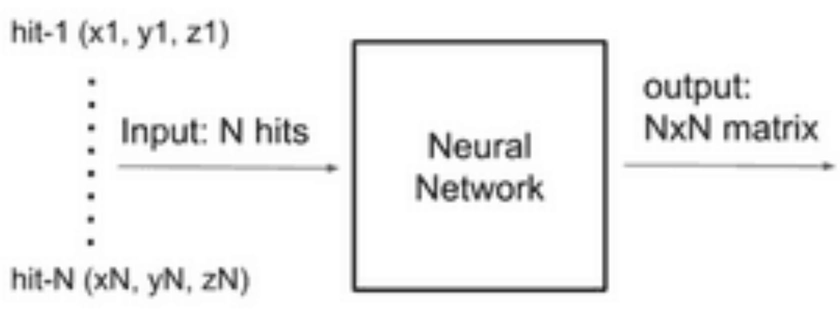
- predict relationship probability

Reconstruct

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

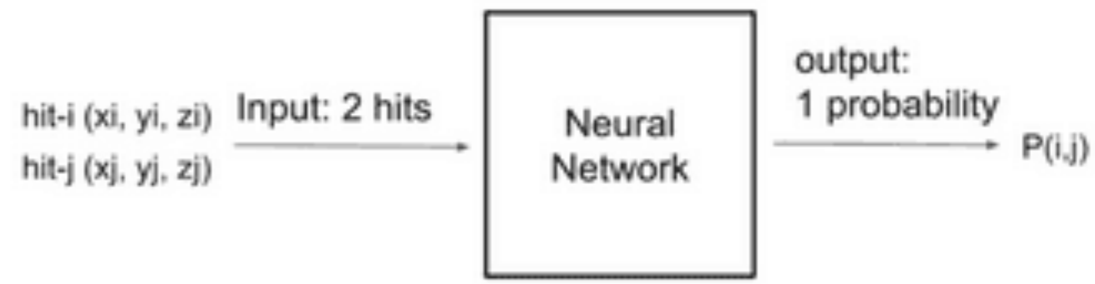
Phase 1 outrunner

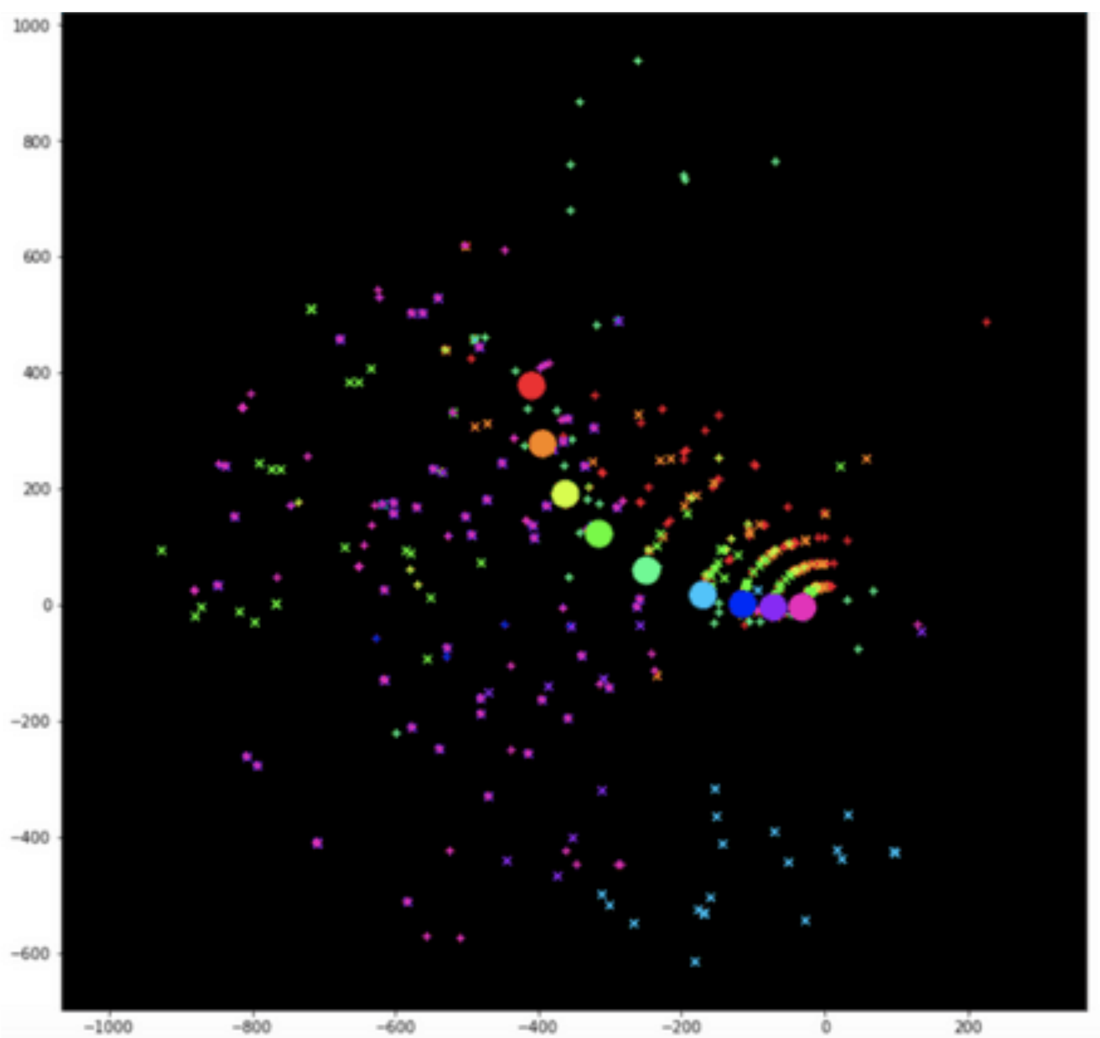
Training

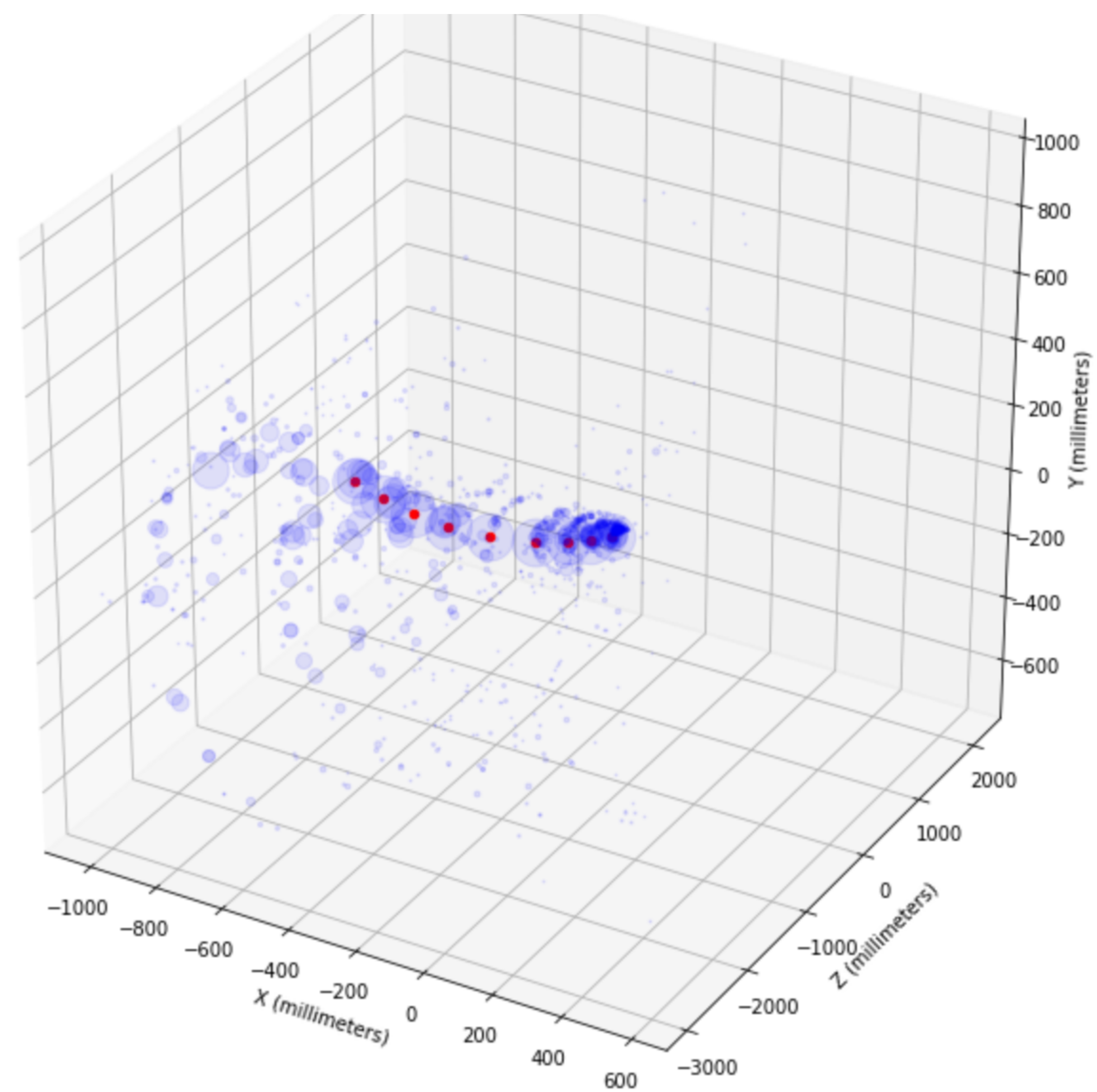


	1	2	3	4	5	~	N
1	0	1	0	0	1	~	1
2	1	0	0	0	1	~	1
3	0	0	0	1	0	~	0
4	0	0	1	0	0	~	0
5	1	1	0	0	0	~	1
~	~	~	~	~	~	~	~
N	1	1	0	0	1	~	0

Prediction







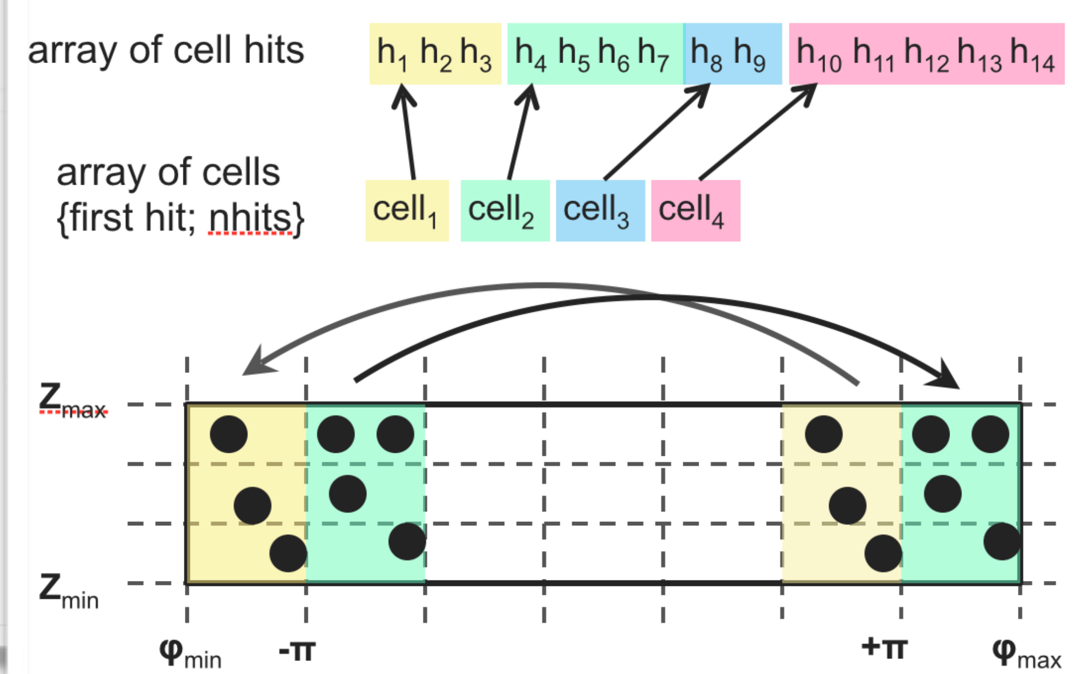
Phase 1 3rd place

Author: *Sergey Gobrunov*

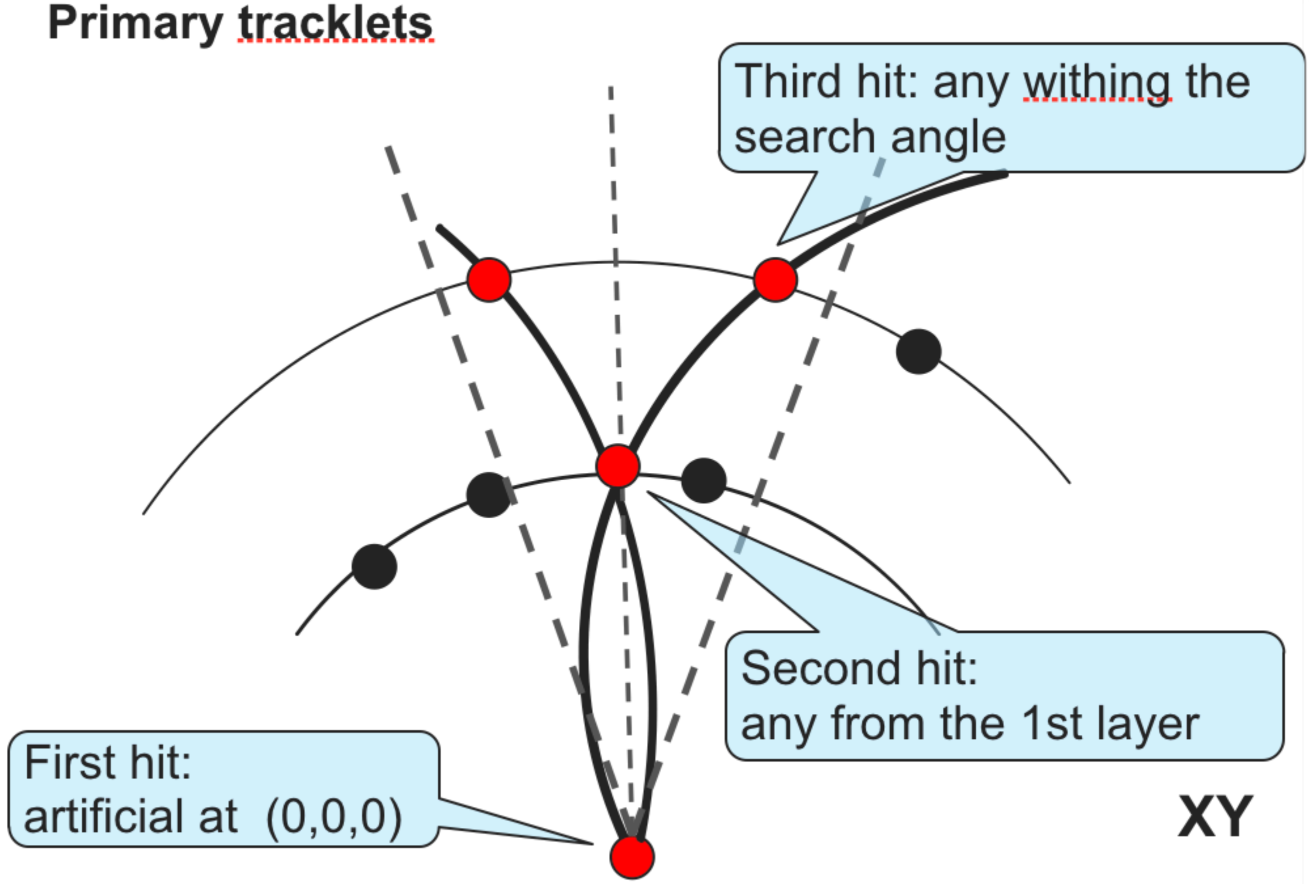
Execution time
1.2 min on single core 2.6 GHz CPU

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

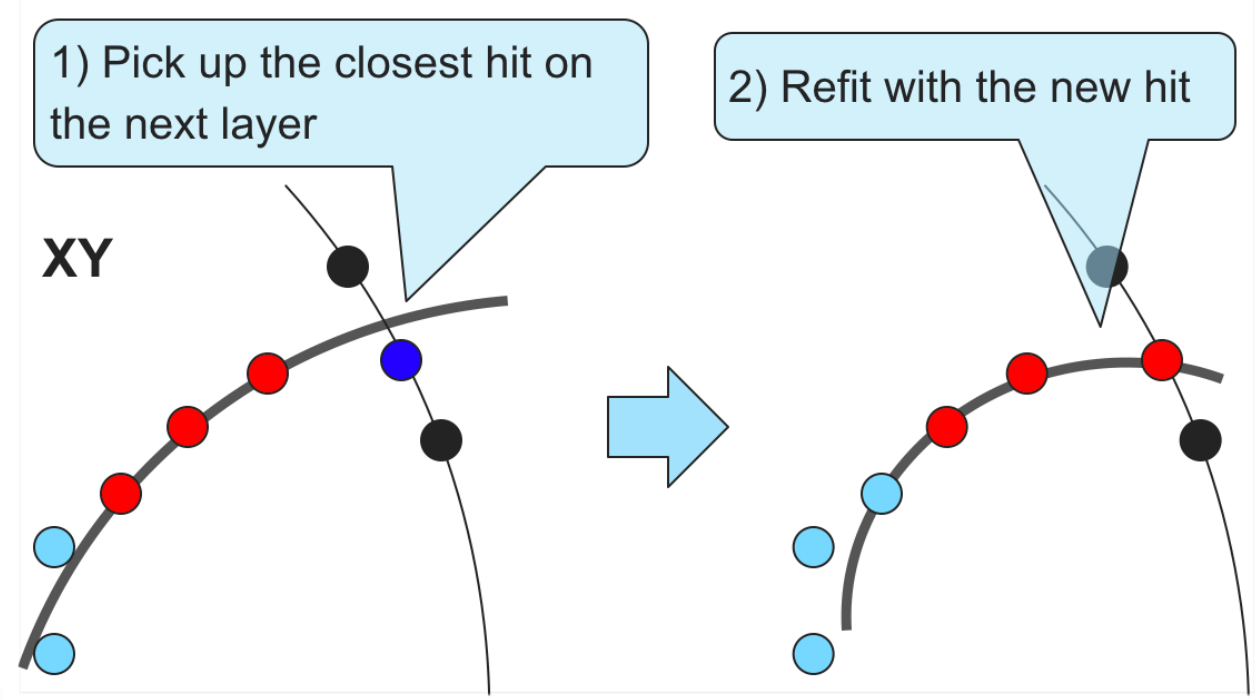
Regular grid with overlaps



Primary tracklets



Prolongation of tracklets



Phase 1 Jury prices

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

Deep Learning prize

Diogo R. Ferreira
Innovative pattern matching

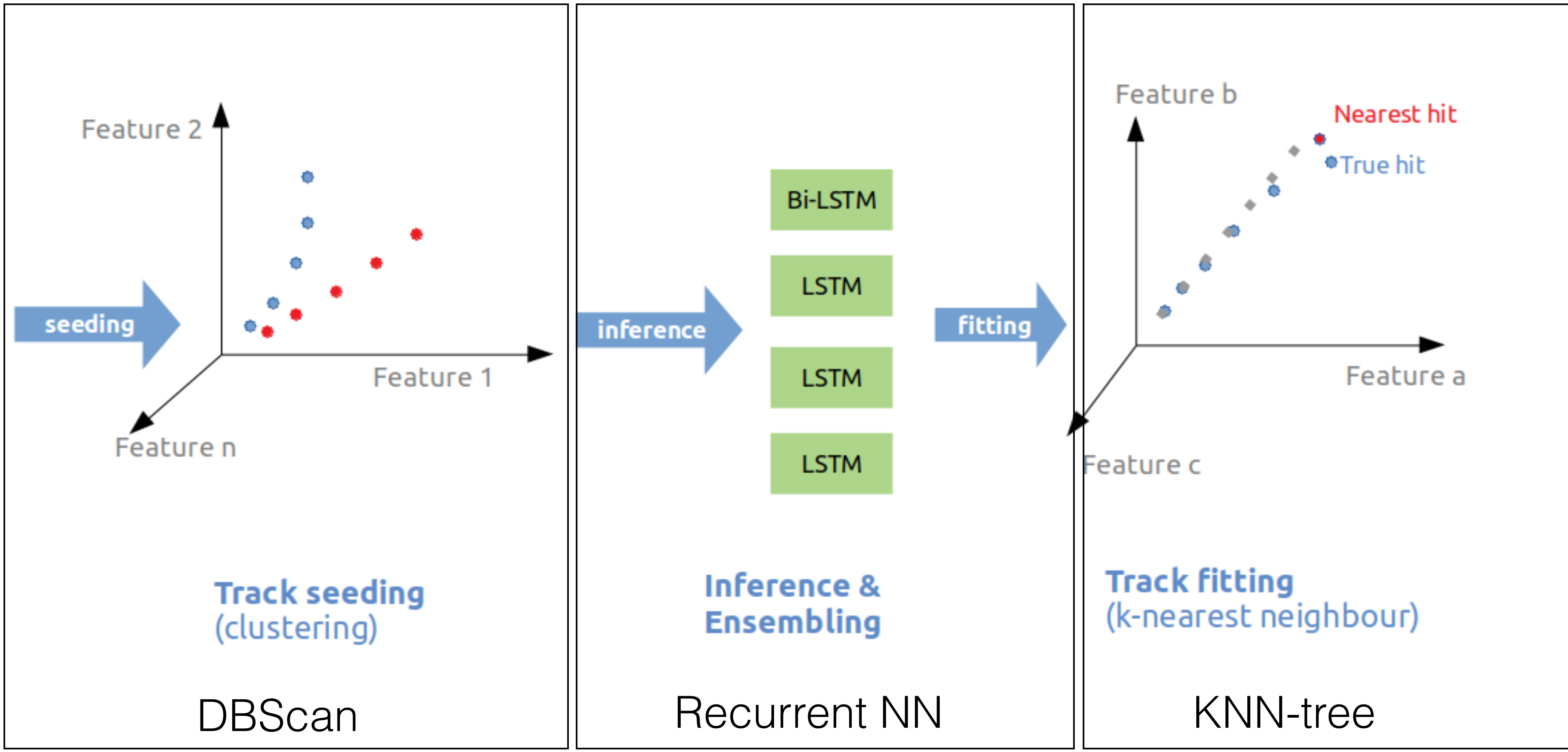
In the money Gold Silver Bronze

#	△pub	Team Name	Kernel	Team Members	Score
1	—	Top Quarks			0.92182
2	—	outrunner	In the money		0.90302
3	—	Sergey Gorbunov			0.89353
4	—	demelian			0.87079
5	—	Edwin Steiner			0.86395
6	—	Komaki			0.83127
7	—	Yuval & Trian	Jury pick		0.80414
8	—	bestfitting			0.80341
9	—	DBSCAN forever	Jury pick		0.80114
10	—	Zidmie & KhaVo			0.76320
11	—	Andrea Lonza			0.75845
12	—	Finnies	Jury pick		0.74827
13	—	Rei Matsuzaki			0.74035
14	—	Mickey			0.73217
15	—	Vicens Gaitan			0.70429
16	—	Robert			0.69955
100	▲2	Diogo	Jury pick		0.55480

Phase 1 Deep Learning Prize

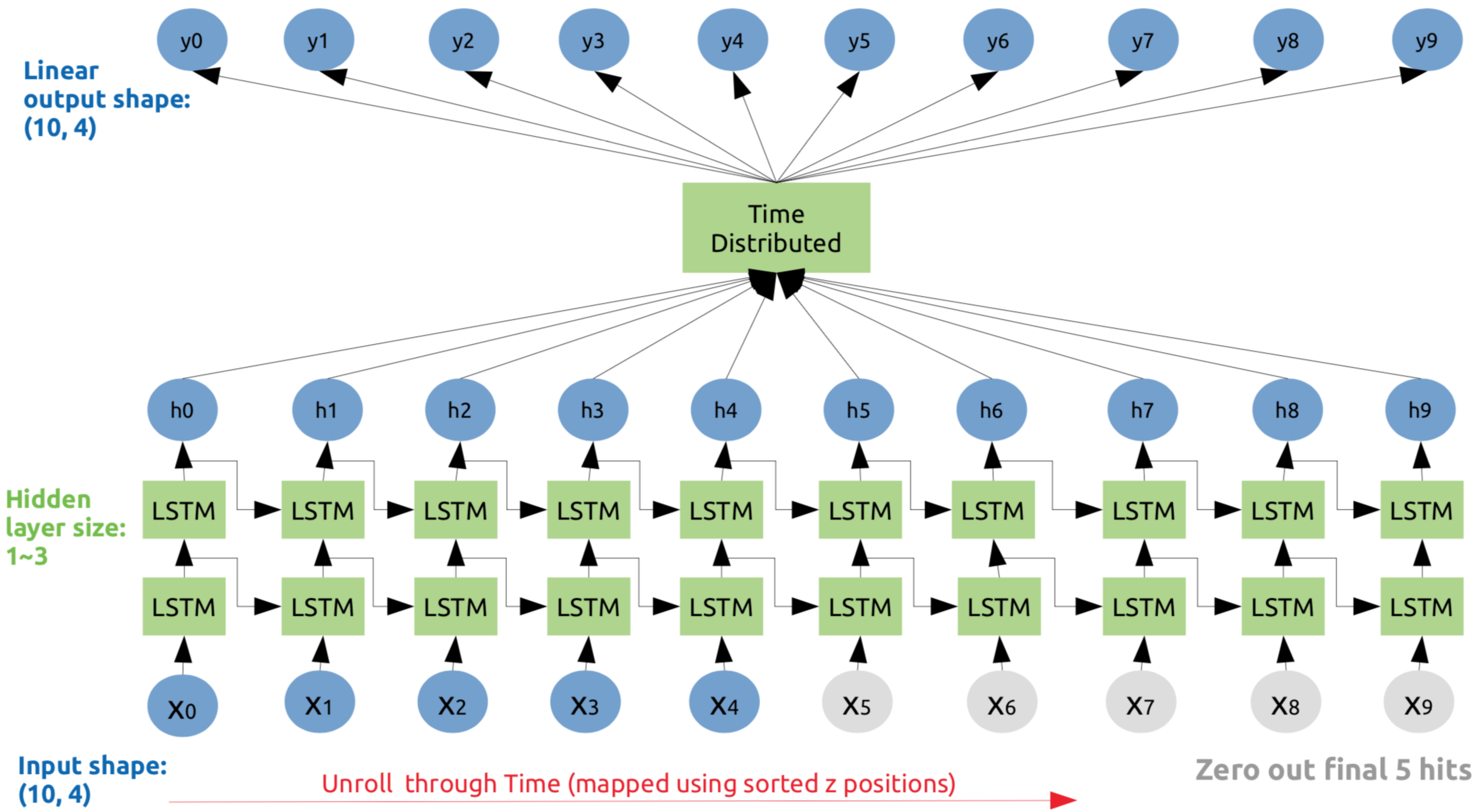
Author: *Nicole and Liam Finnie*

Three step approach



Phase 1 Deep Learning Prize

10 output sets of hit quadruplets



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

5 “empty” sets of hit quadruplets

Phase 1 Organizer's prize

Author: ***Diogo R. Ferreira***

Algorithm outline

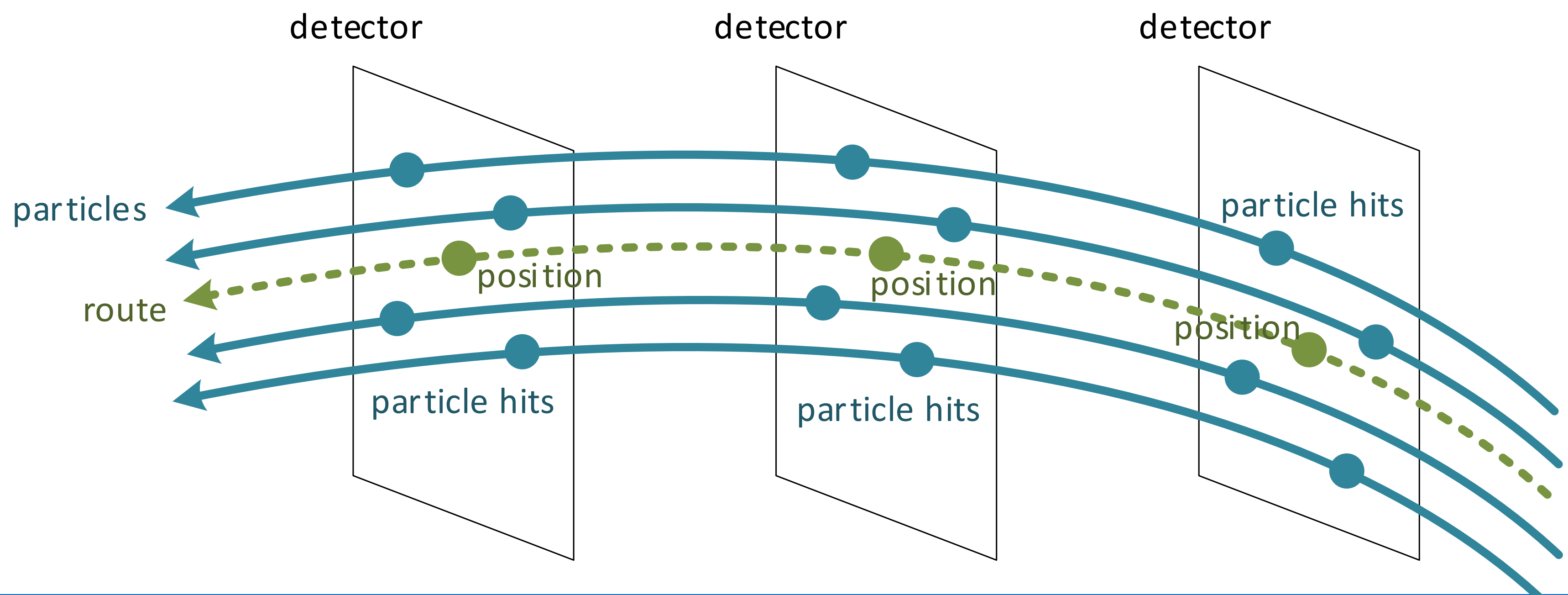
- First step is a route data bank building

  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



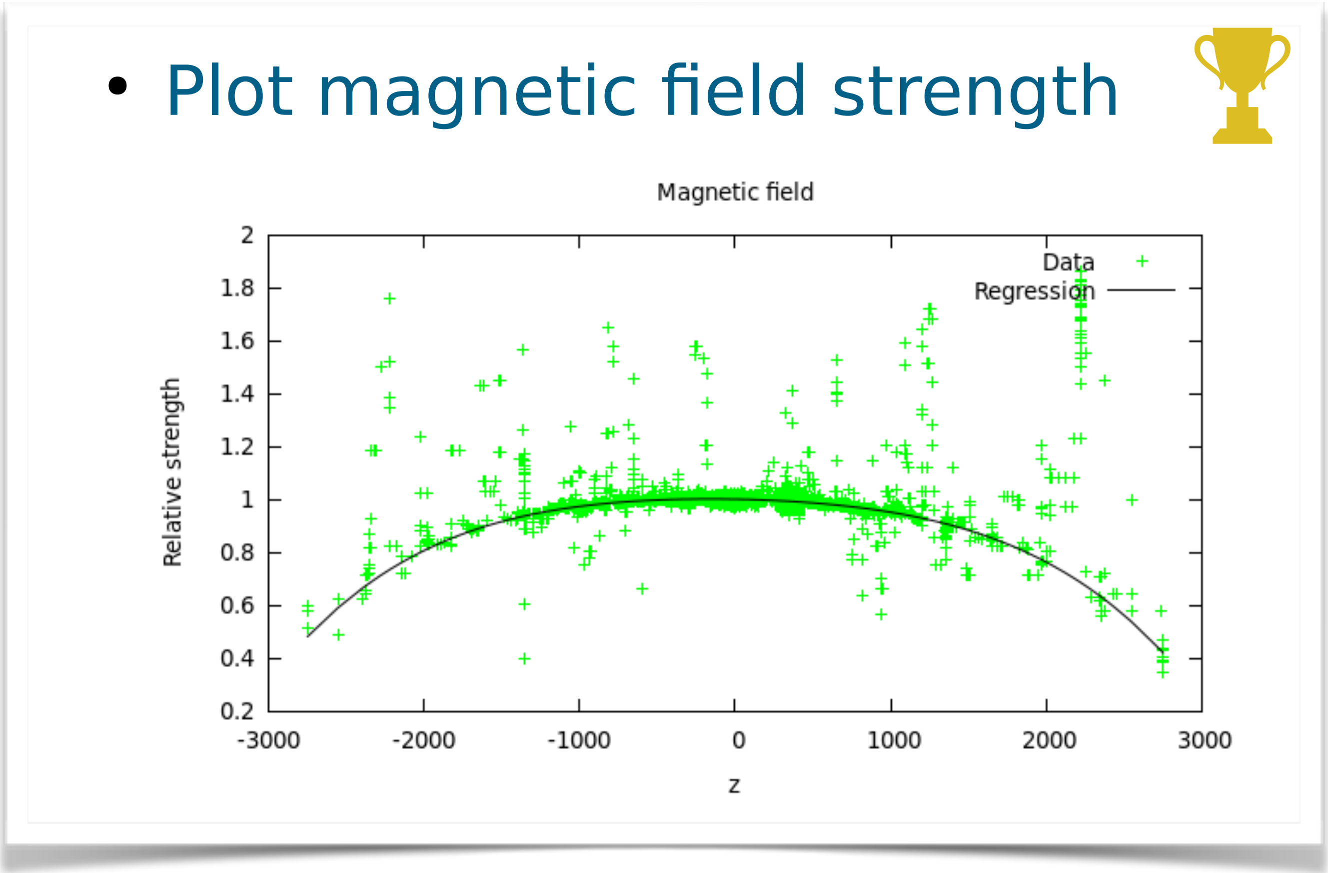
Phase 1 Some lessons learned

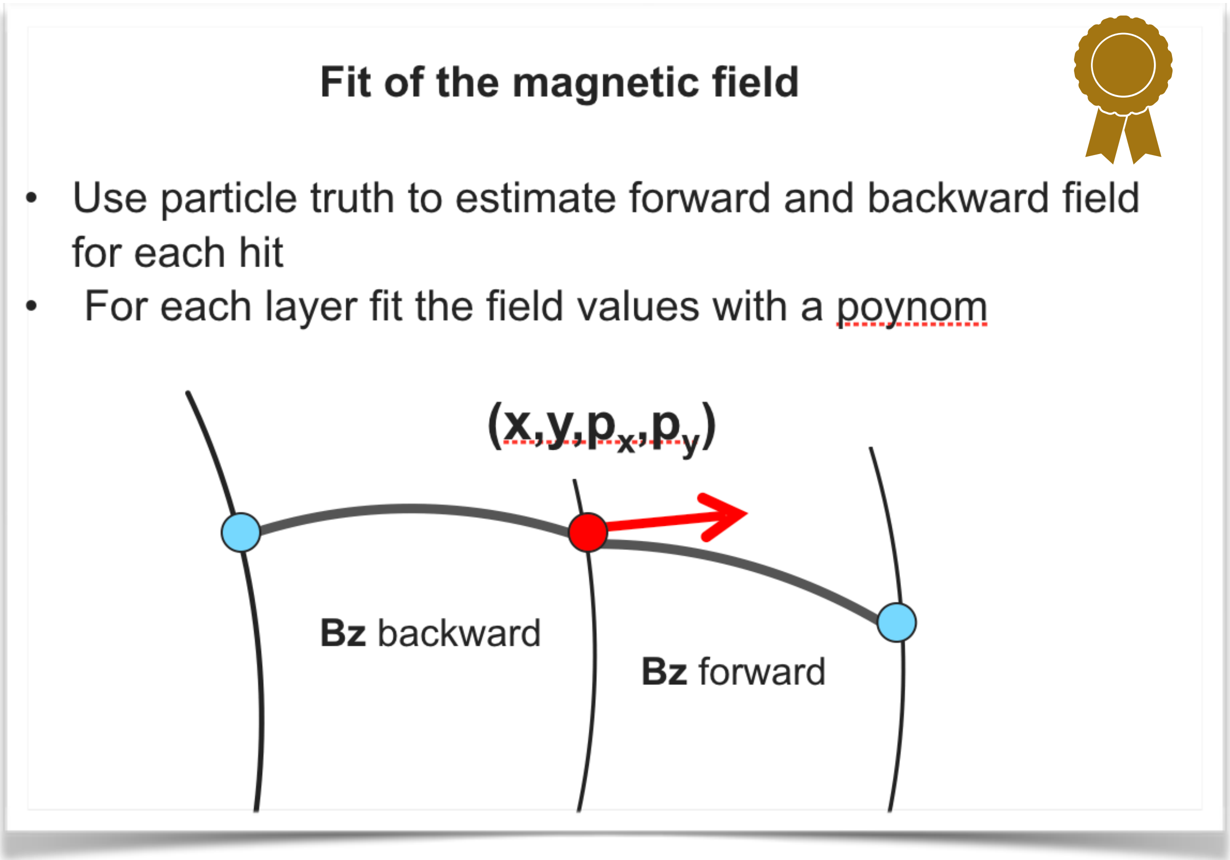
The threshold has been scary for some

- 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





Phase 1 Some lessons learned

The threshold has been scary for some

- 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

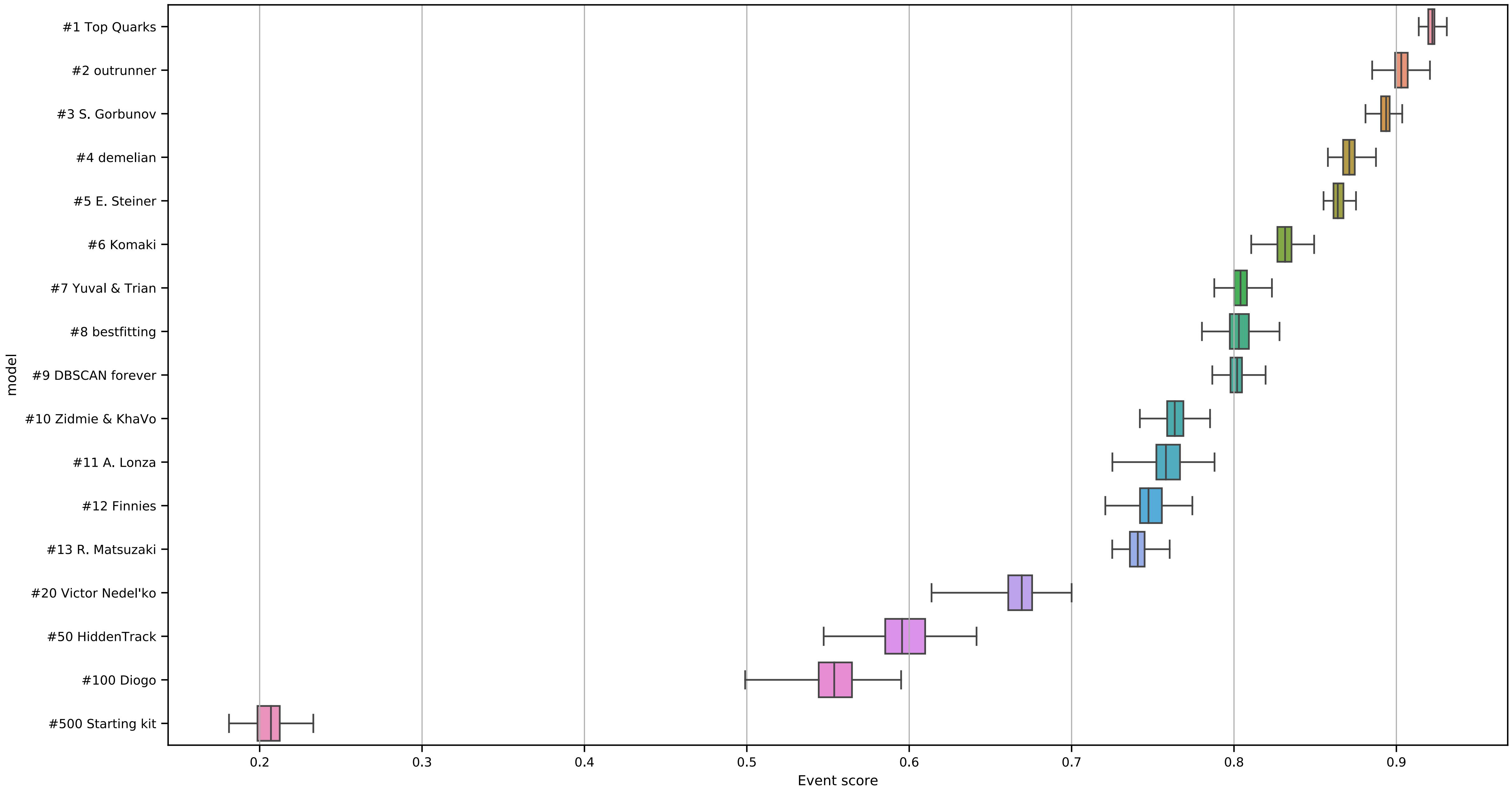
- Very good slides for beginners
- Lecture of particles tracking
- Full helix equations for ATLAS - All equations you need!
- Diplom thesis of Andreas Salzburger (Wow, he started in this field as a CERN student already in 2001 :p)
- Doctor thesis of Andreas Salzburger
- CERN tracking software Acts - 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!!

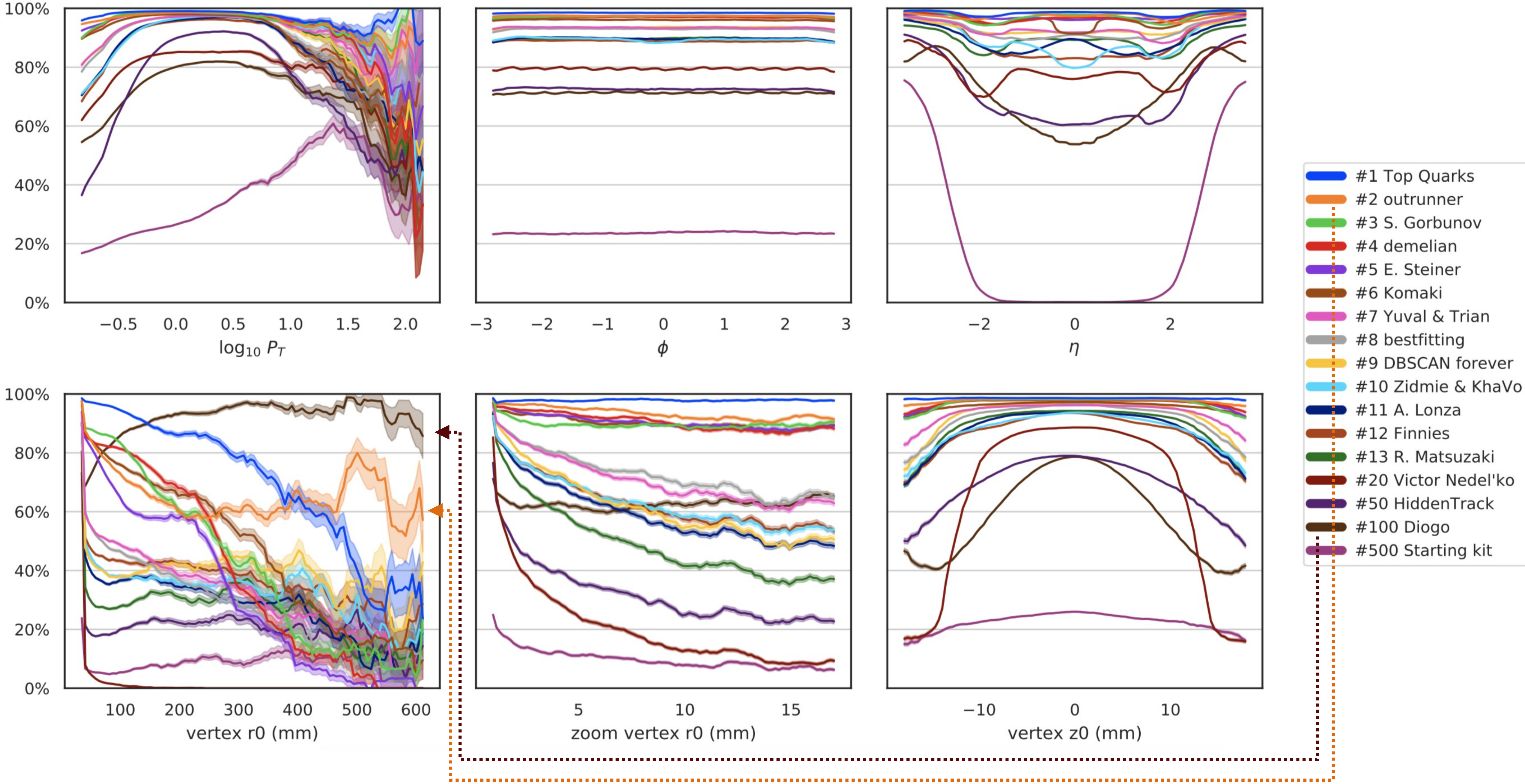
Phase 1 Aftermath Score stability



Score quartiles and extrema of the submitted solutions.

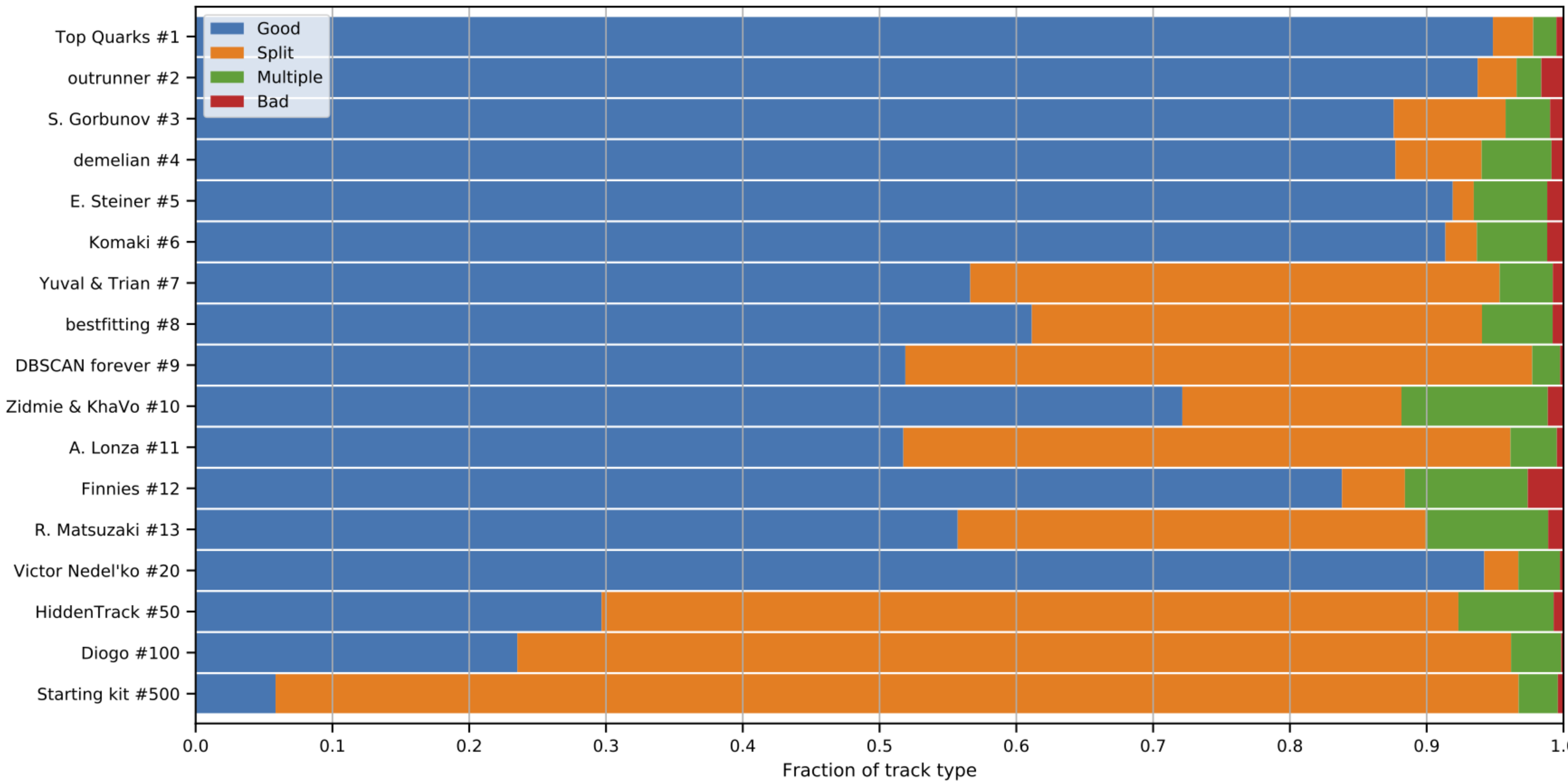
Summary of Phase-1 submitted as NeurIPS2018 Competition book.

Phase 1 Aftermath Tracking efficiency



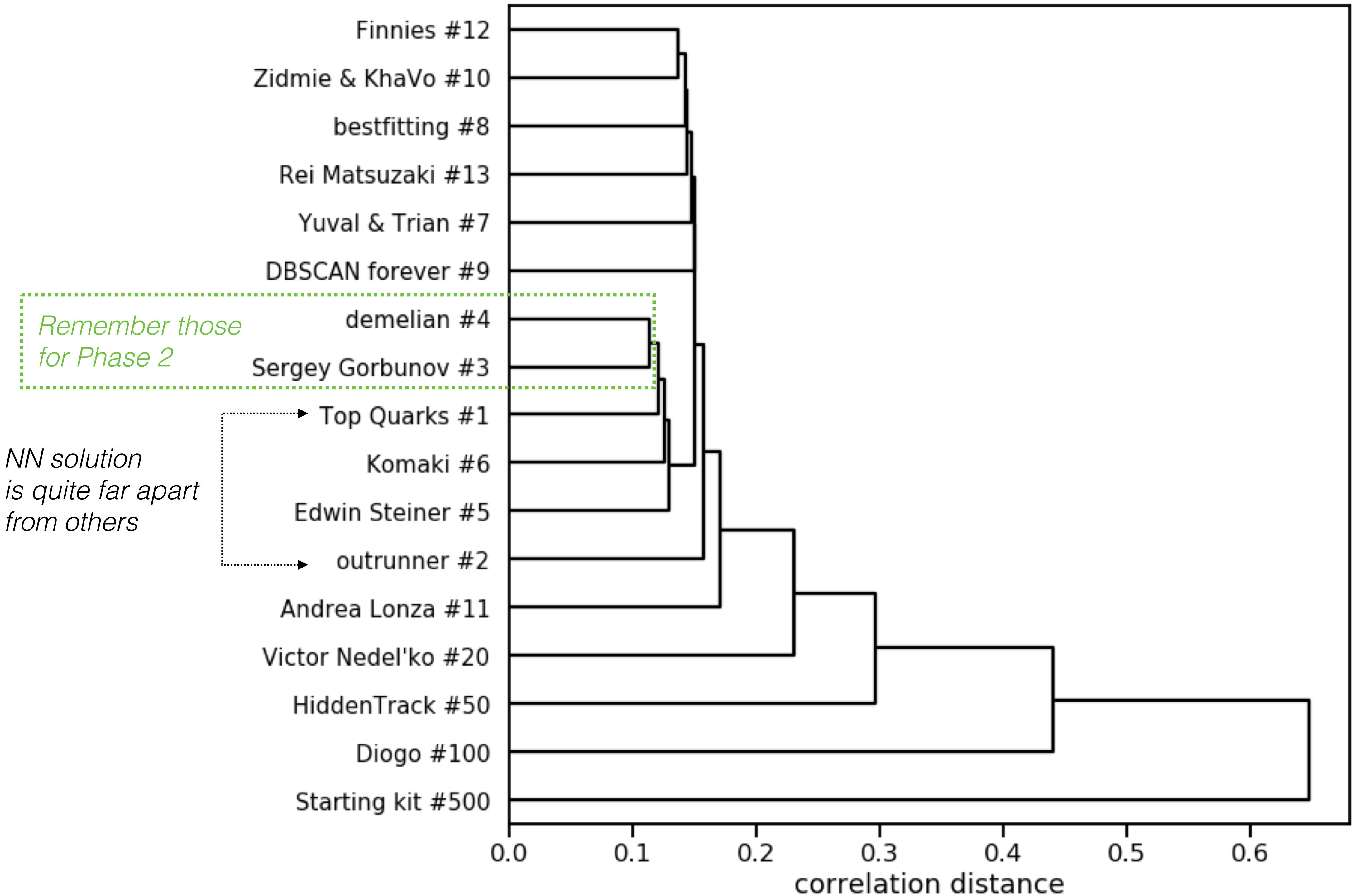
Efficiency correlates very strongly with score … good!

Phase 1 Aftermath Track types



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 Solution correlation



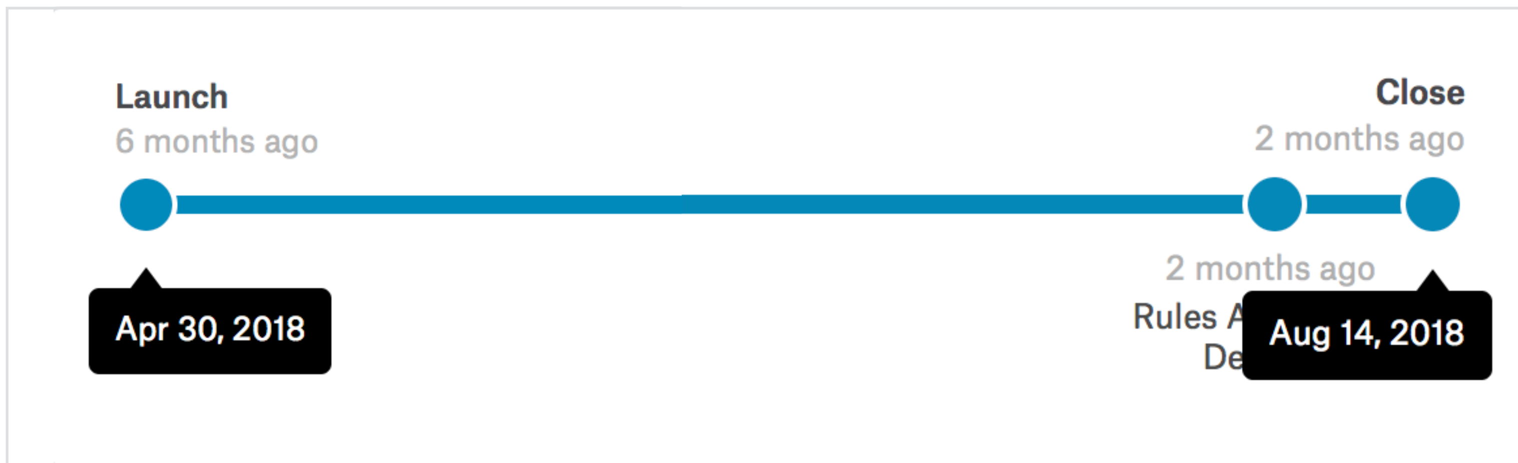
Phase 2 Throughput

CodaLab

The challenge in 2 phases

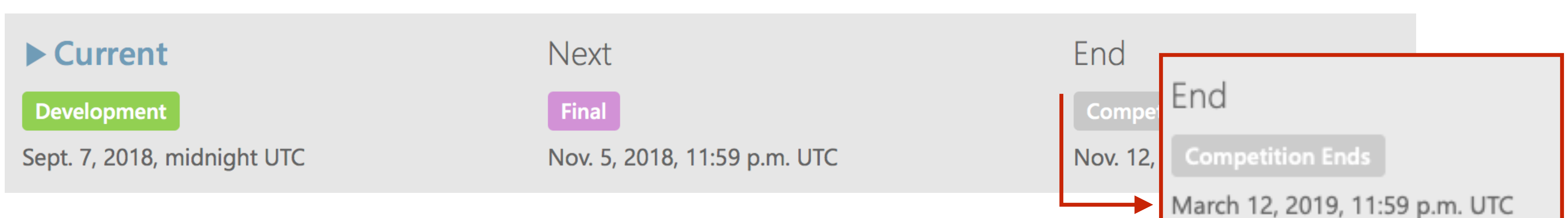
Phase 1: **accuracy phase**

kaggle



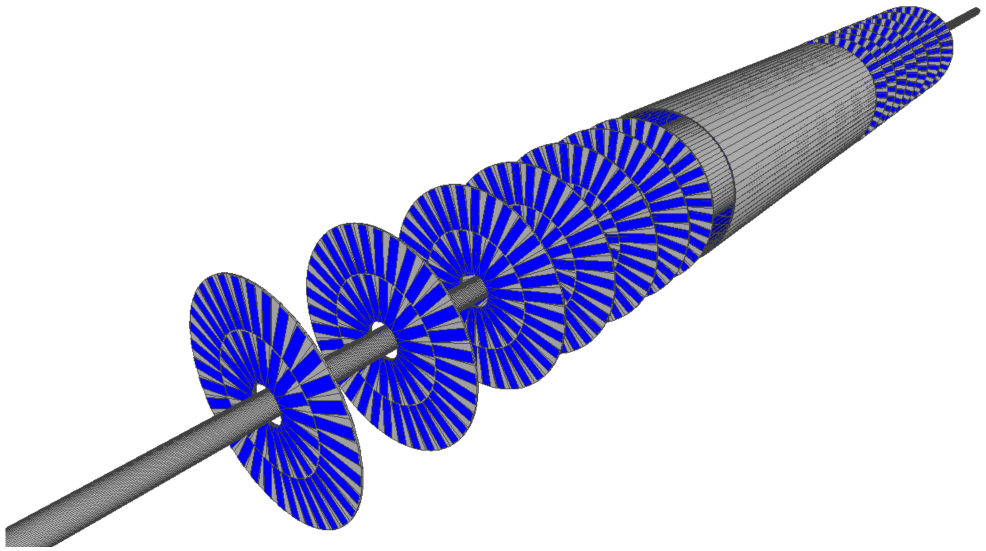
Phase 2: **throughput phase**

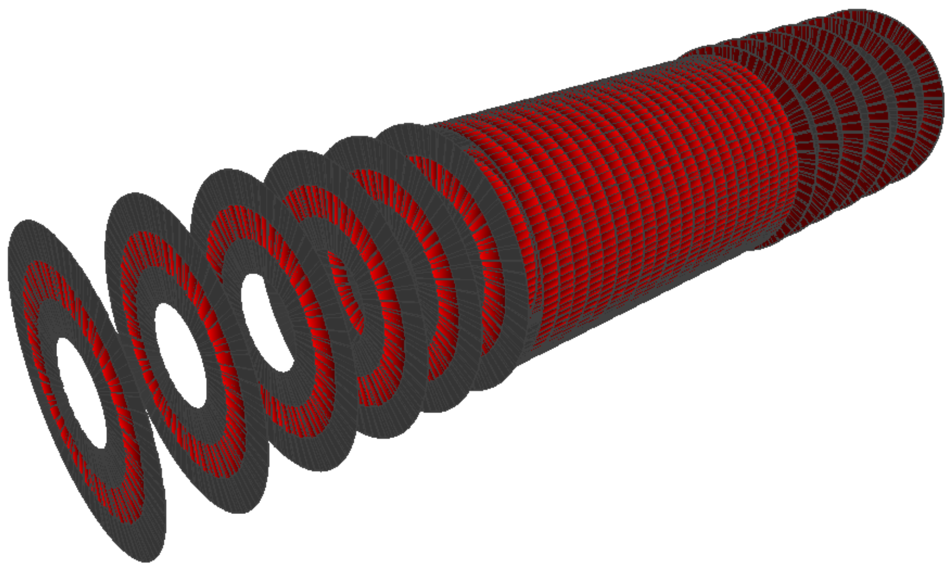
CodaLab

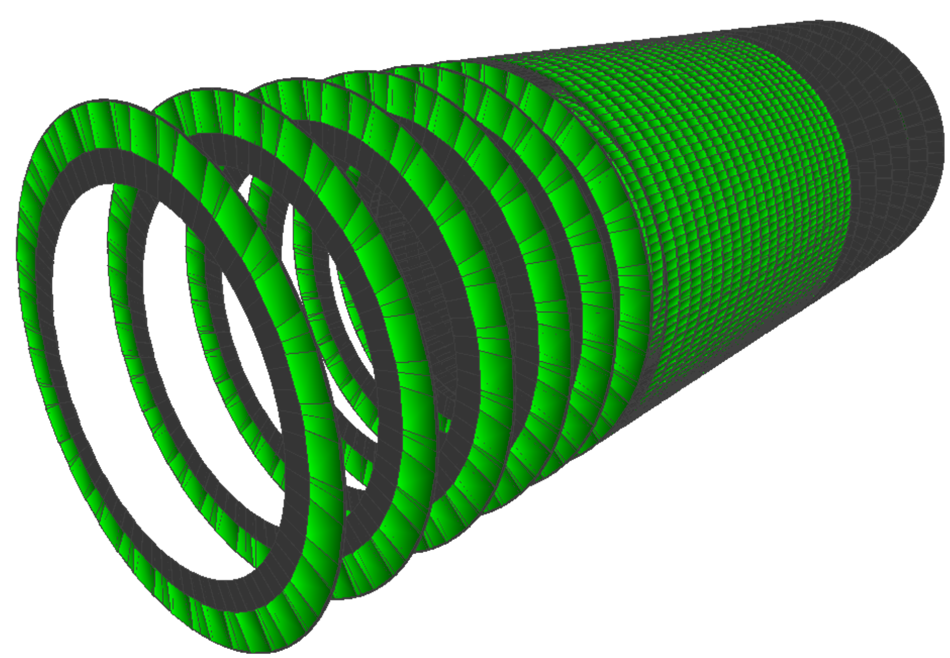


after initial very low participation & in agreement with contestants 4 months extra duration

Phase 2 Dataset







Detector remained unchanged

- served us well

Objective was slightly simplified

- only primary particles enter the scoring

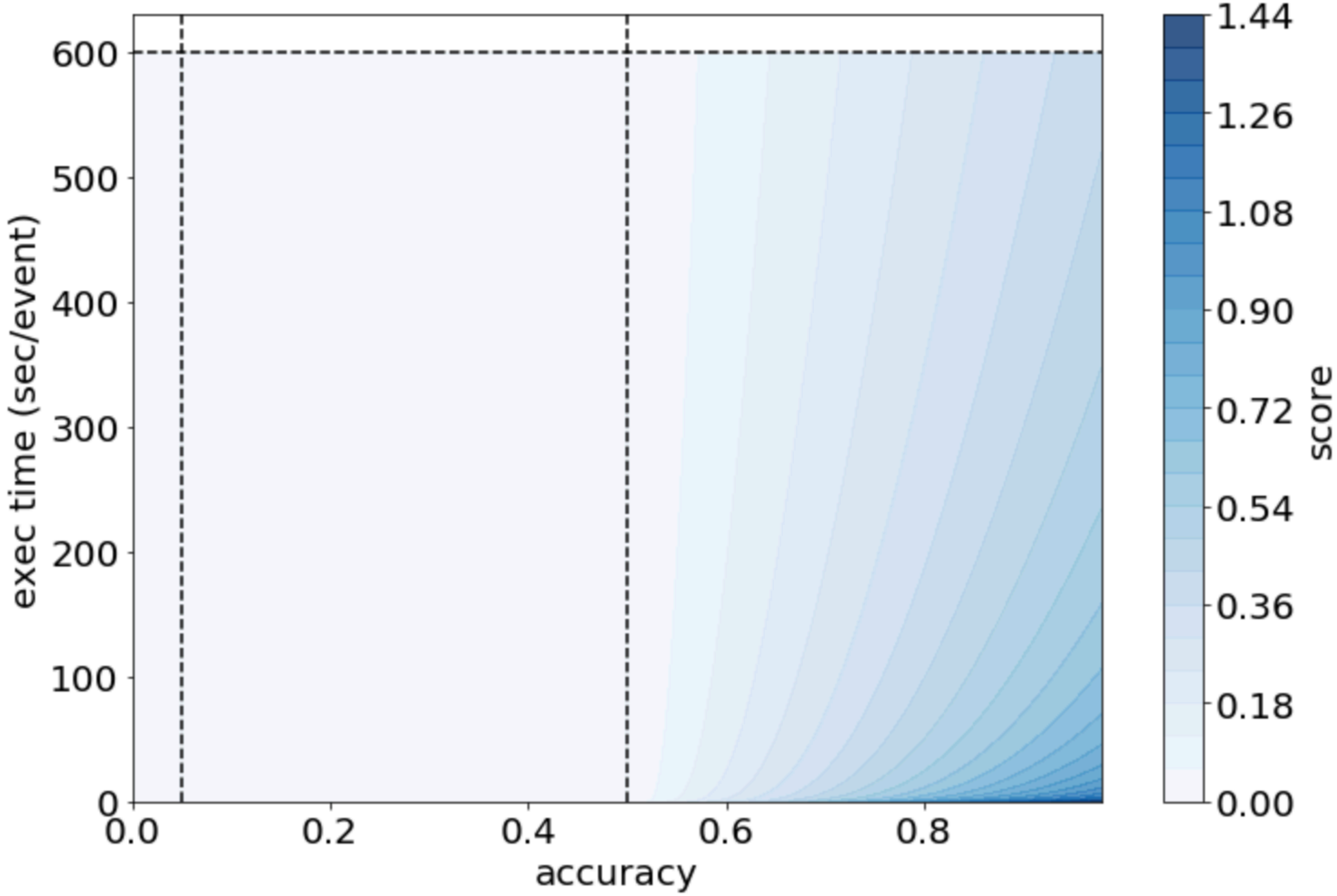
Some "features" have been fixed

- module thickness is corrected
  was wrong for cluster size evaluation
- too narrow beam spot in Phase 1
  corrected from s=5.5 mm to s=5.5 cm
- looping particles (present in Phase 1) have been removed
- overshooting scattering for electrons (0.5 % effect in 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 and with codalab)



Phase 2 Control of 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

start

API to call CPP PY

User executable

stop

solution

VM 2 cores, 4 Gb memory

Phase 2 Winners

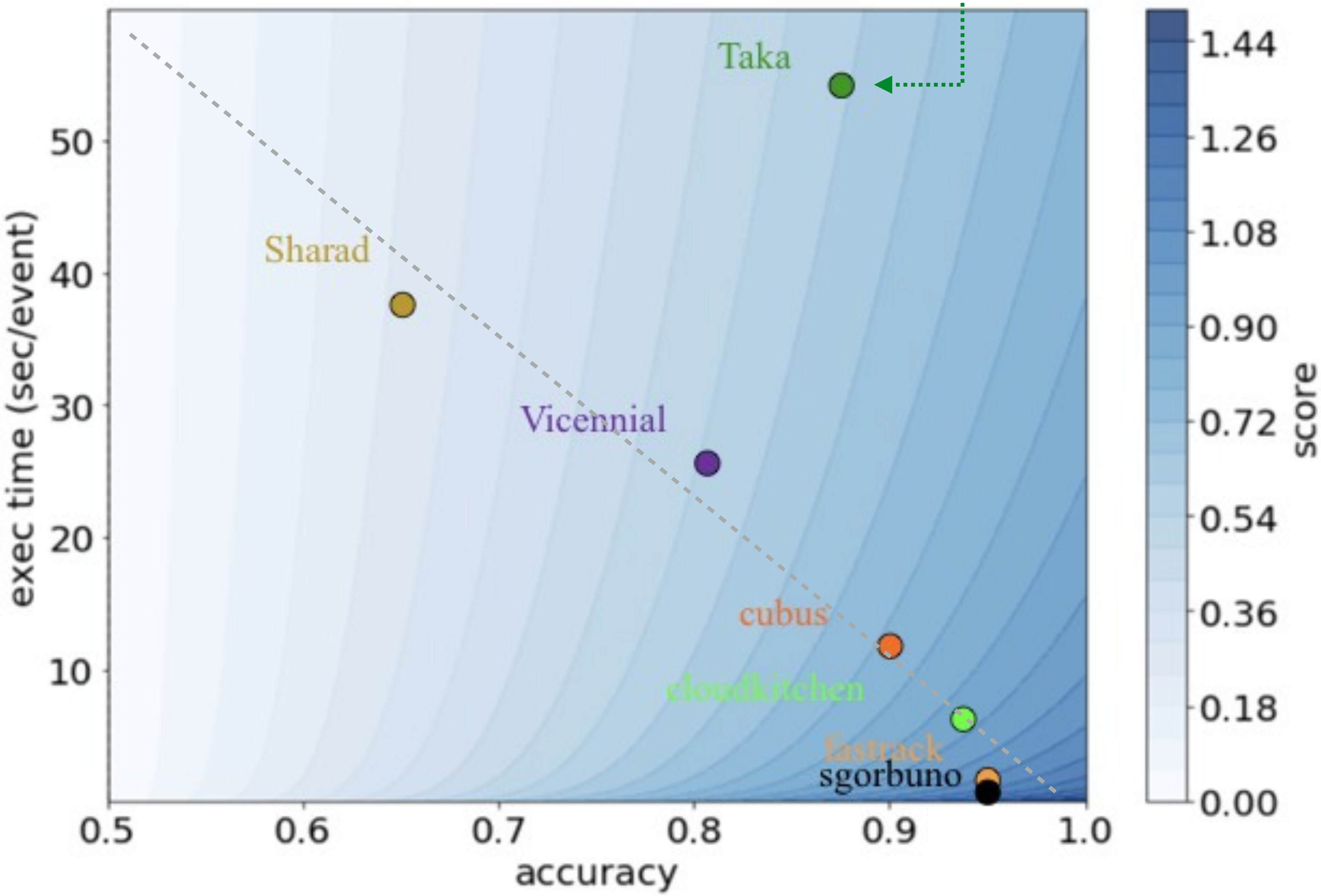
RESULTS									
#	User	Entries	Date of Last Entry	score ▲	accuracy_mean ▲	accuracy_std ▲	computation time (sec) ▲	computation speed (sec/event) ▲	Duration ▲
1	sgorbuno	9	03/12/19	1.1727 (1)	0.944 (2)	0.00 (14)	28.06 (1)	0.56 (1)	64.00 (1)
2	fastrack	53	03/12/19	1.1145 (2)	0.944 (1)	0.00 (15)	55.51 (16)	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 Resulting 2D scoring map

Impressive trend

- generally fastest solutions are also the best

  lesson from winner of Phase-1: *the faster, the more time to tune!*

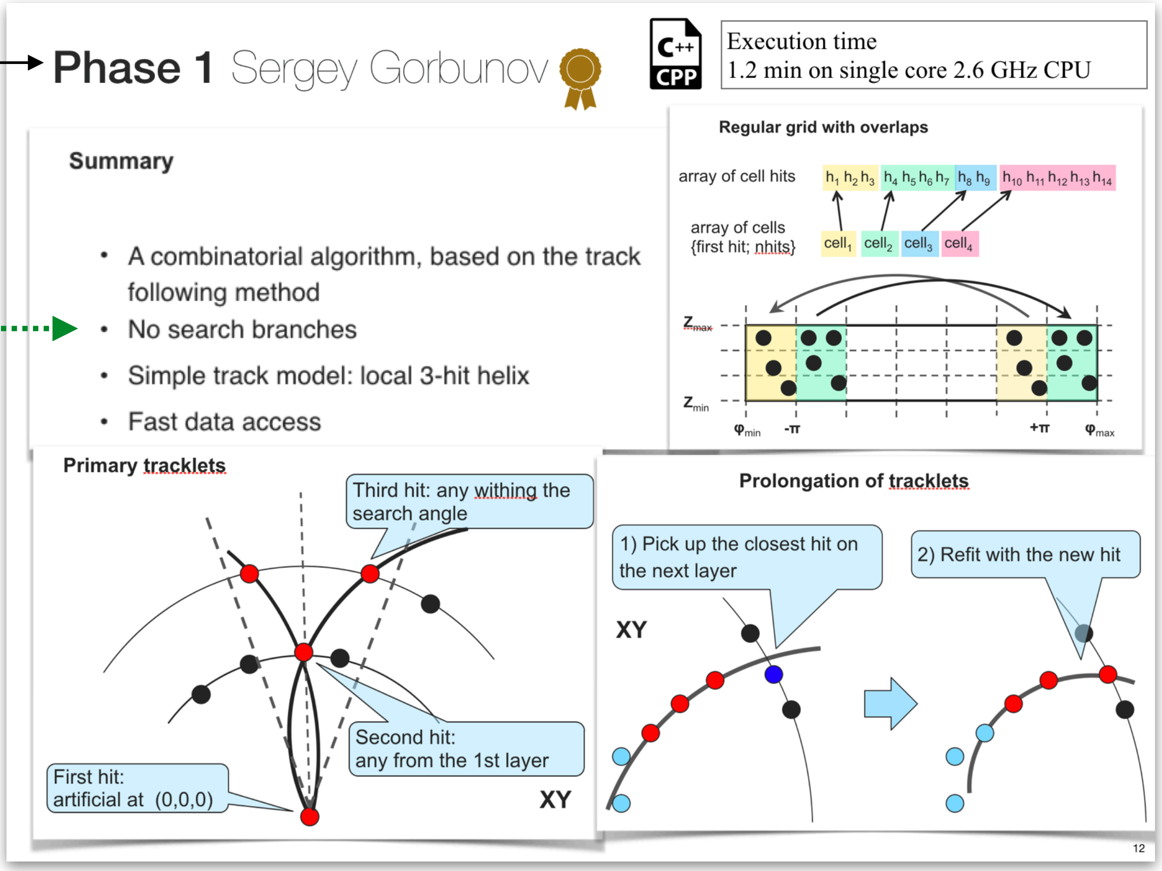


Phase 2 *Mikado*

Author: **Sergey Gorbunov**

Accuracy: 0.944
Time/event: 0.56 sec
Memory: 0.1/0.178 Gb (1core/2 cores)

third in Phase-1



Based on Phase-1 algorithm

- runs iteratively in 80 passes
  & hit removal from high to low pT
- modifications with respect to Phase 1
  search branches enabled
- every pass has optimised parameters
  results in $O(10^4)$ parameters to be tuned,
  tuning done semi-automated

Phase 2 *FASTrack*

Author: **Dmitry Emeliyanov**

+ OpenMP

Accuracy: 0.944
Time/event: 1.11 sec → 0.8 sec
Memory: 0.6 Gb
recently down to

first runner-up to podium in Phase-1

4	—	demelian	0.87079	35	2mo

Phase-1 w/o measurement shapes

Algorithm outline

- using measurement shapes to predict intervals of track inclination
- segment based track following network with embedded Kalman Filter
  - **connection graph** pre-build (&compiled) from `Detector.csv` file
  - run with a **Cellular Automaton (CA), parallelised** with **OpenMP**
  - **candidate building:** graph traversal with applied simplified KF
- combinatorial track following for track completion
  - fast **combinatorial** Kalman Filter using **3rd oder RK** & **simplified field** includes **clone identification** & **track merging**

3 passes (hit removal):
- high momentum
- low momentum
- rest

Phase 2 cloudkitchen

Author: **Marcel Kunze**

Accuracy: 0.93
Time/event: ~7 sec
Memory: 0.7 Gb

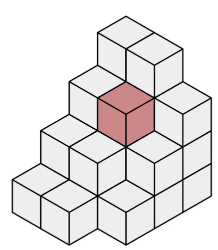
partly based on top quarks Phase 1 solution

1	—	**Top Quarks**	0.92182	10	2mo

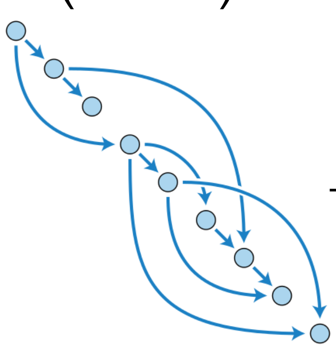
Algorithm outline

hits

sorted in voxels

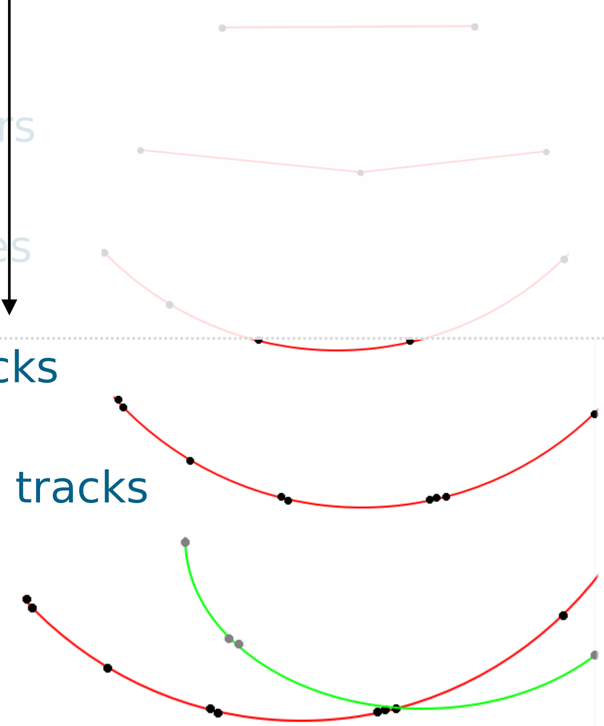


organised in
direct acyclic graphs
(DAG)



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



DAGs are pre-trained on ~25 events ground truth

DAGs are used to fast navigate through voxel space

$\pm z$ graph set

$\eta - \phi$ graph set

Triplet finder

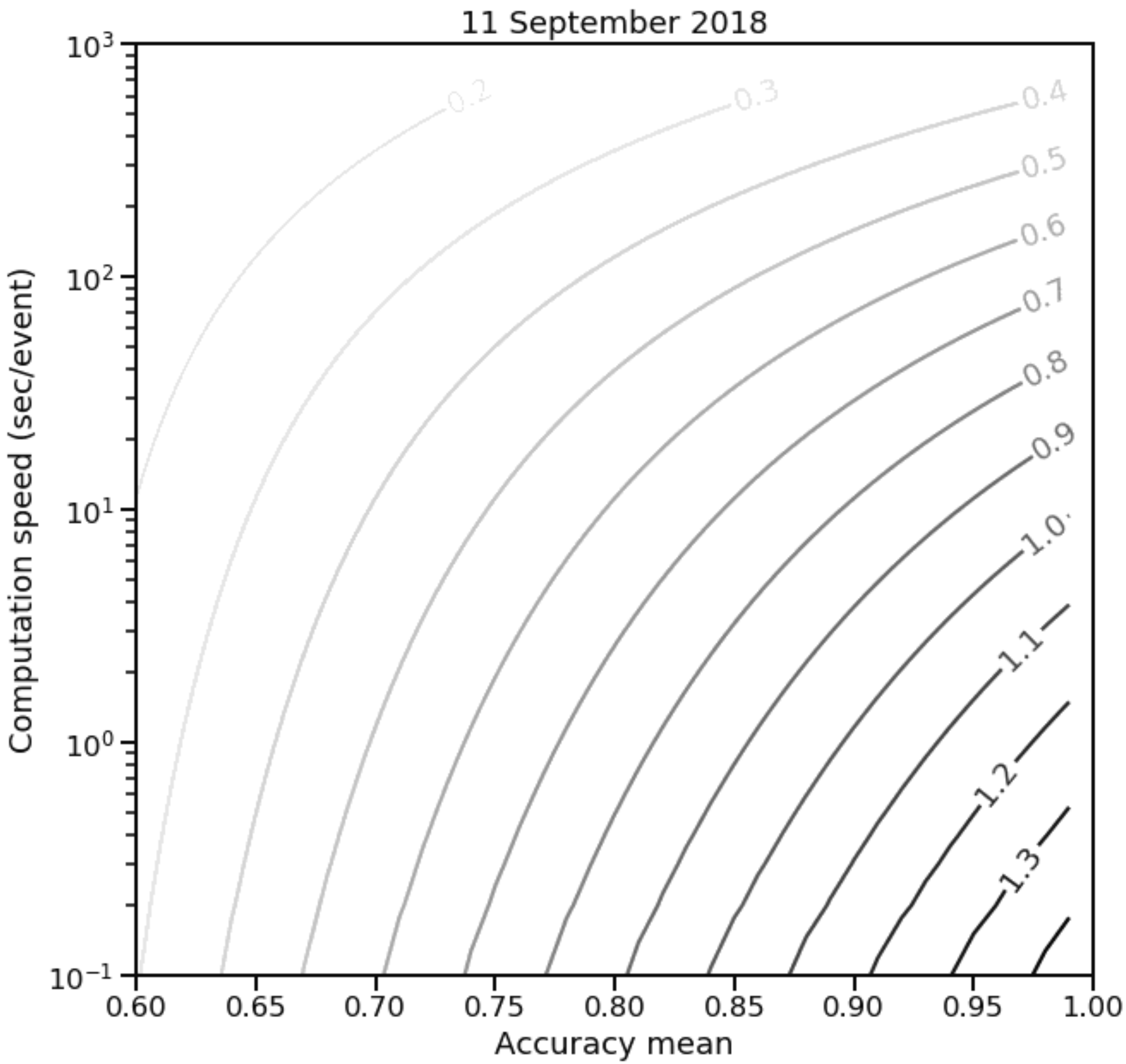
NN3

doublet finder

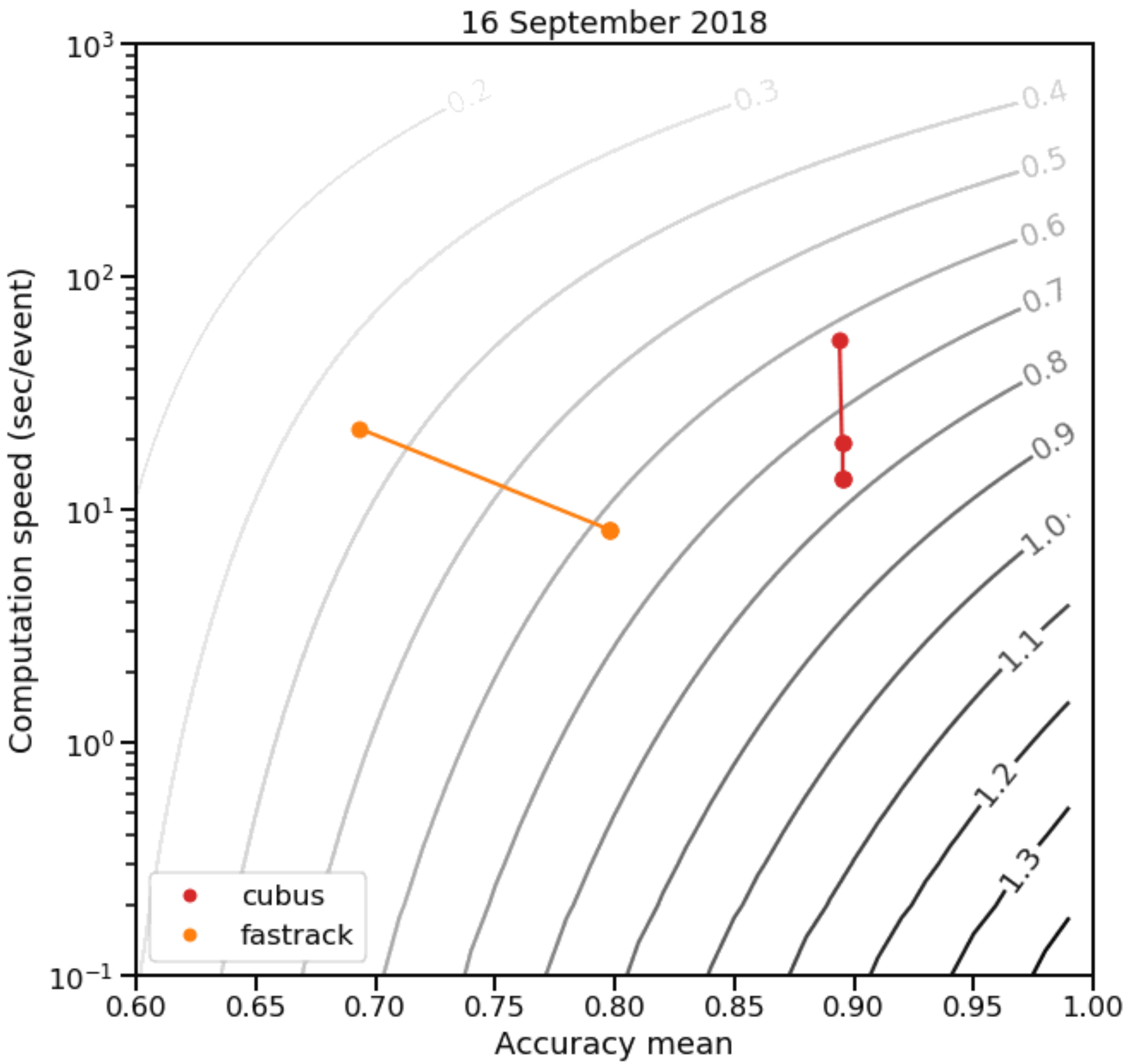
NN1 NN2

Threaded

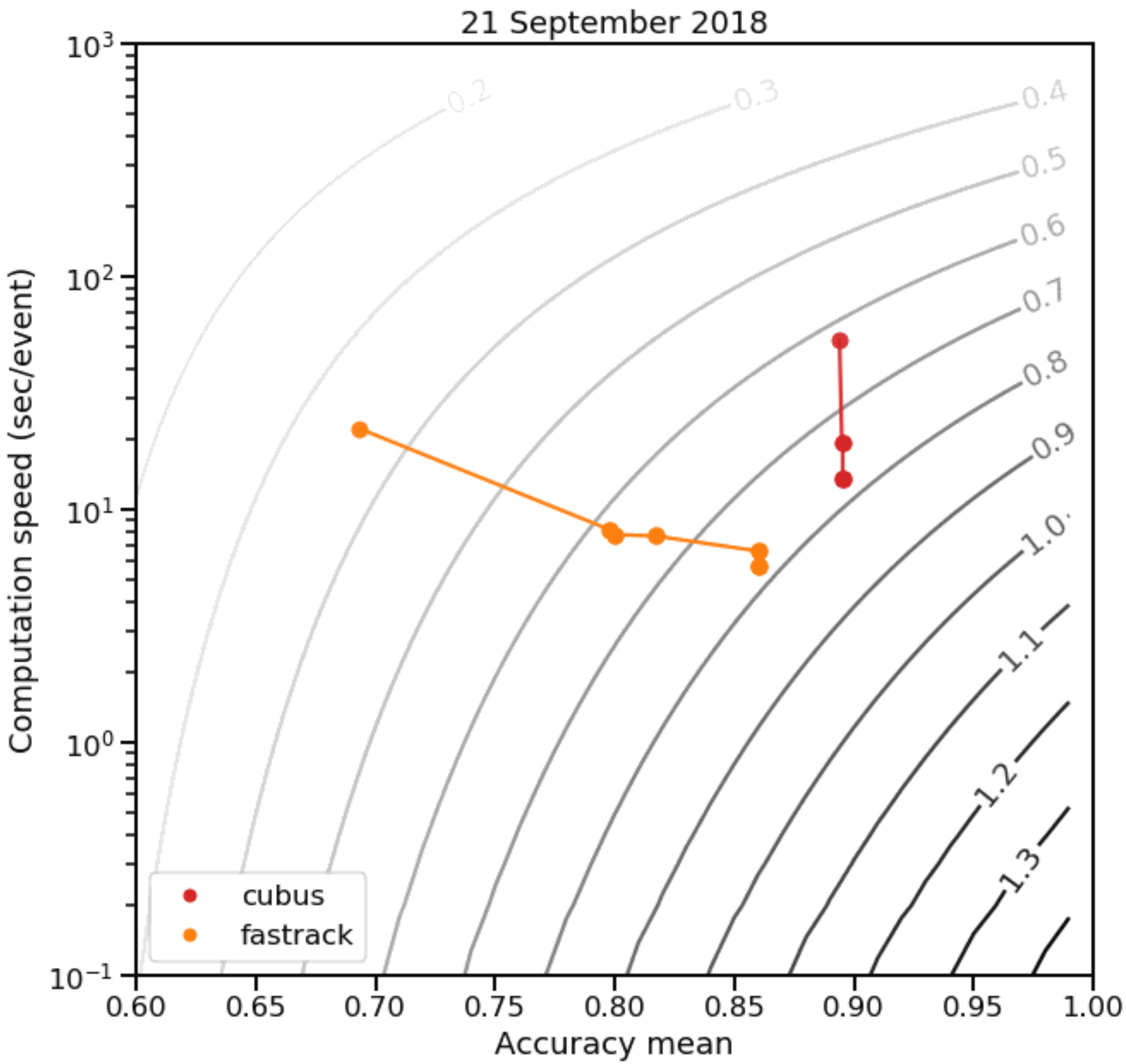
Phase 2 Aftermath Score evolution with time



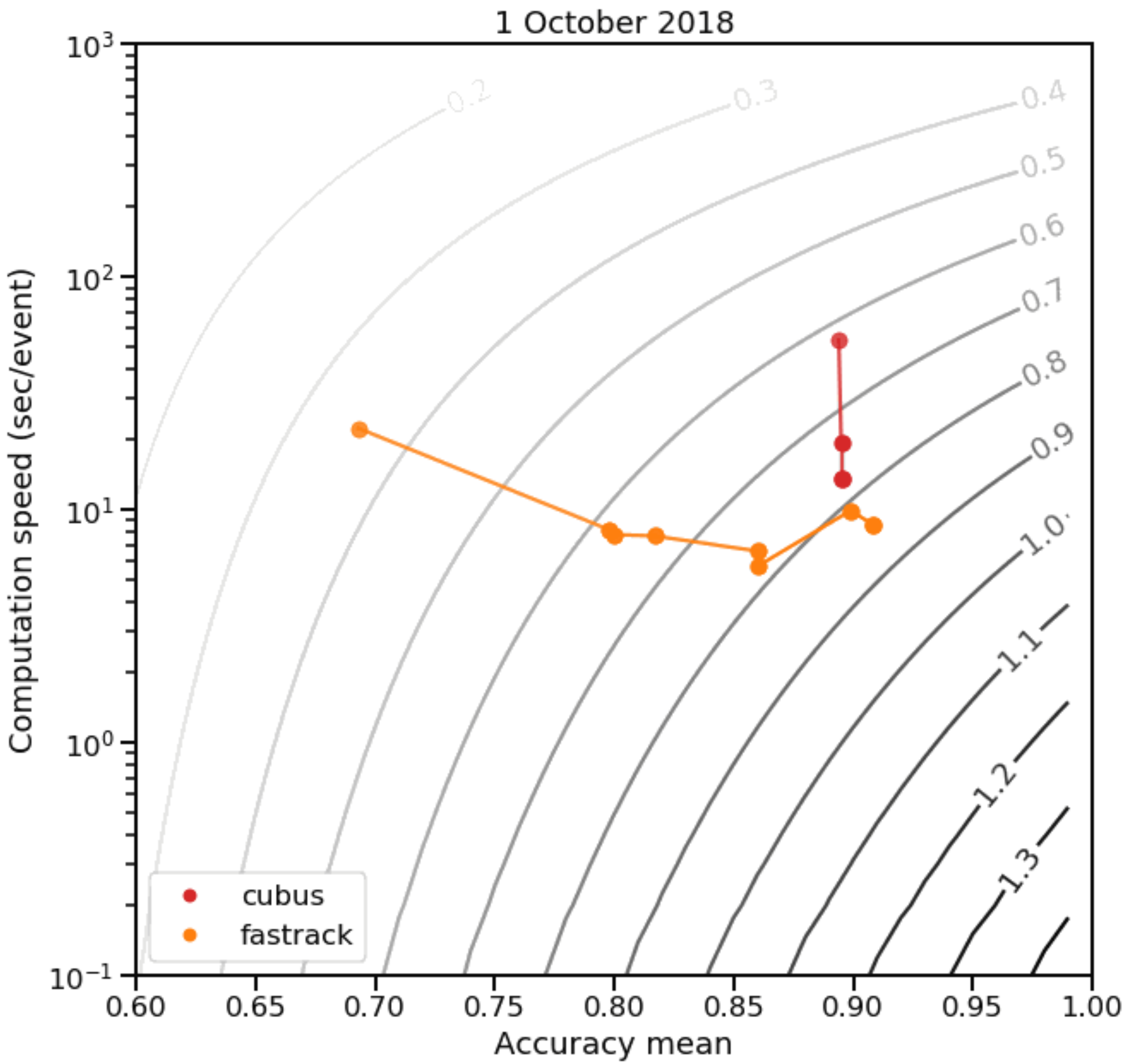
Phase 2 Aftermath Score evolution with time



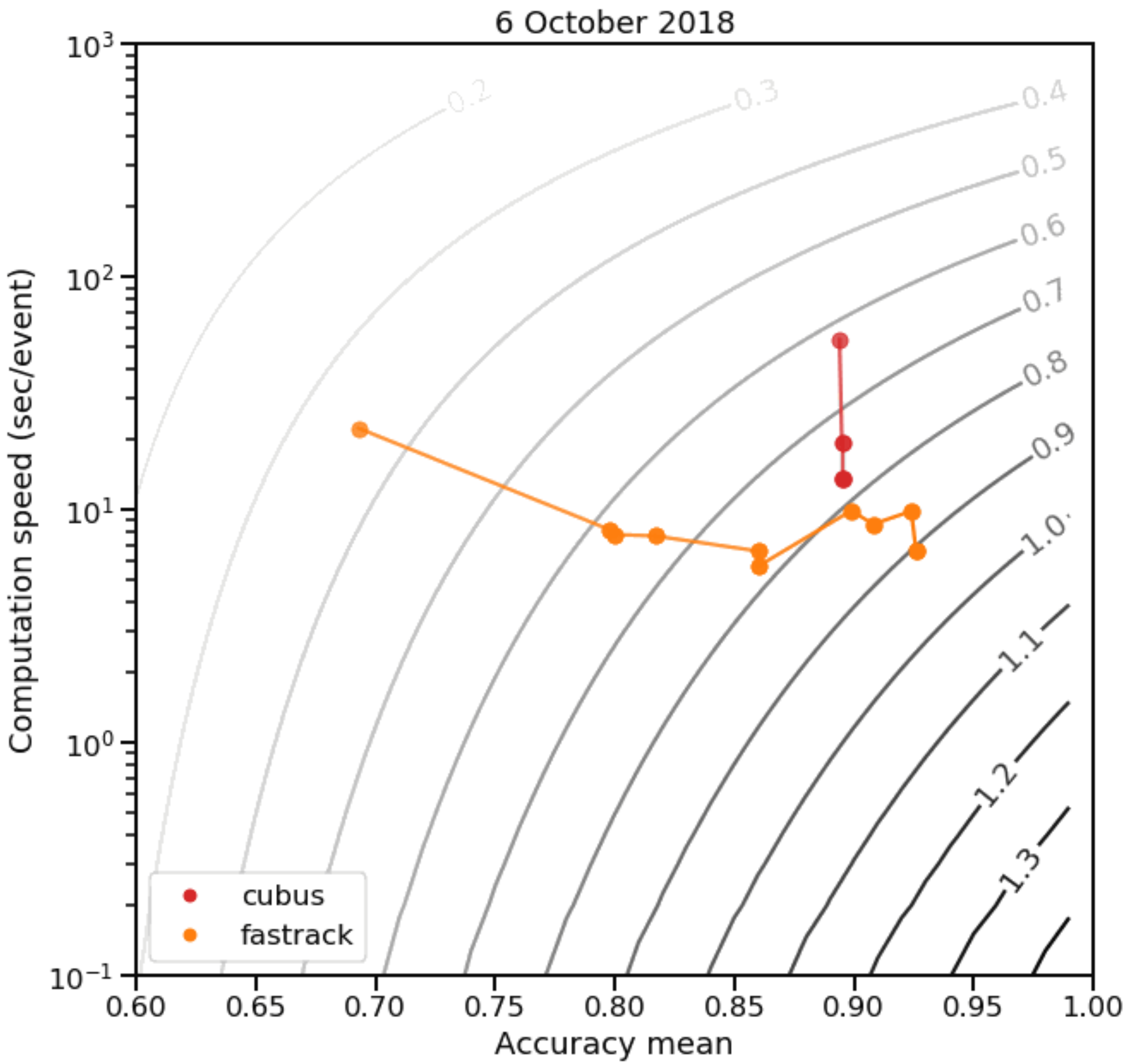
Phase 2 Aftermath Score evolution with time



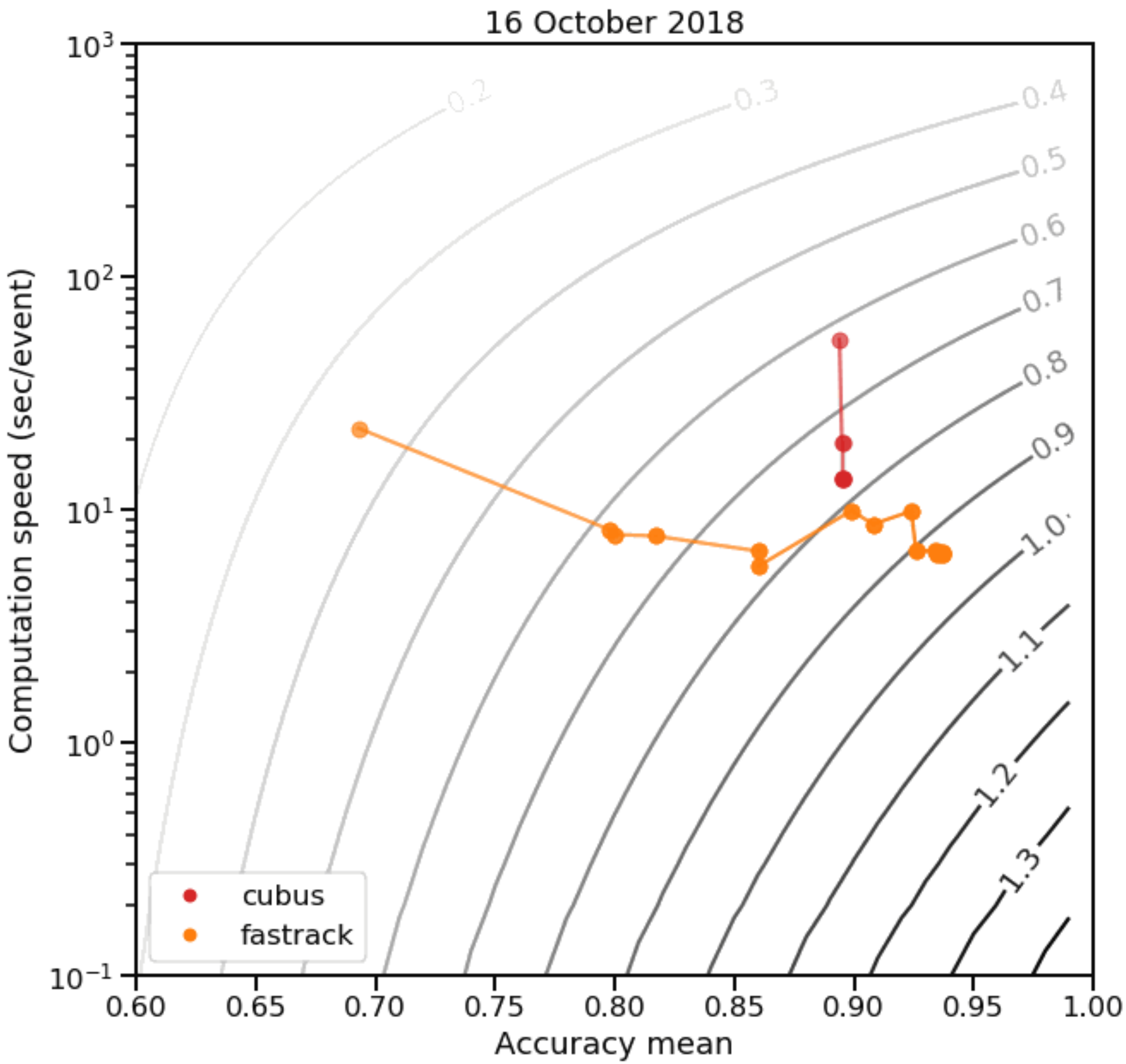
Phase 2 Aftermath Score evolution with time



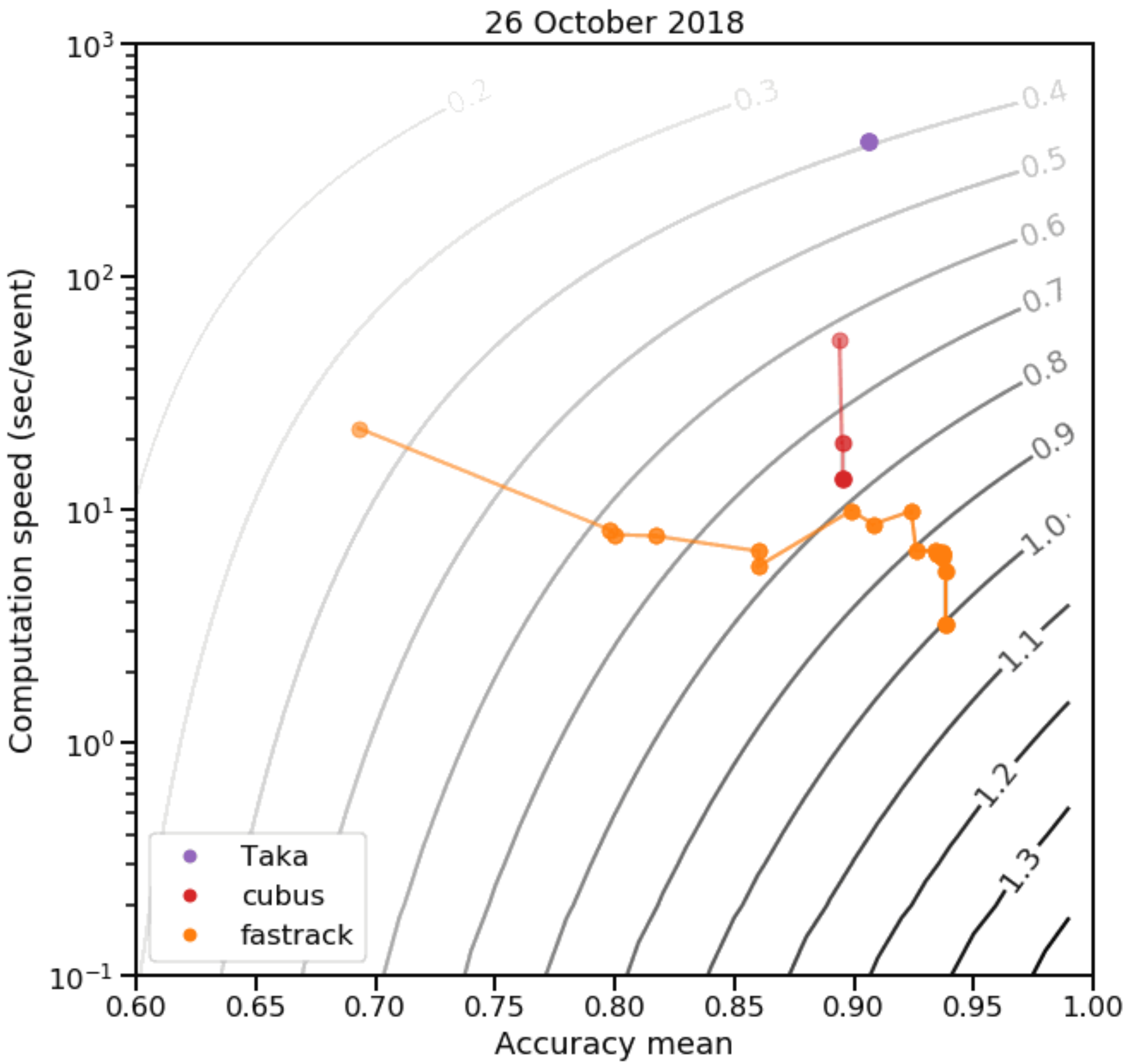
Phase 2 Aftermath Score evolution with time



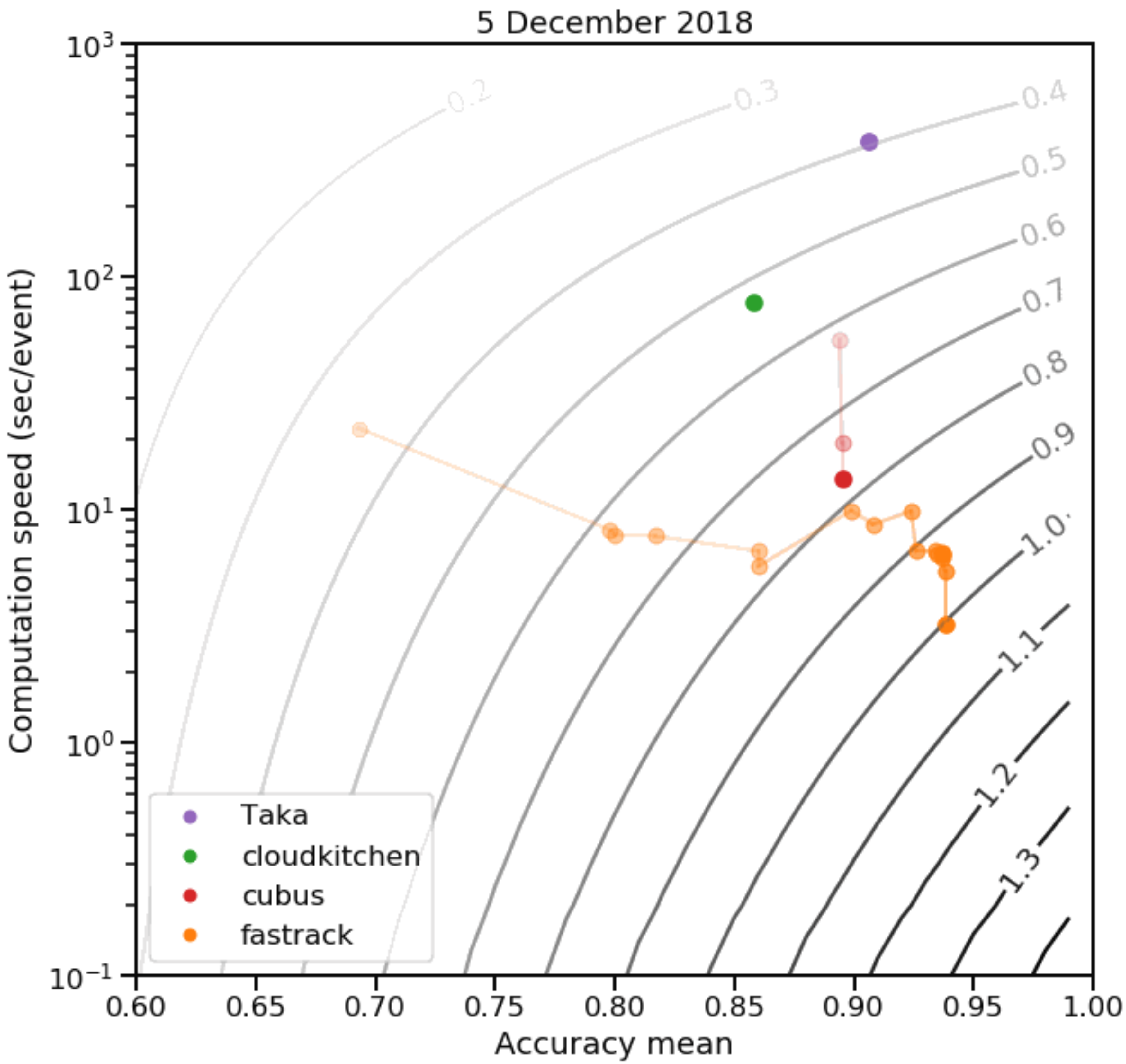
Phase 2 Aftermath Score evolution with time



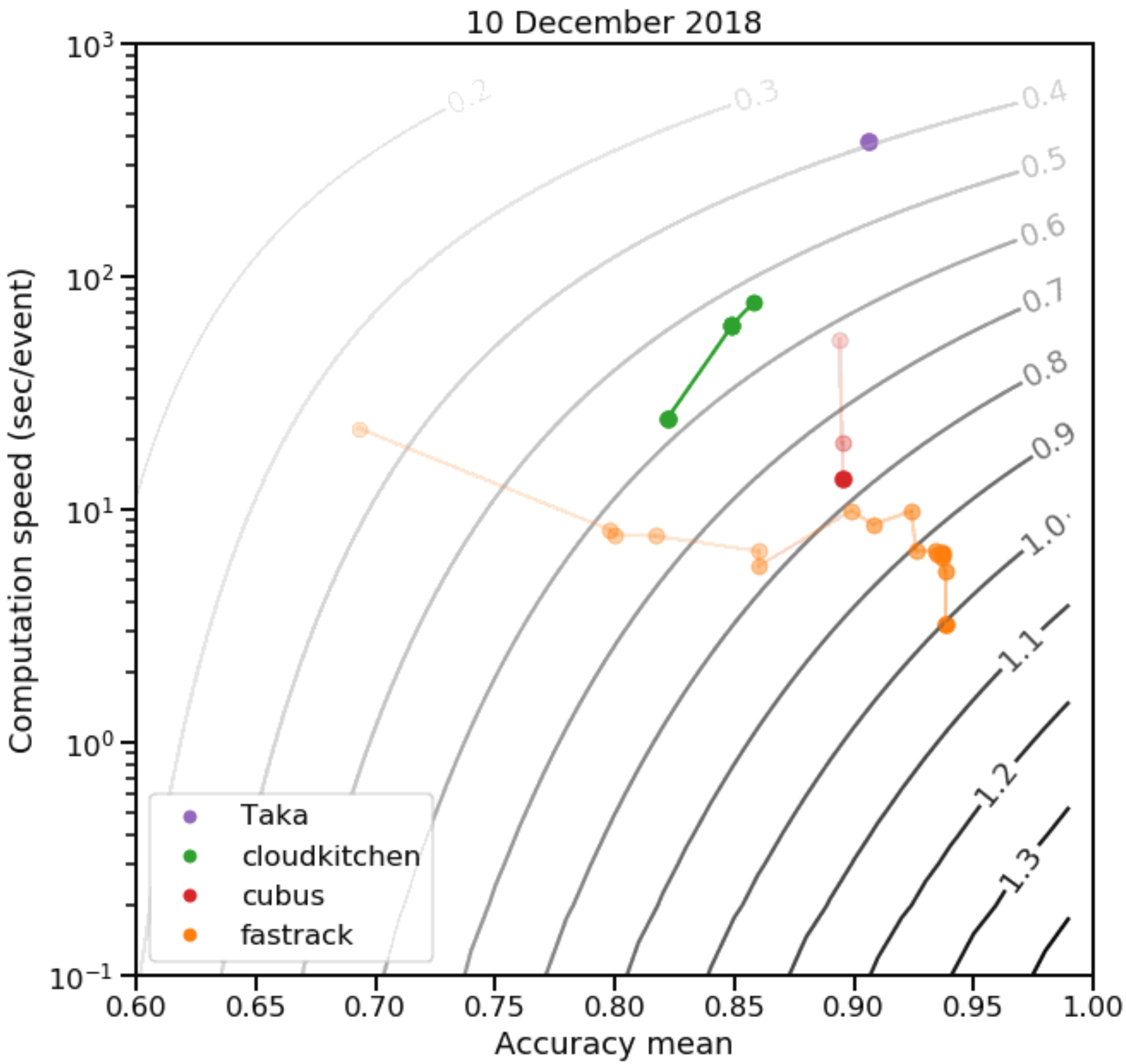
Phase 2 Aftermath Score evolution with time



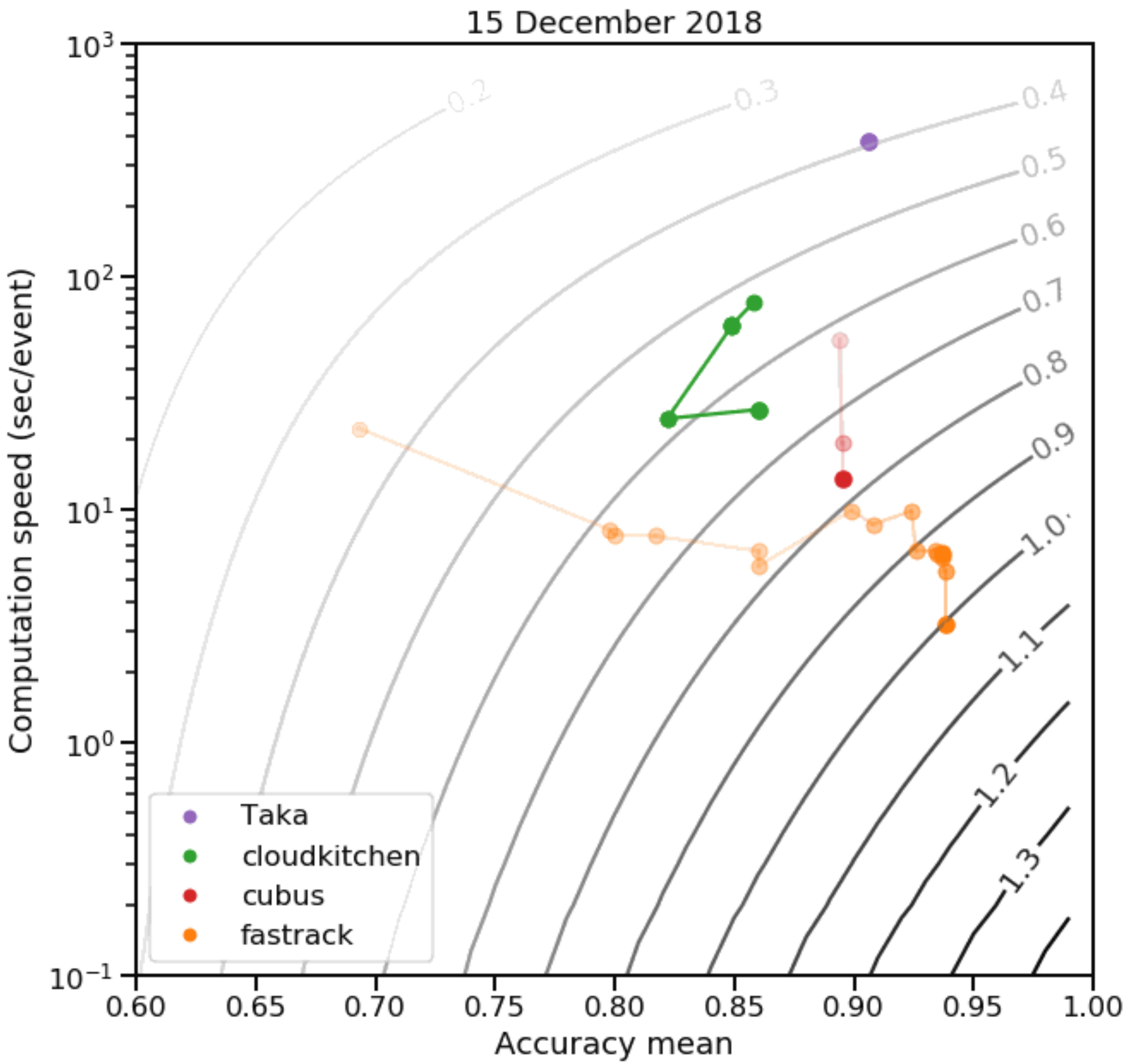
Phase 2 Aftermath Score evolution with time



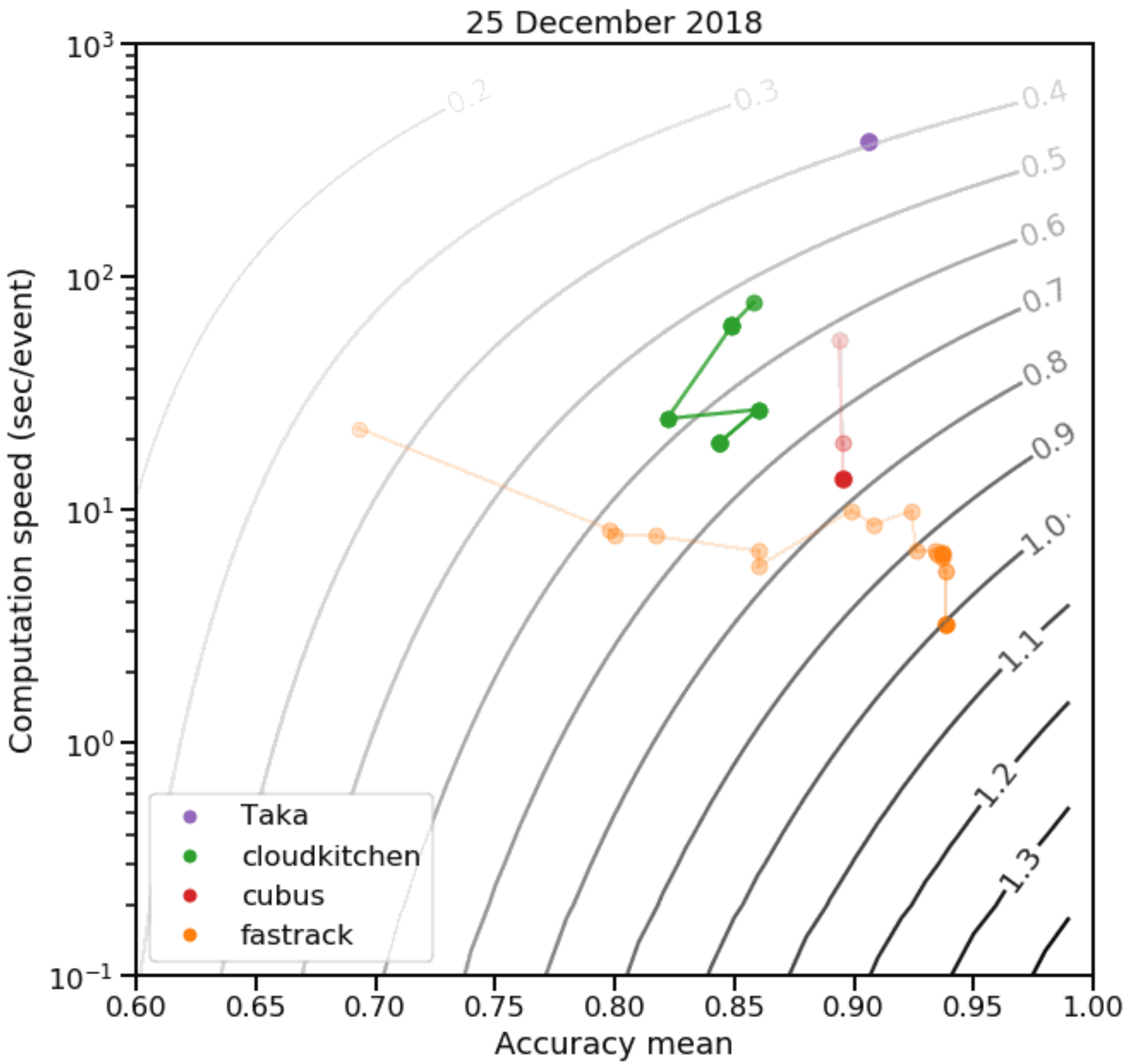
Phase 2 Aftermath Score evolution with time



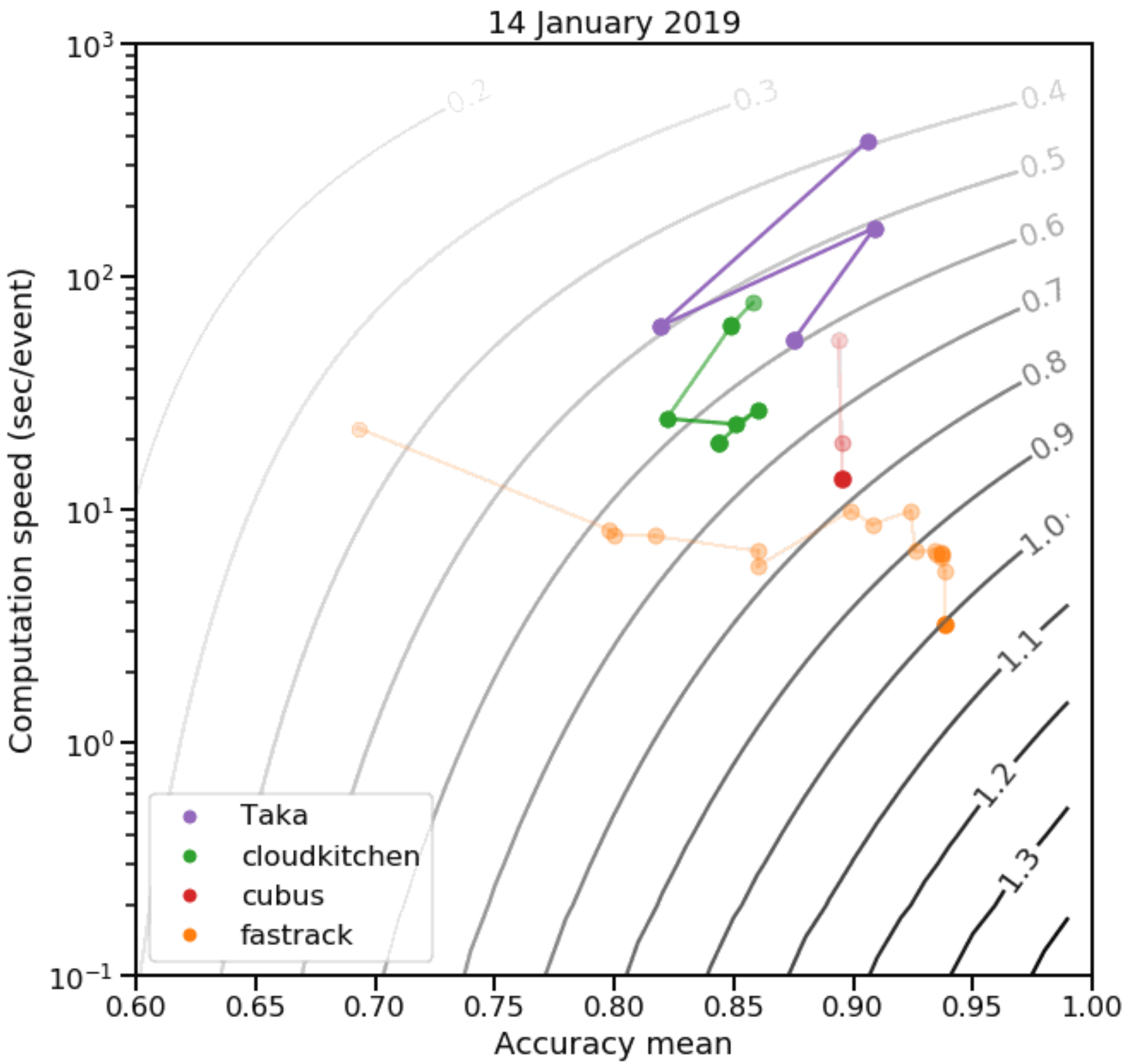
Phase 2 Aftermath Score evolution with time



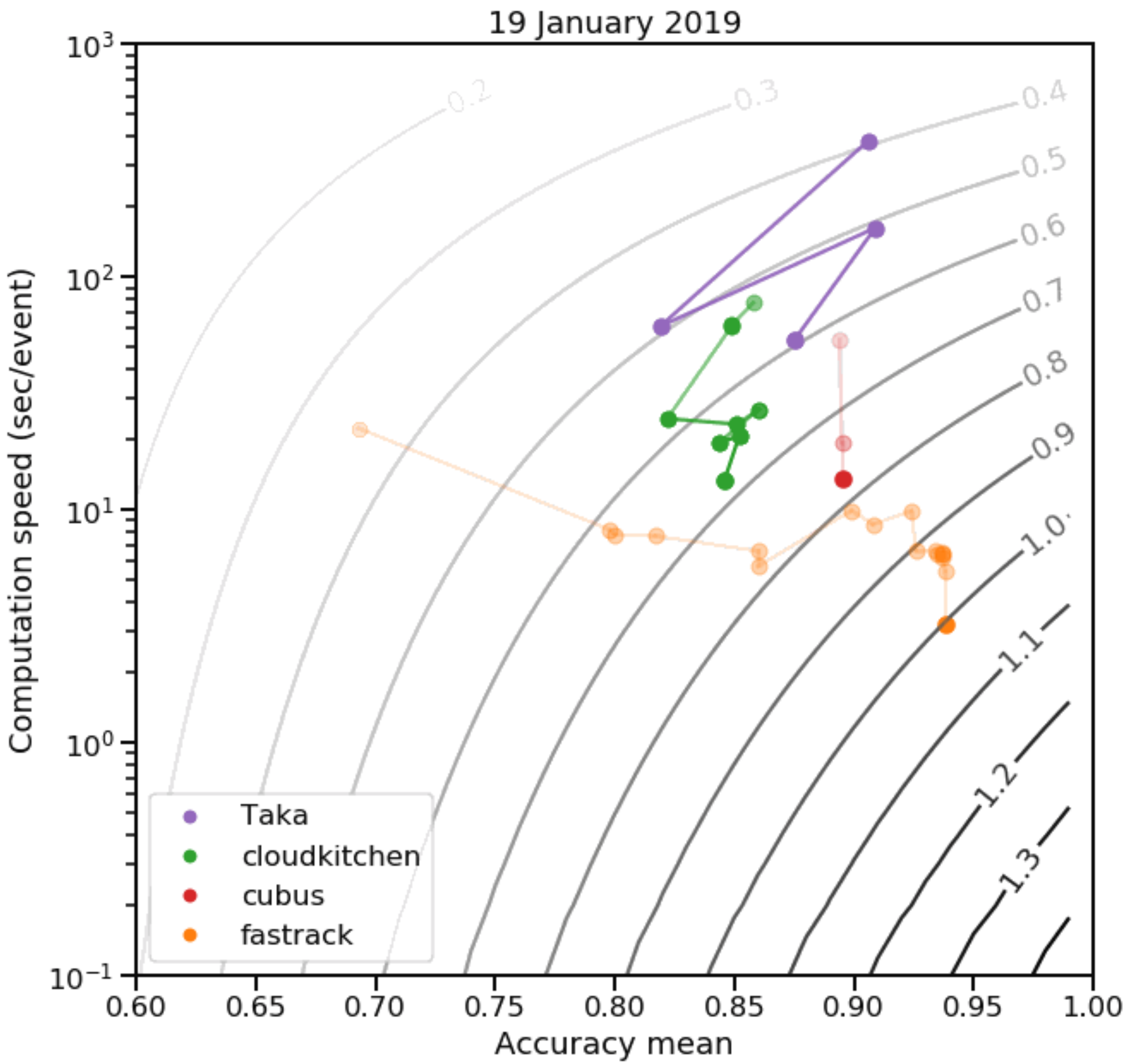
Phase 2 Aftermath Score evolution with time



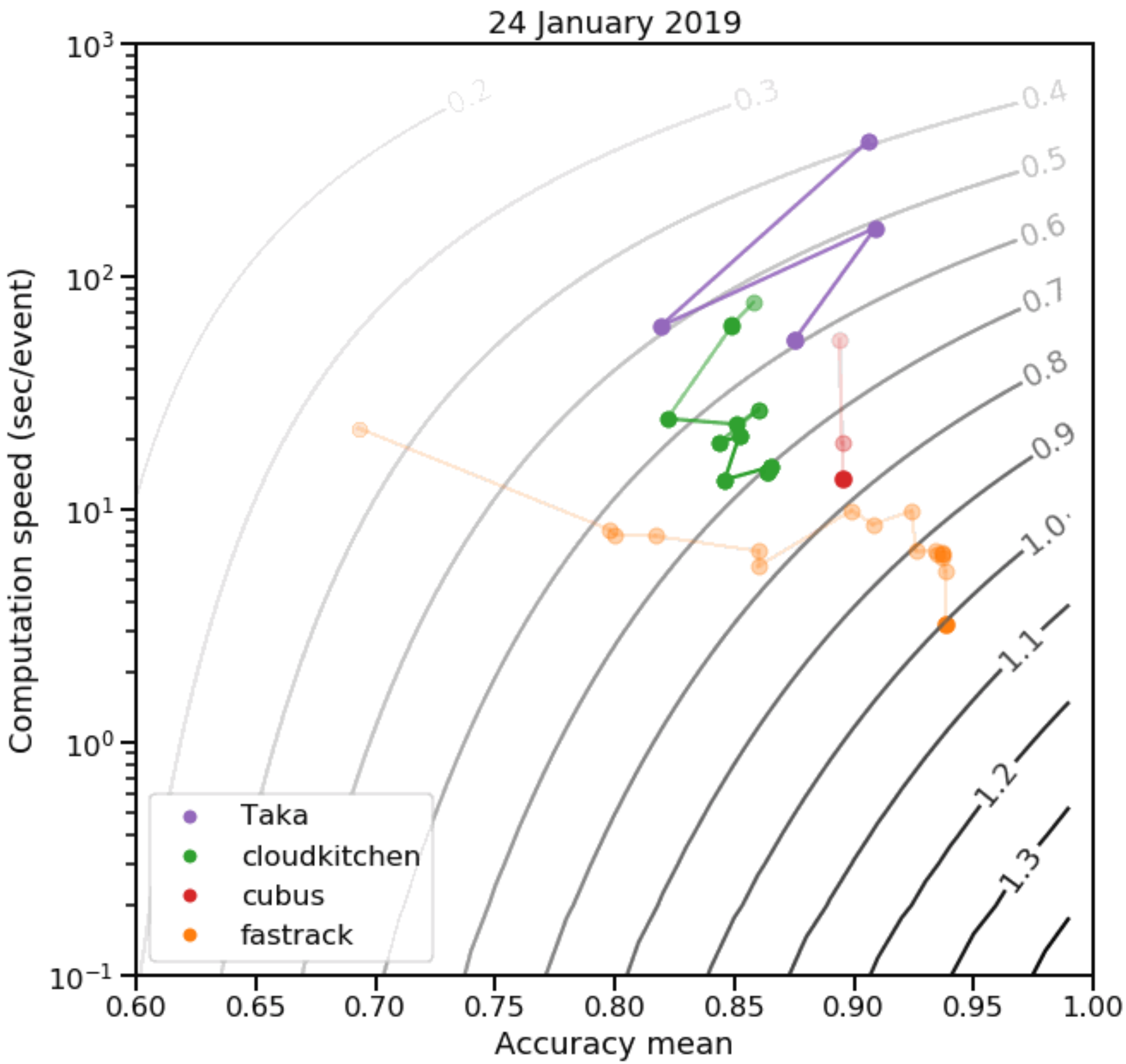
Phase 2 Aftermath Score evolution with time



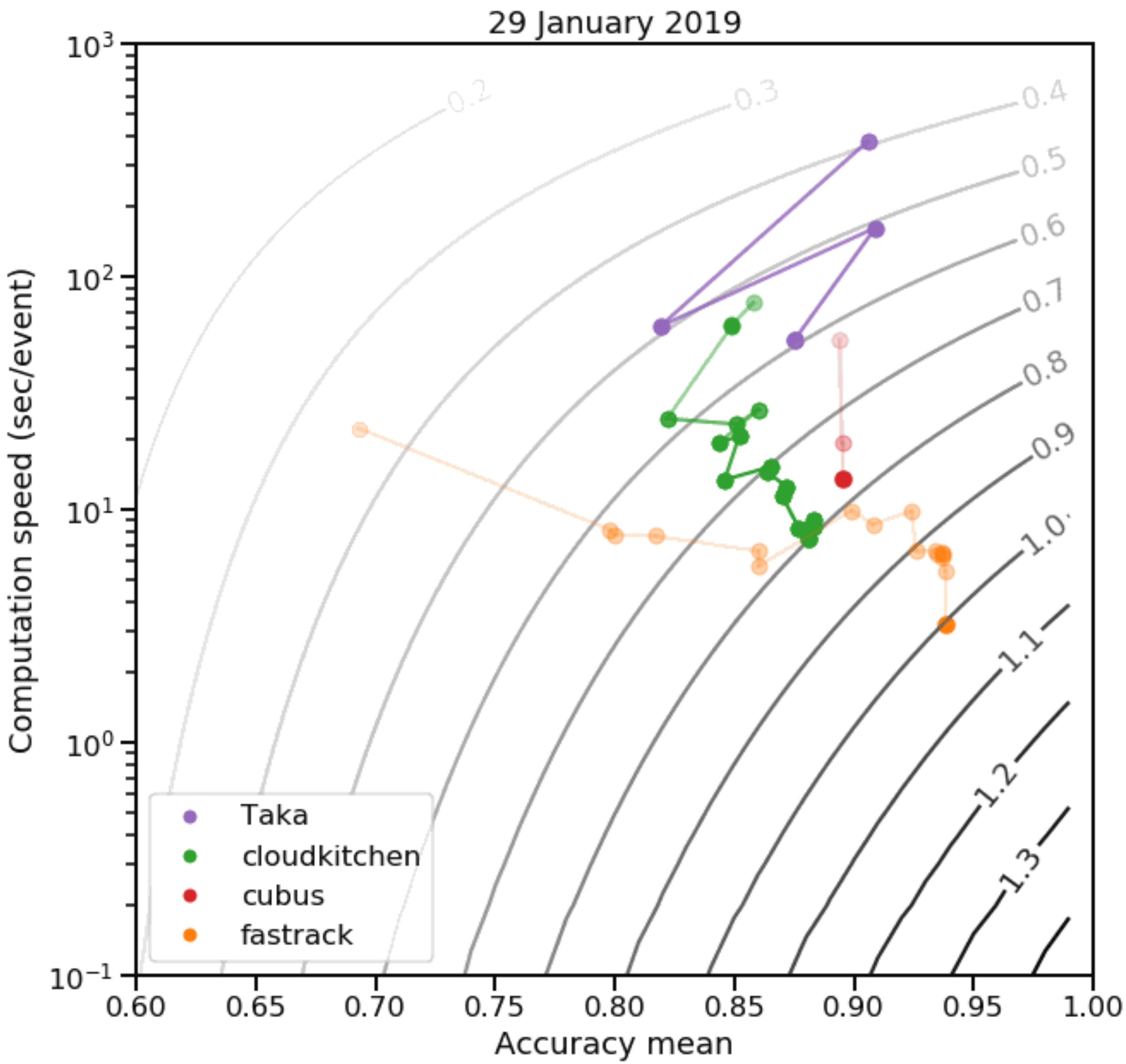
Phase 2 Aftermath Score evolution with time



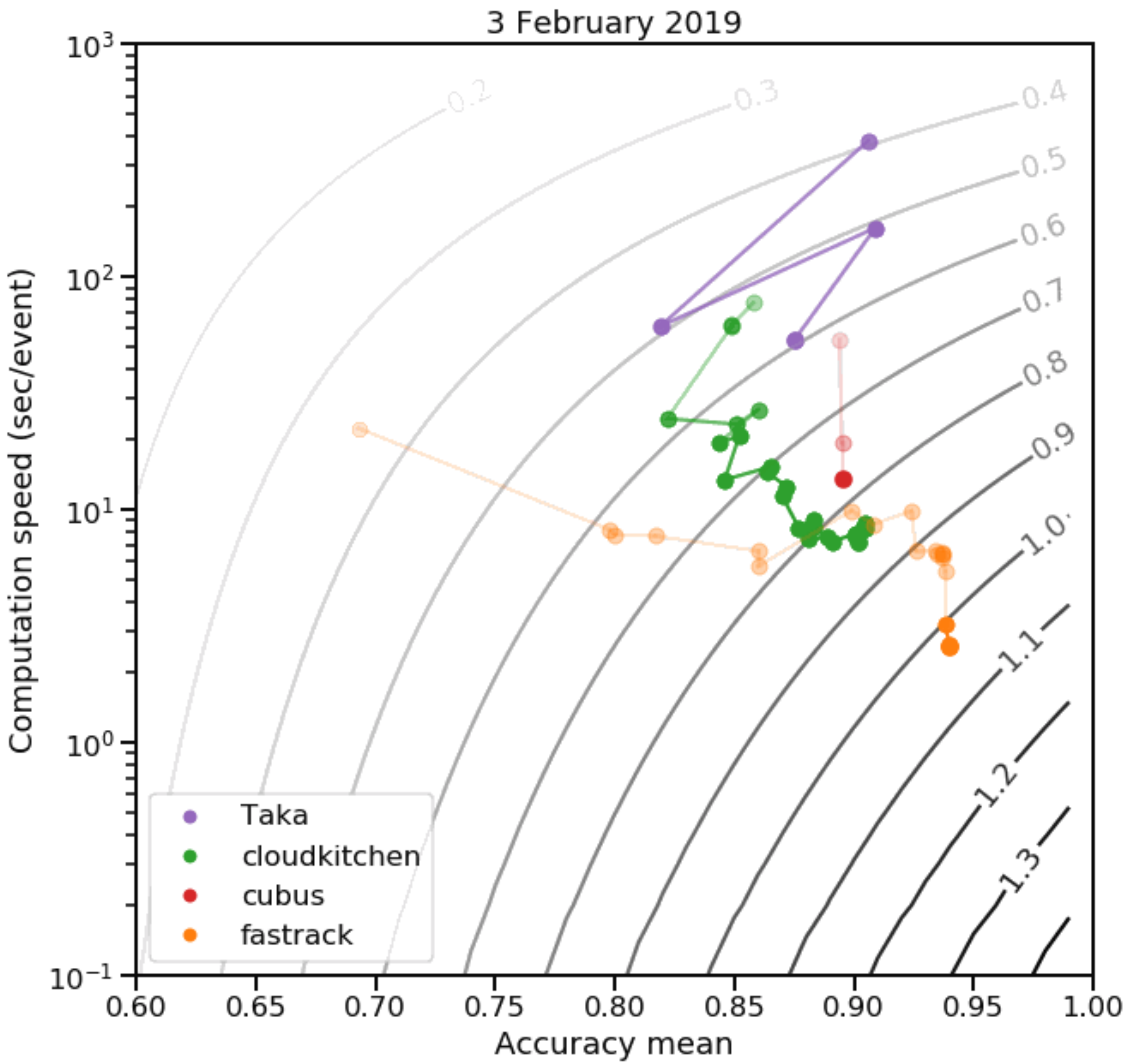
Phase 2 Aftermath Score evolution with time



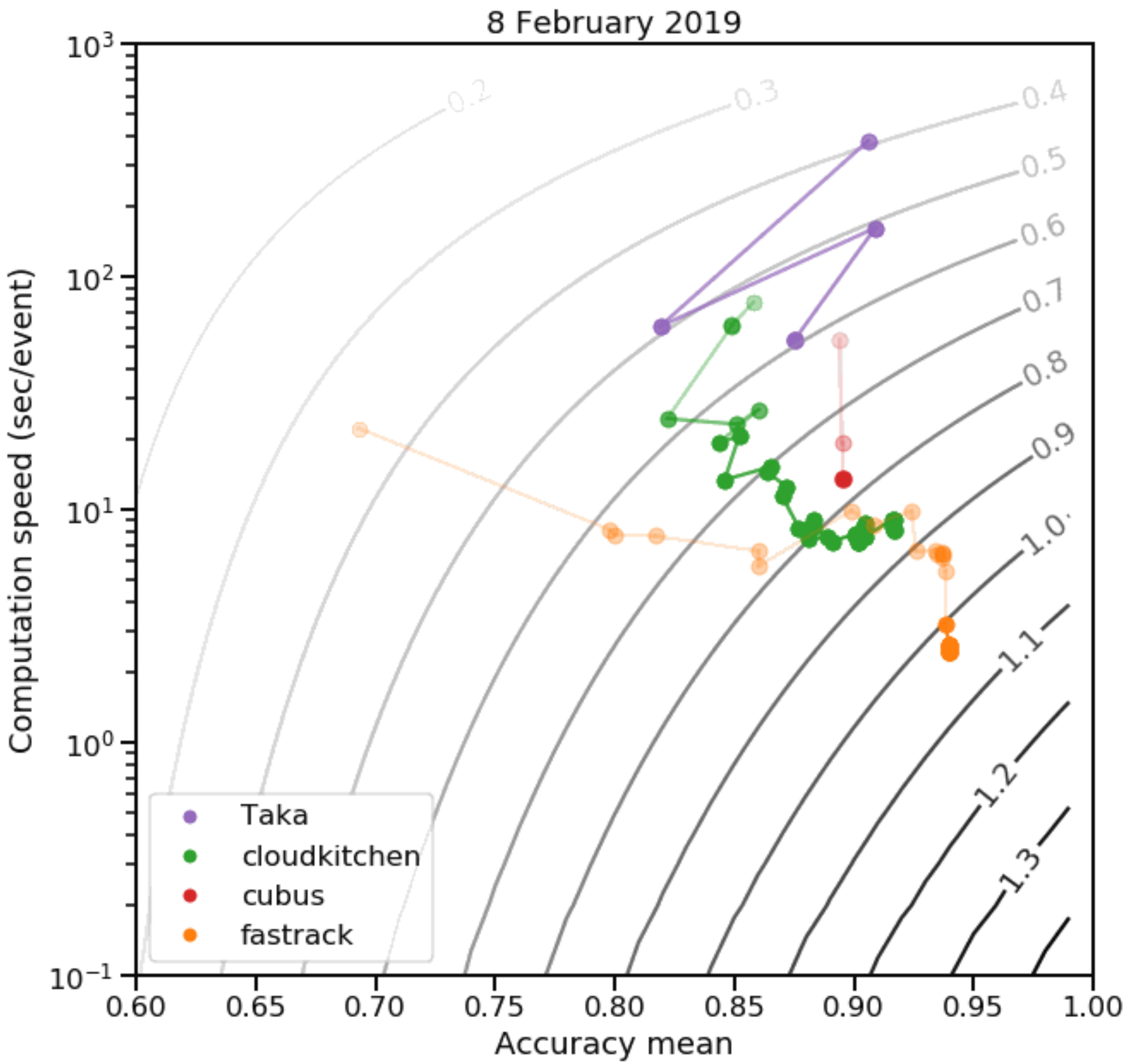
Phase 2 Aftermath Score evolution with time



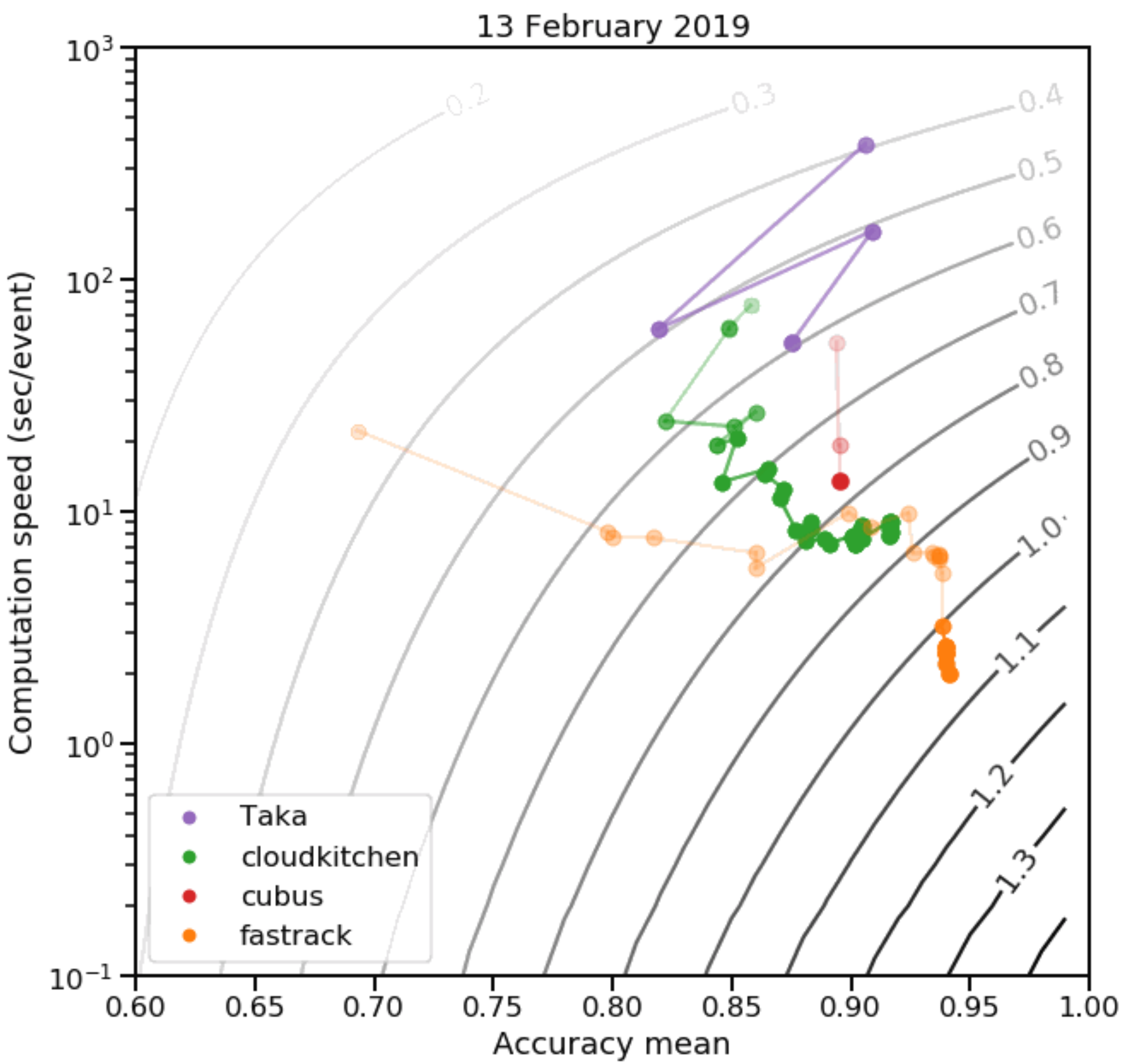
Phase 2 Aftermath Score evolution with time



Phase 2 Aftermath Score evolution with time

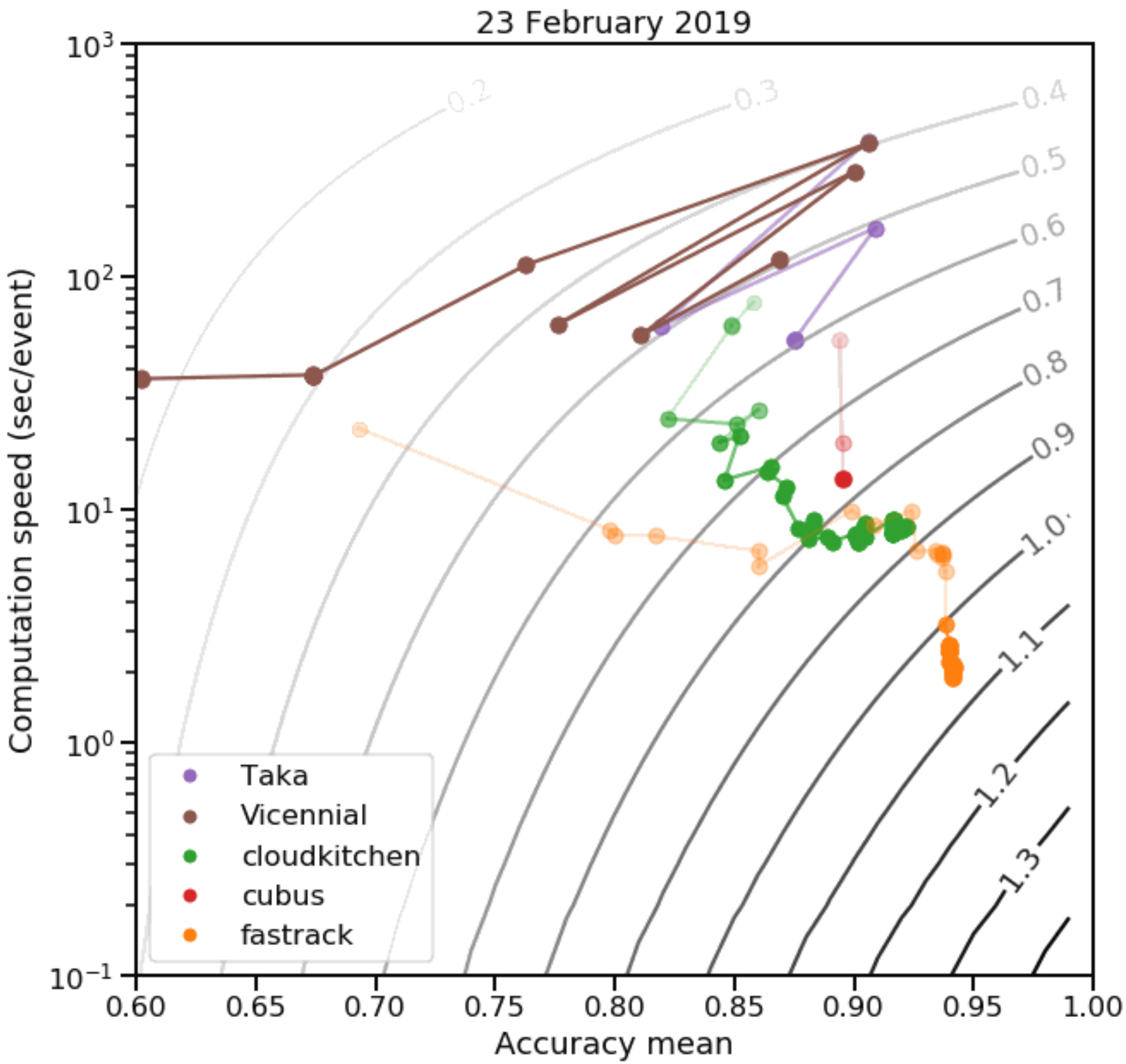


Phase

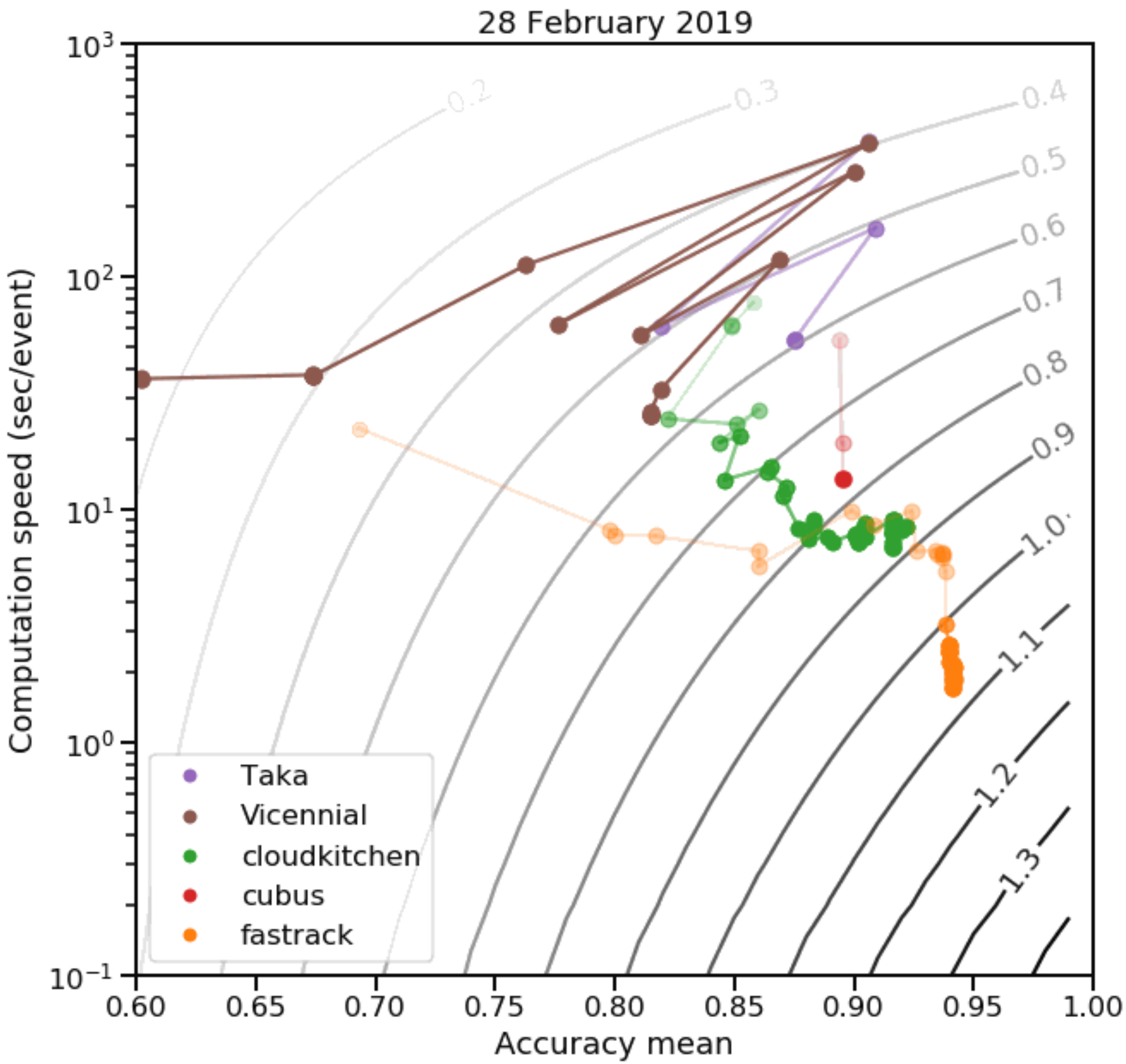


13 February 2019
Computation speed (sec/event)
Accuracy mean
10^3
10^2
10^1
10^0
10^{-1}
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
Taka
cloudkitchen
cubus
fastrack

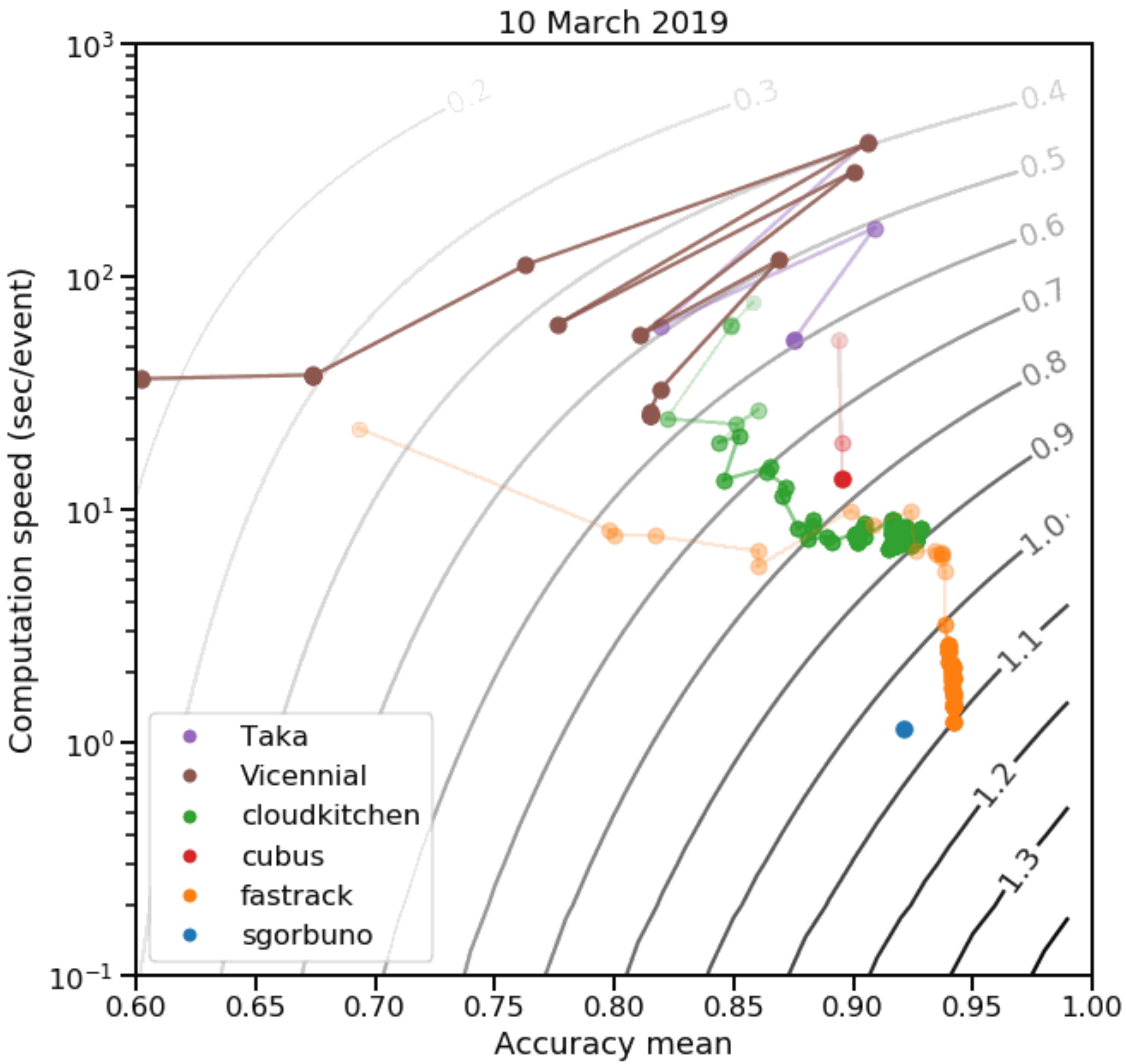
Phase 2 Aftermath Score evolution with time



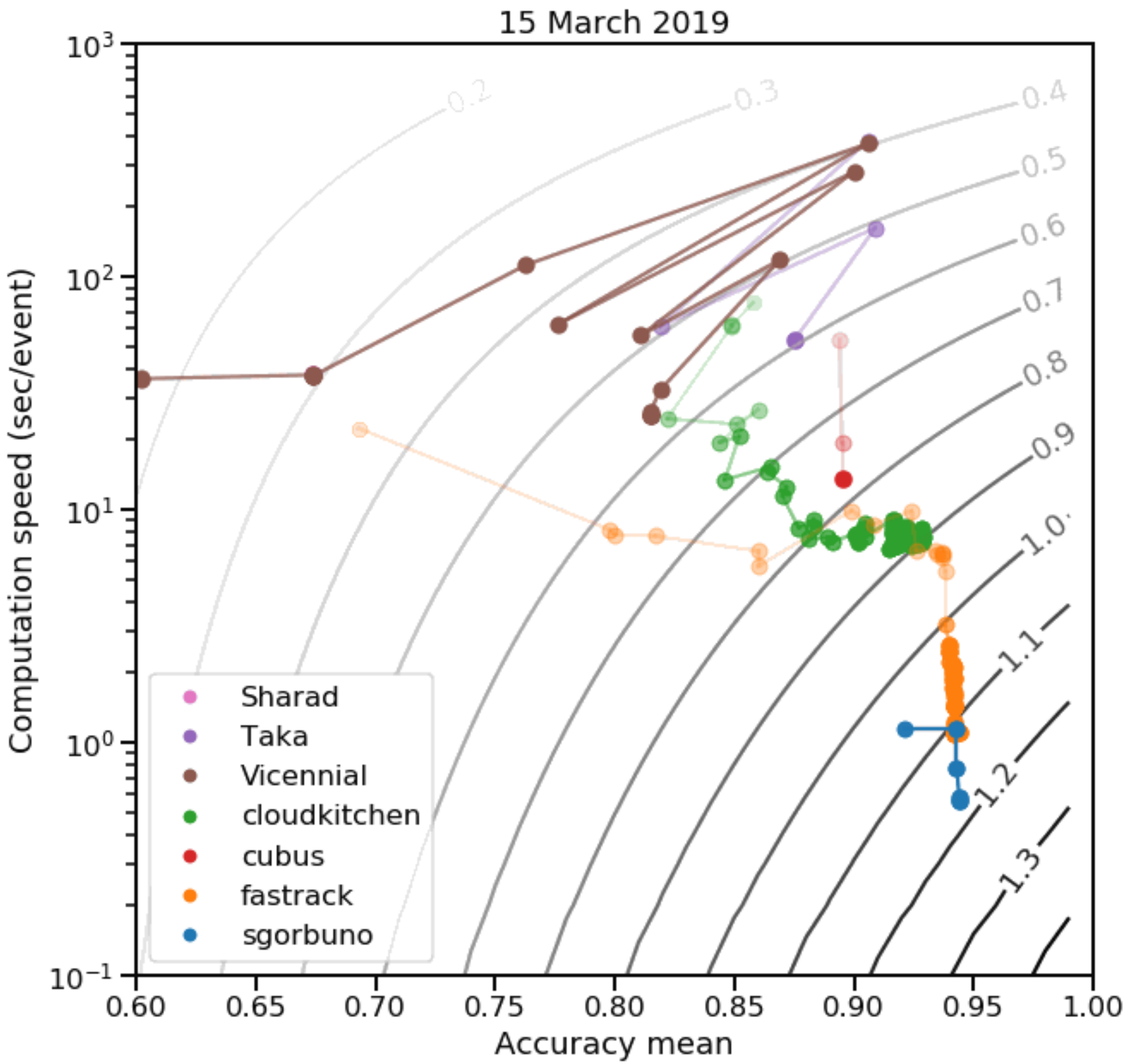
Phase 2 Aftermath Score evolution with time



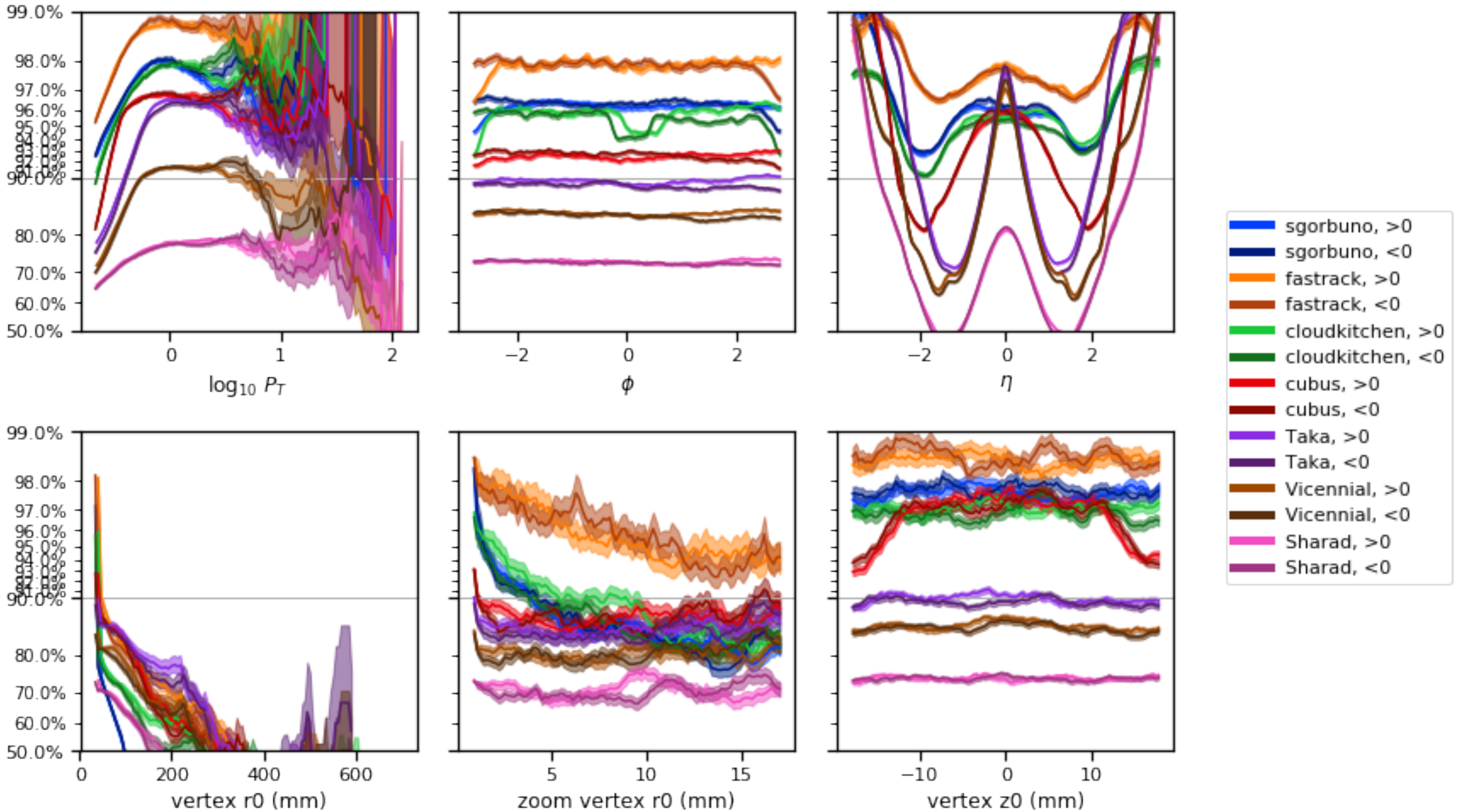
Phase 2 Aftermath Score evolution with time



Phase 2 Aftermath Score evolution with time



Phase 2 Aftermath Tracking efficiency



Phase 2 Aftermath

Phase 2 closed a fortnight ago - just starting

- there are way fewer submissions though
- currently collecting code and submission contributions

Longer term projects

- GSoC (embedded in CERN-HSF context) project submitted to re-implement the algorithms as parts of the ACTS project
- Would allow to run to test on a variety of detectors

Announcement:

Final TrackML Workshop, July 1^{st} & 2^{nd}, 2019

@CERN

Phase 1 & Phase 2

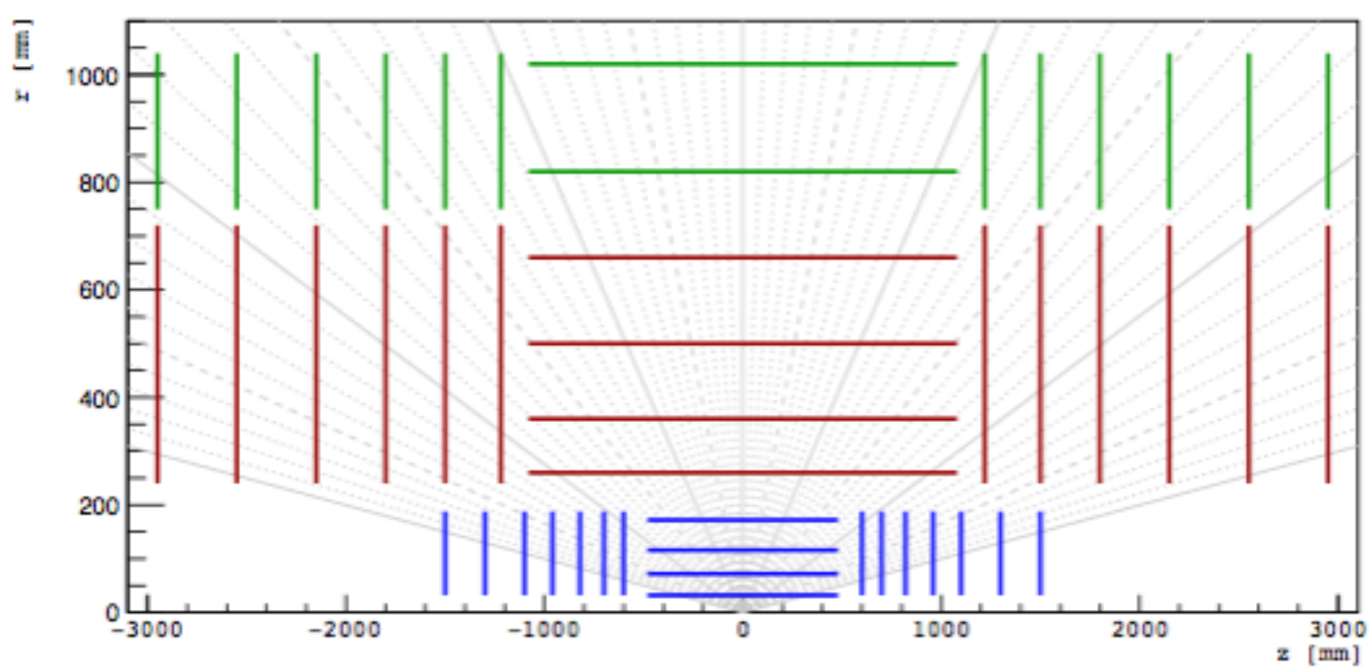
Spin-off

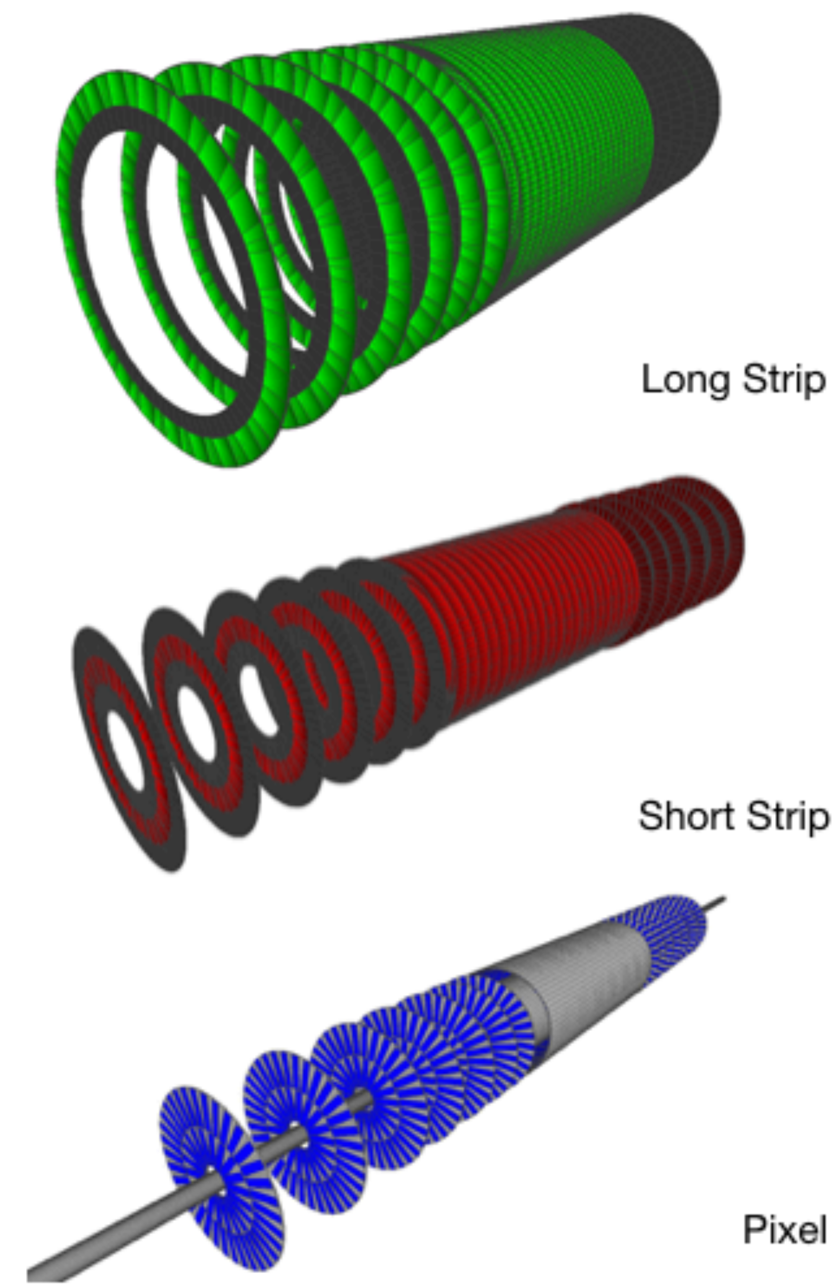
Spin-off

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

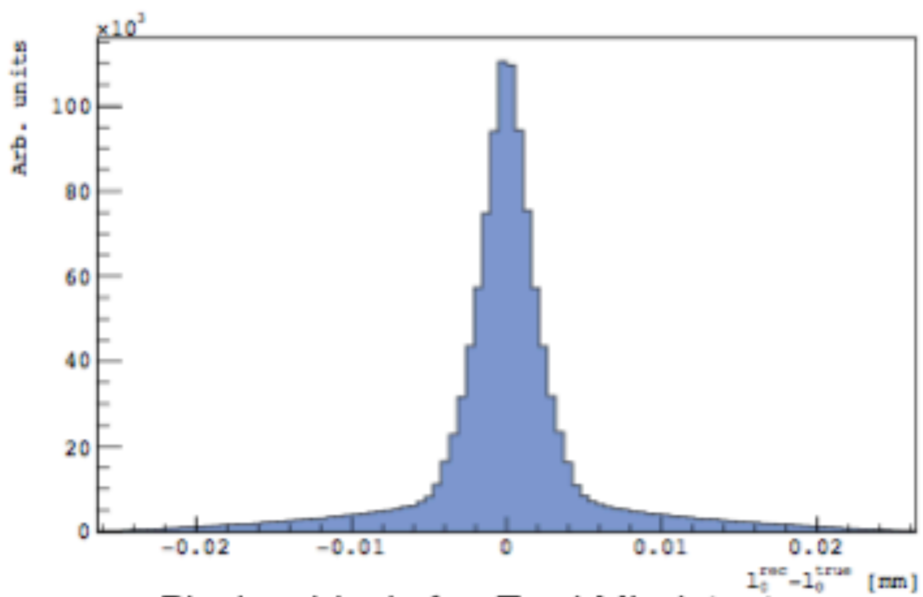




Spin-off

Reference detector&dataset

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



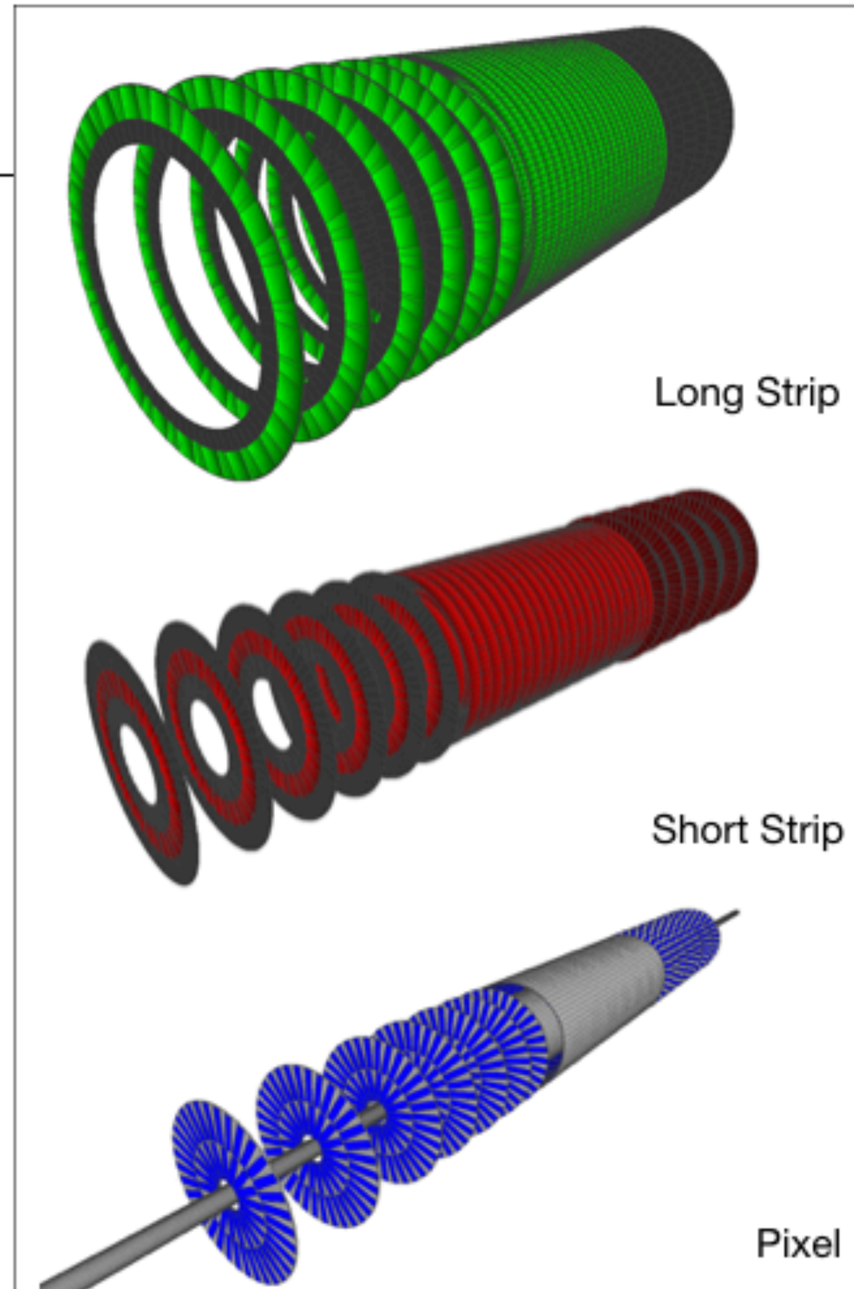
Pixel residuals for TrackML detector
(50 µm x 50 µm pixel size)

opendata CERN

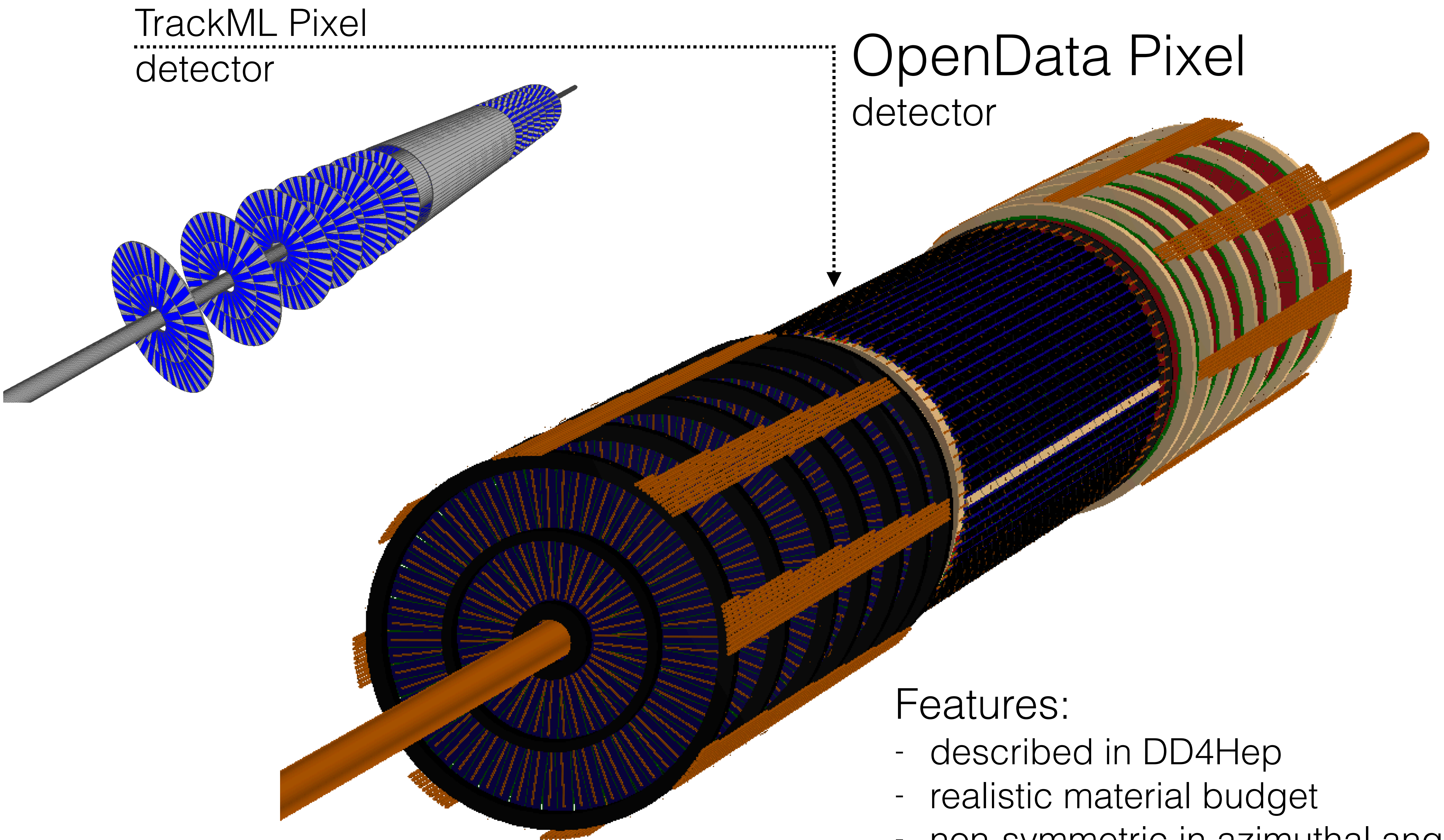
development & testing ecosystem

Experiment SW ecosystem

publish



Spin-off Sneak Preview



Features:

- described in DD4Hep
- realistic material budget
- non-symmetric in azimuthal angle
- full (G4) and fast (ACTS) simulation
- misalignment possibility

… to be released soon!

More Information & links

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

Announcement:

Final TrackML Workshop, July 1st & 2nd, 2019

@CERN

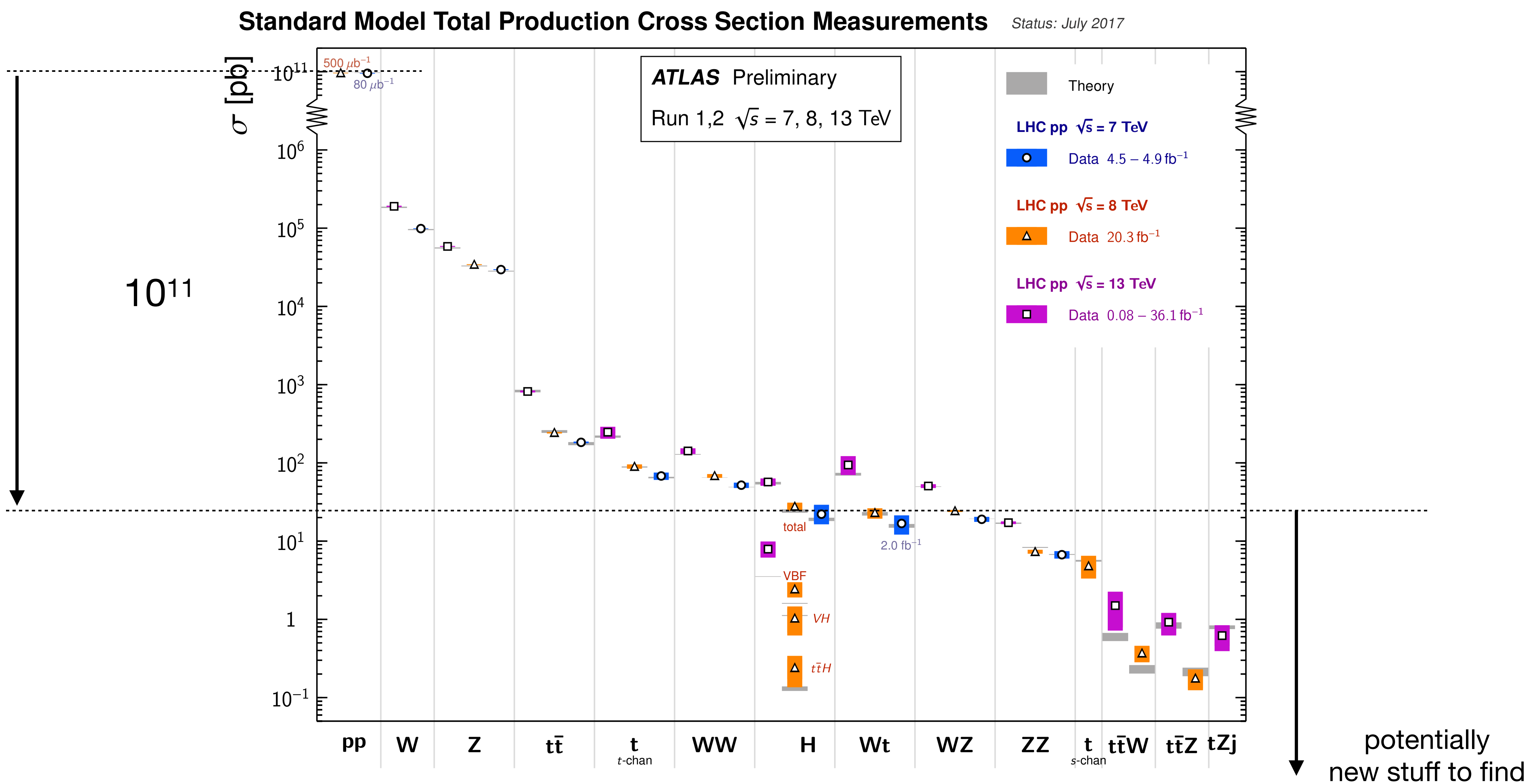
Phase 1 & Phase 2

Backup slides

Introduction Physics

Focus on hadron colliders as the LHC

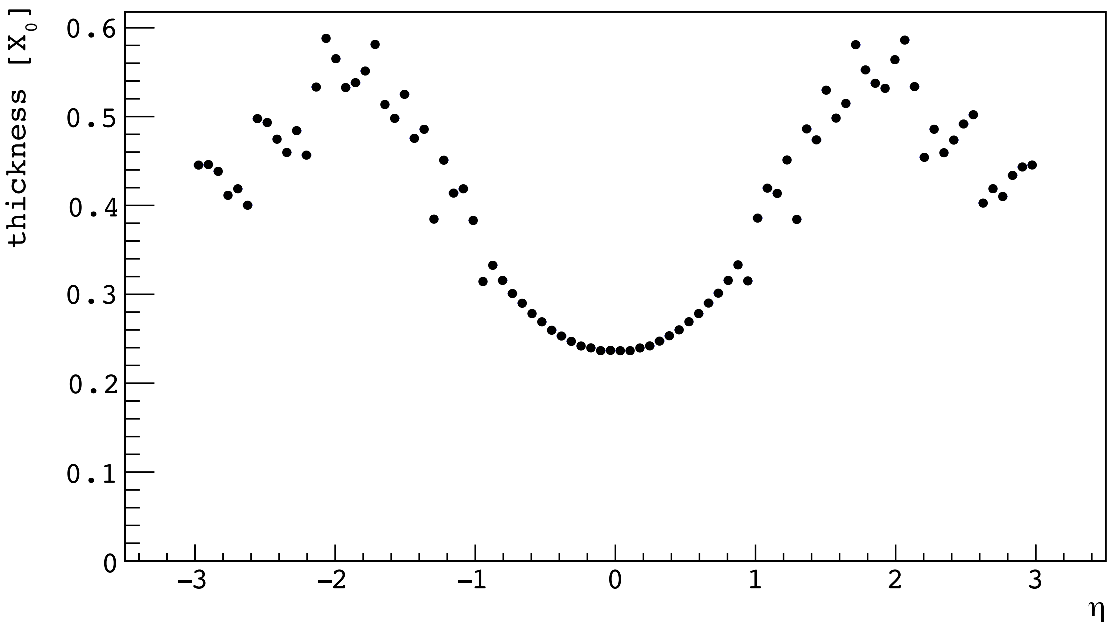
- High luminosity (HL-)LHC
- Future FCC-hh design study in preparation

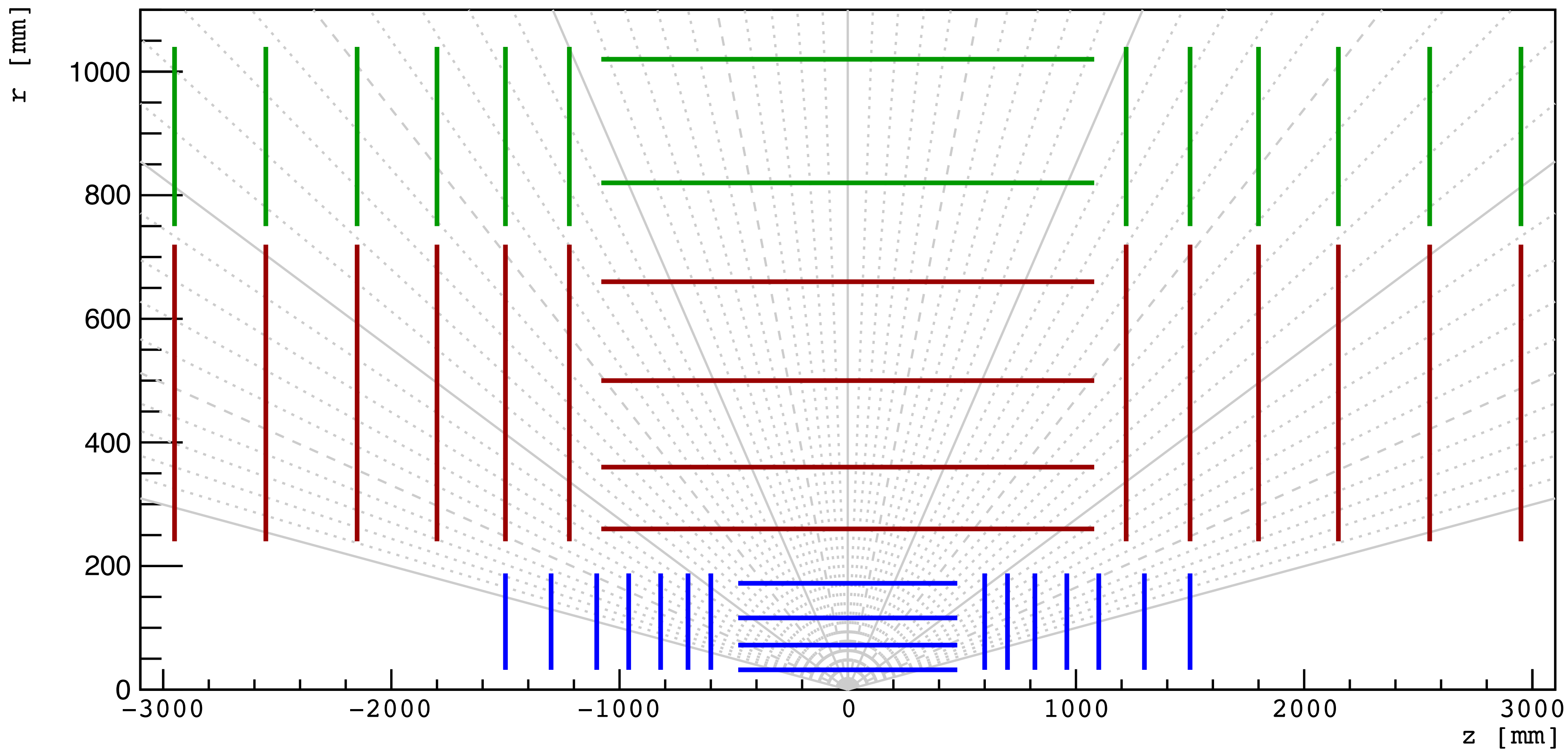


The **detector**

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



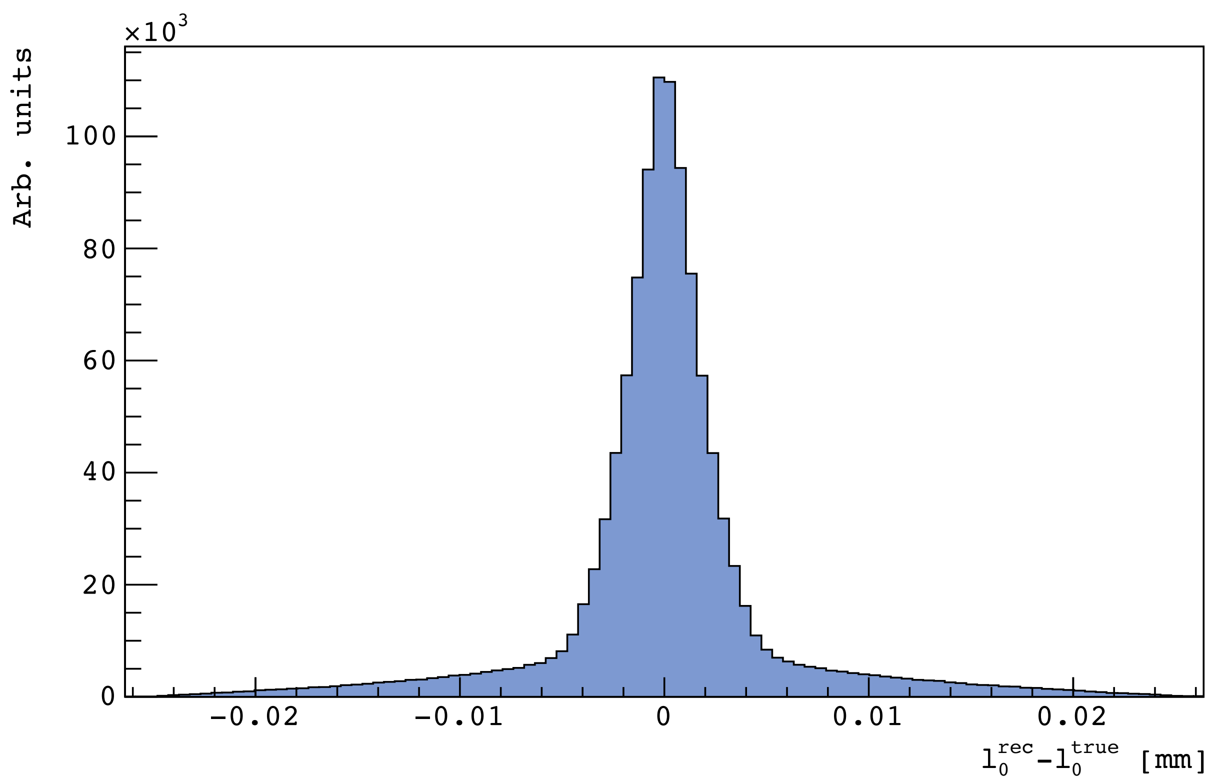


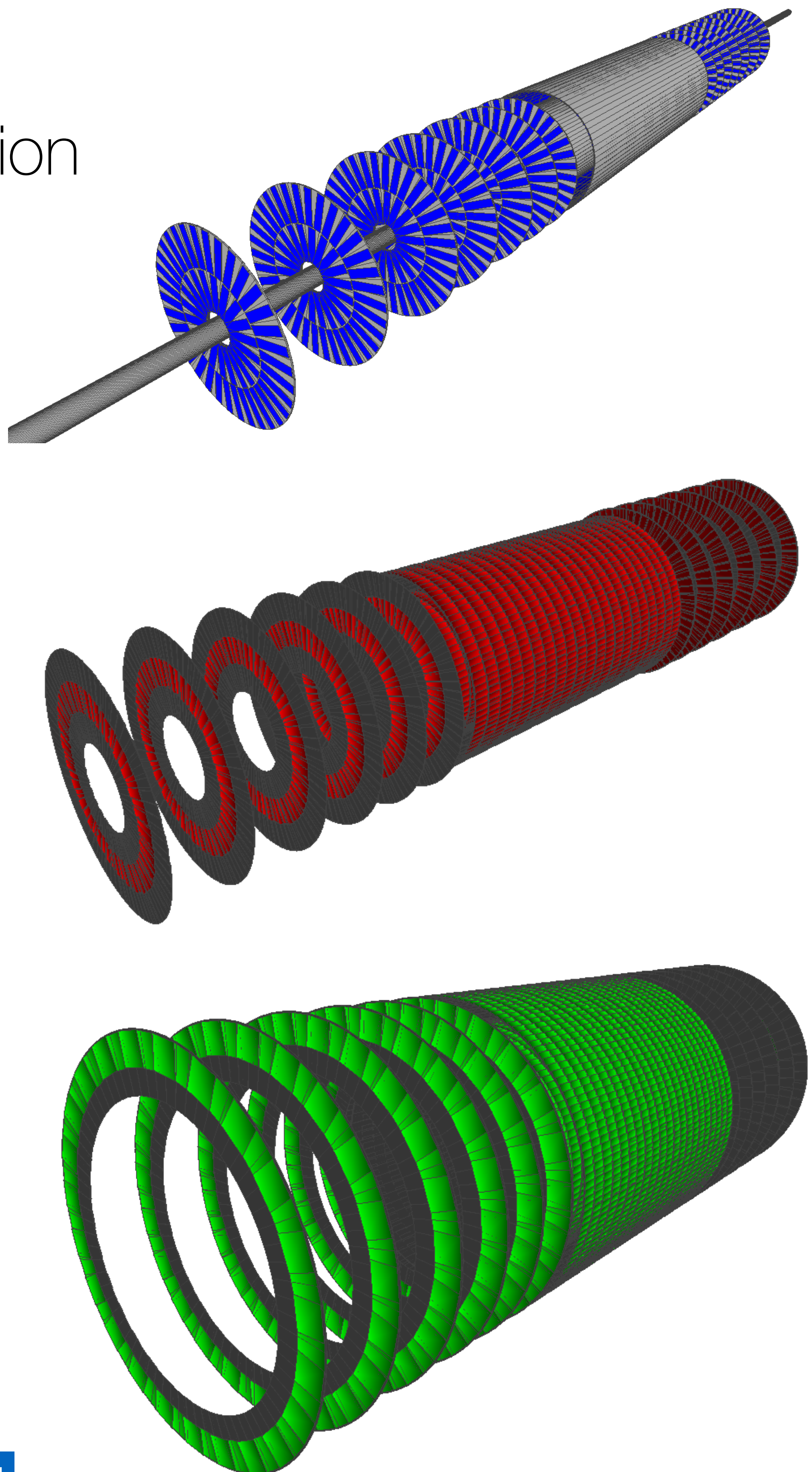
plot & image
(left) X0 distribution of the trackML detector
(right) longitudinal view of the trackML detector

The **detector**

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

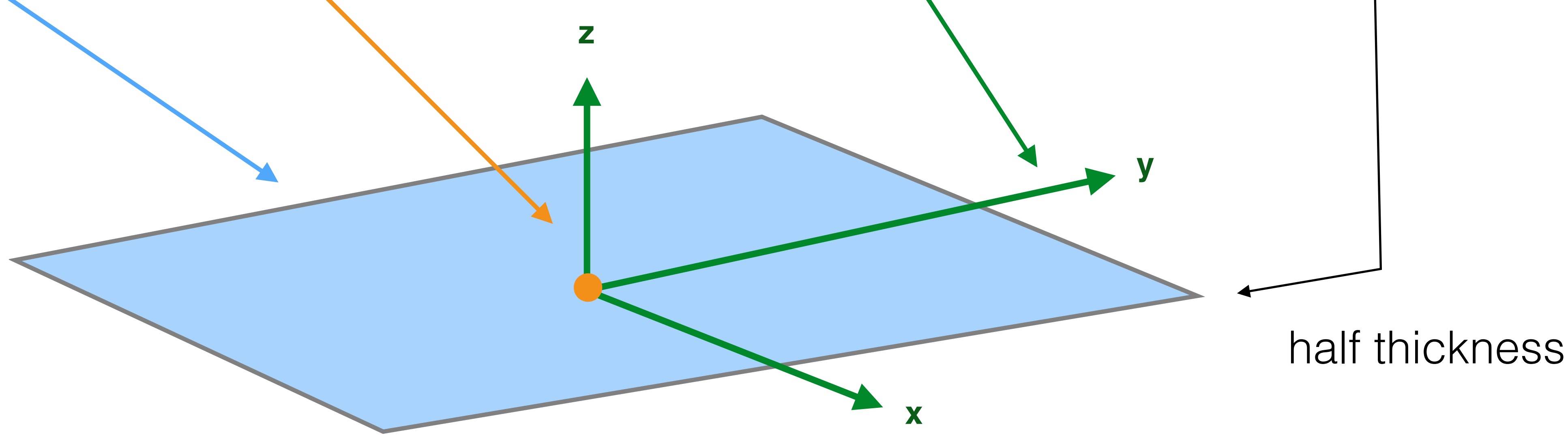
(left) estimated pixel resolution distribution

(right) 3D view of pixel, short strip and long strip detector

The **detector**

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	modu
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:

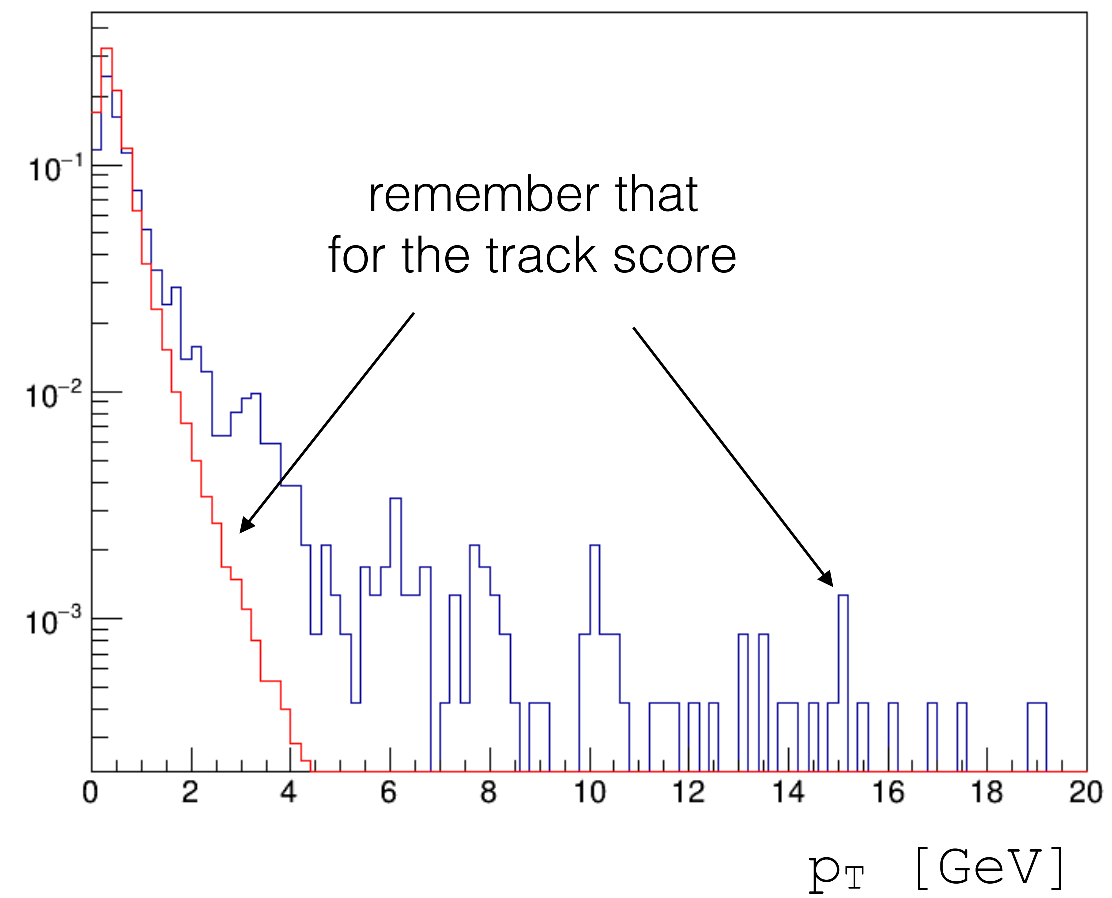
- HS: **"Top:gg2ttbar = on"**
- PU (@200): **"SoftQCD = on"**

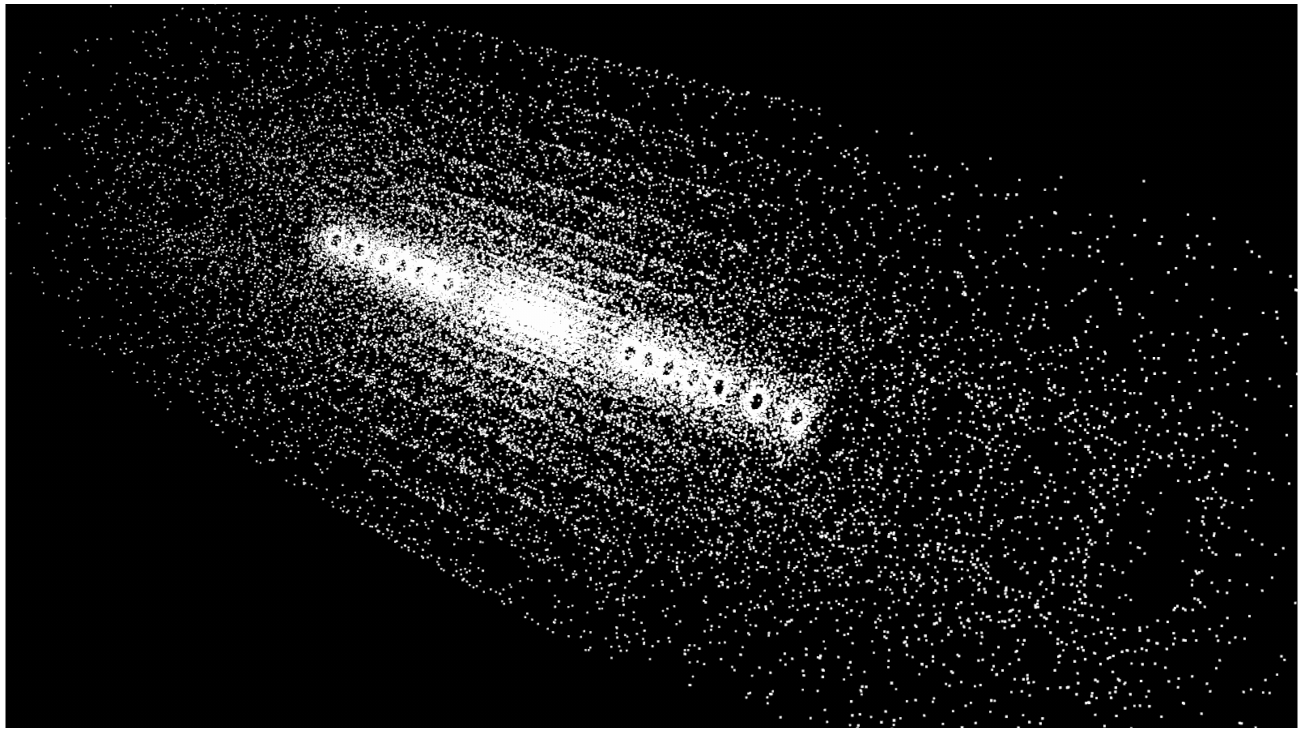
Smeared beam spot

- $\sigma_Z = 5.5$ 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

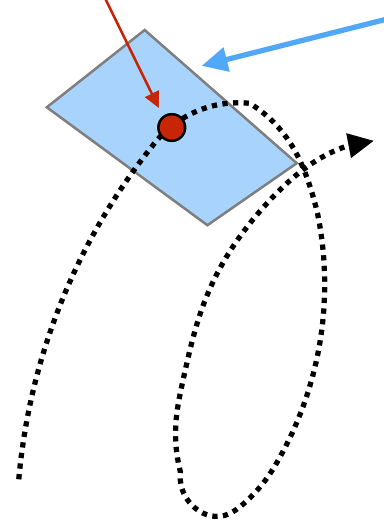


table & images
(top) csv file format for the hit file
(bottom) illustration of the hit information

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

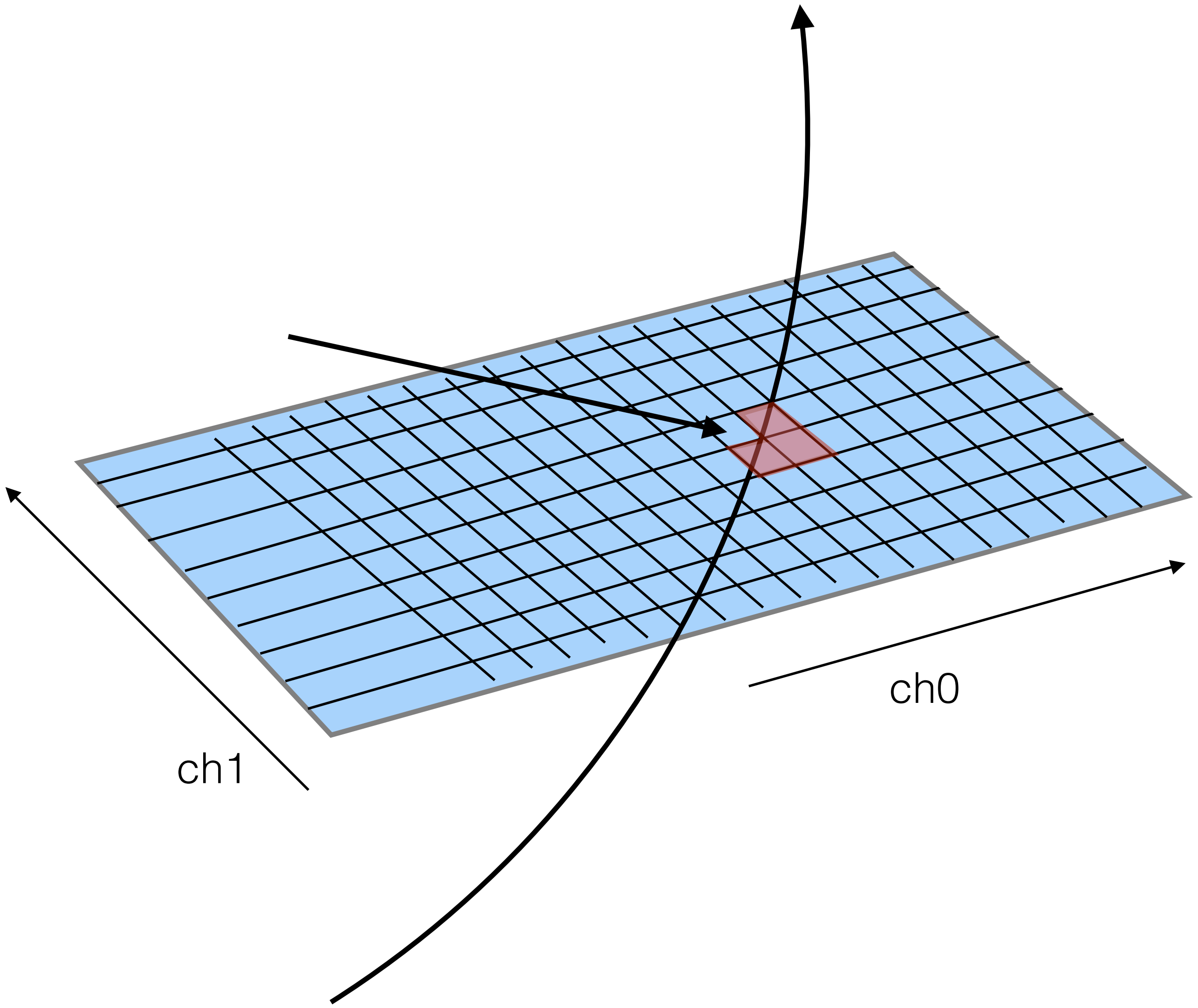


table & images
(top) csv file format for the hit file
(bottom left) csv file format of the cells information
(bottom right) cell information illustration

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

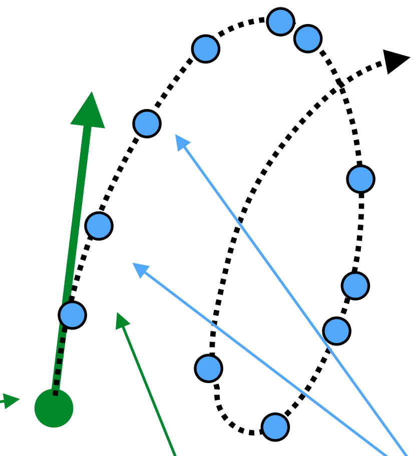
	hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight
0	1	0	-64.411598	-7.164120	-1502.5	250710.000000	-149908.000000	-956385.000000	0.000000
1	2	22525763437723648	-55.338501	0.630805	-1502.5	-0.570605	0.028390	-15.492200	0.000010
2	3	0	-83.828003	-1.145580	-1502.5	626295.000000	-169767.000000	-760877.000000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



charge

	particle_id	vx	vy	vz	px	py	pz	q	nhits
520	22525763437723648	-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**

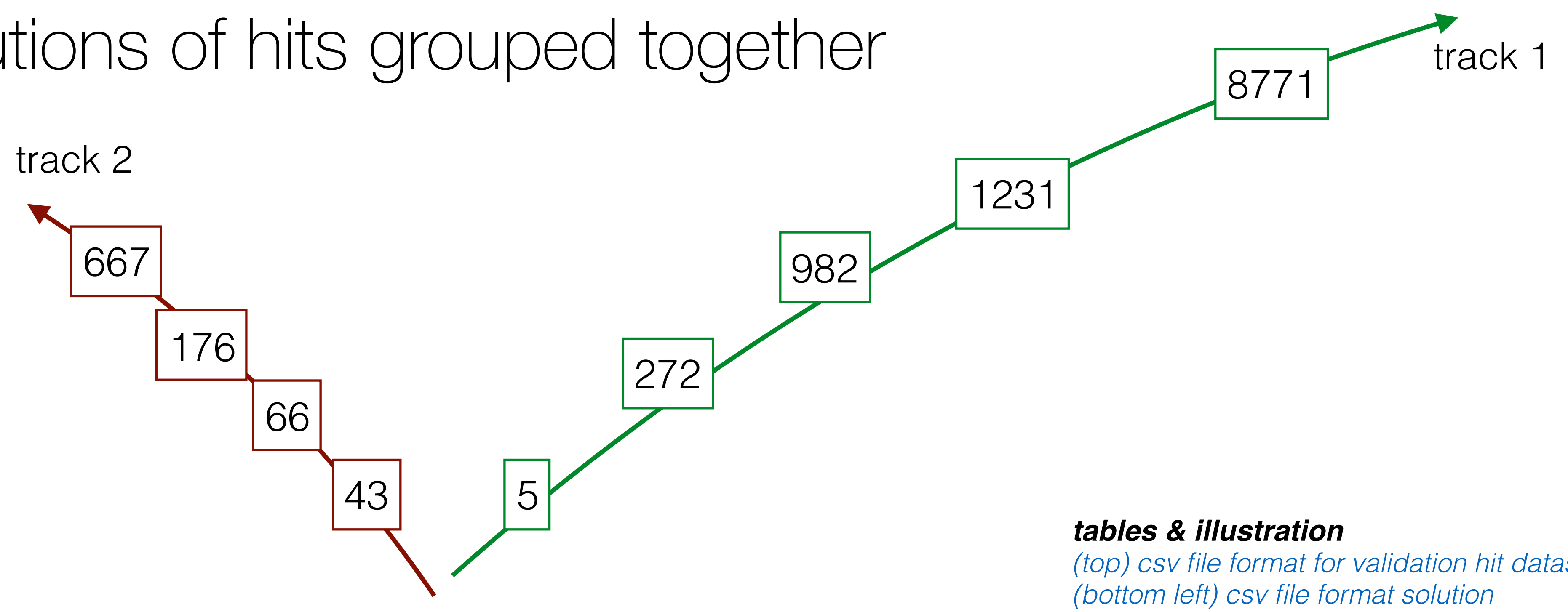
This leaderboard is calculated with approximately 29% of the test data.

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

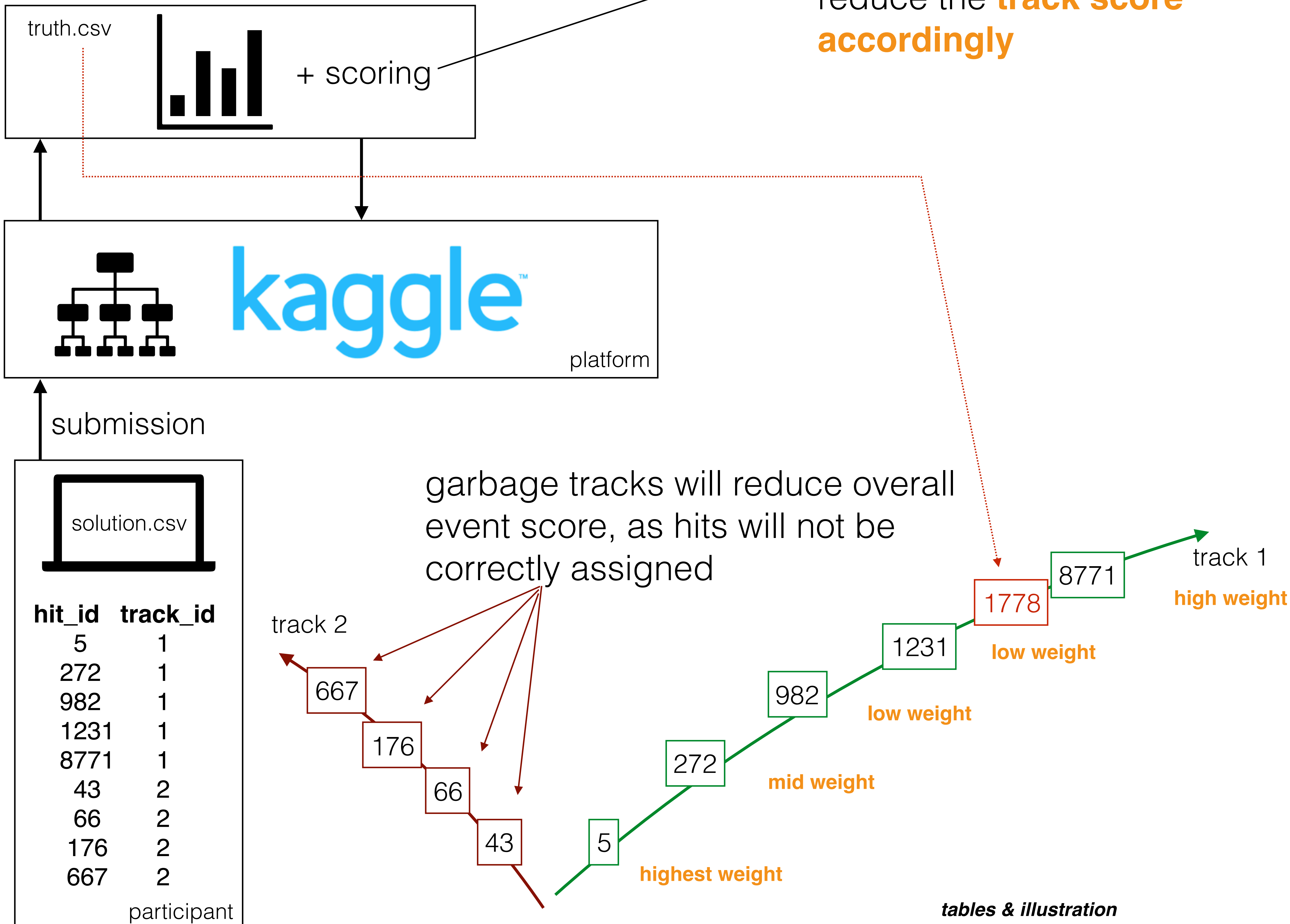


tables & illustration
(top) csv file format for validation hit dataset
(bottom left) csv file format solution
(bottom right) track representation of solutions

Submission & scoring (2)

missing hits reduce the **track score accordingly**

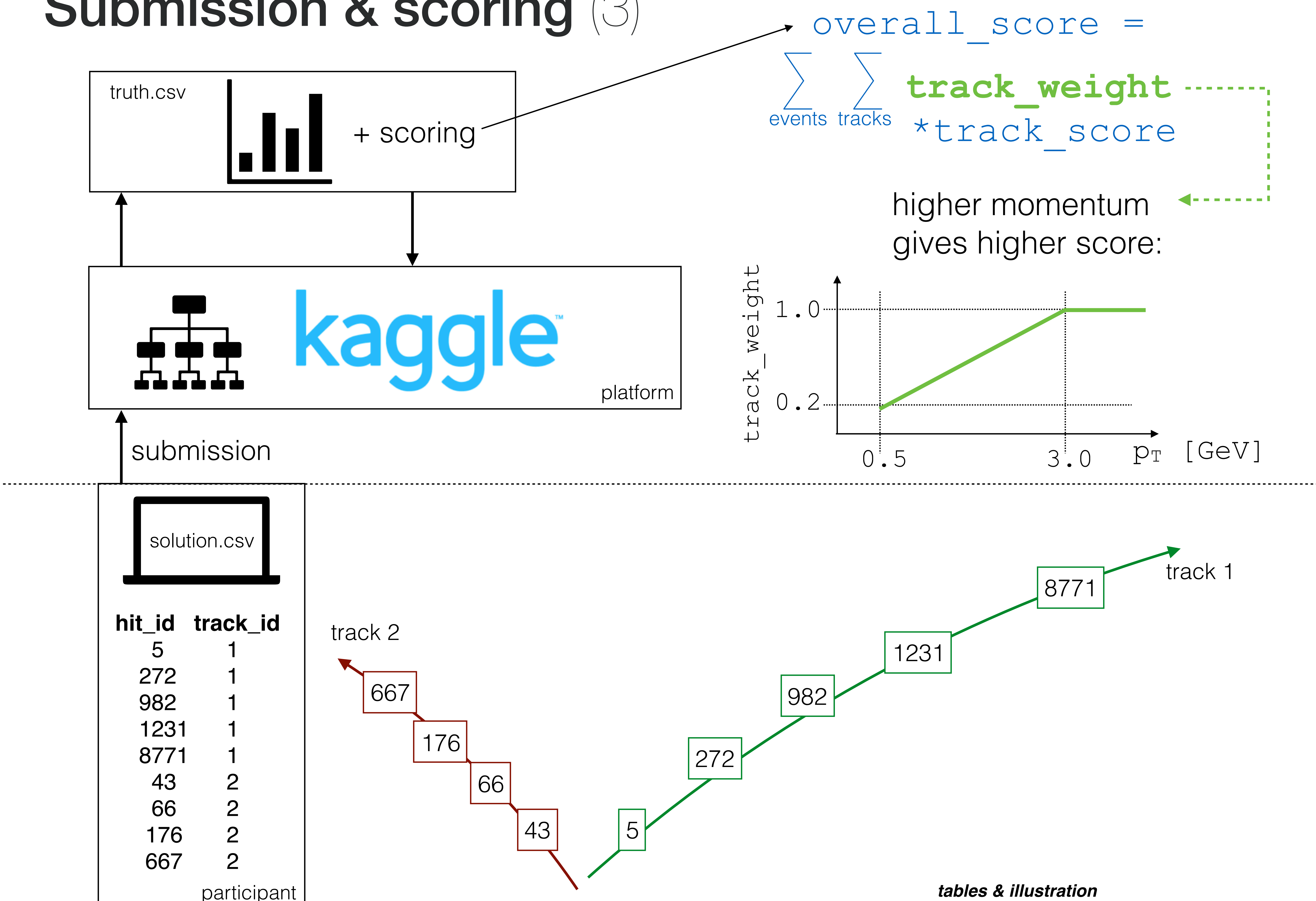
garbage tracks will reduce overall event score, as hits will not be correctly assigned



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

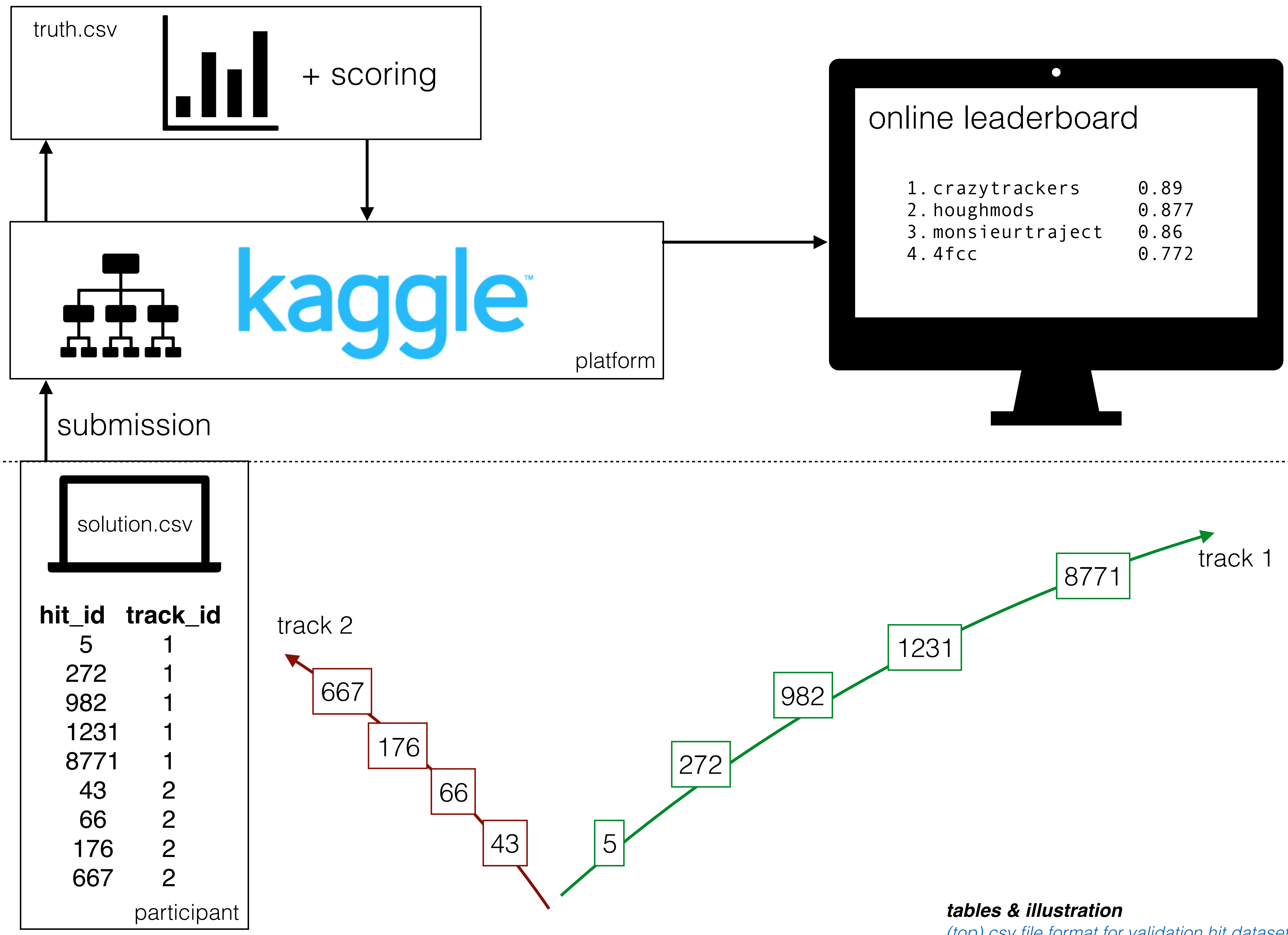
tables & illustration
(top) csv file format for validation hit dataset

Submission & scoring (3)



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

Submission & scoring (4)

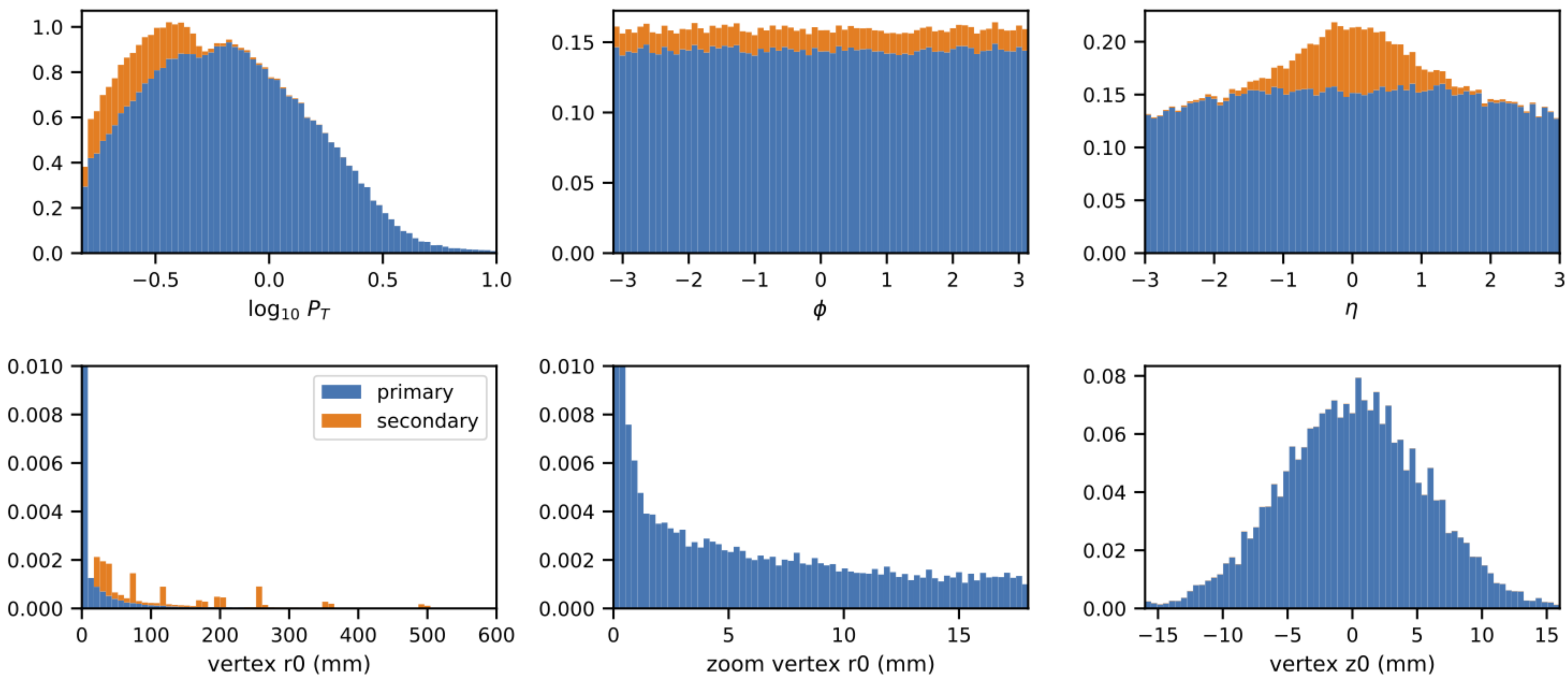


hit_id	track_id
5	1
272	1
982	1
1231	1
8771	1
43	2
66	2
176	2
667	2

tables & illustration

(top) csv file format for validation hit dataset

Phase 1 Dataset - what's there to find



Phase 1 Top Quarks