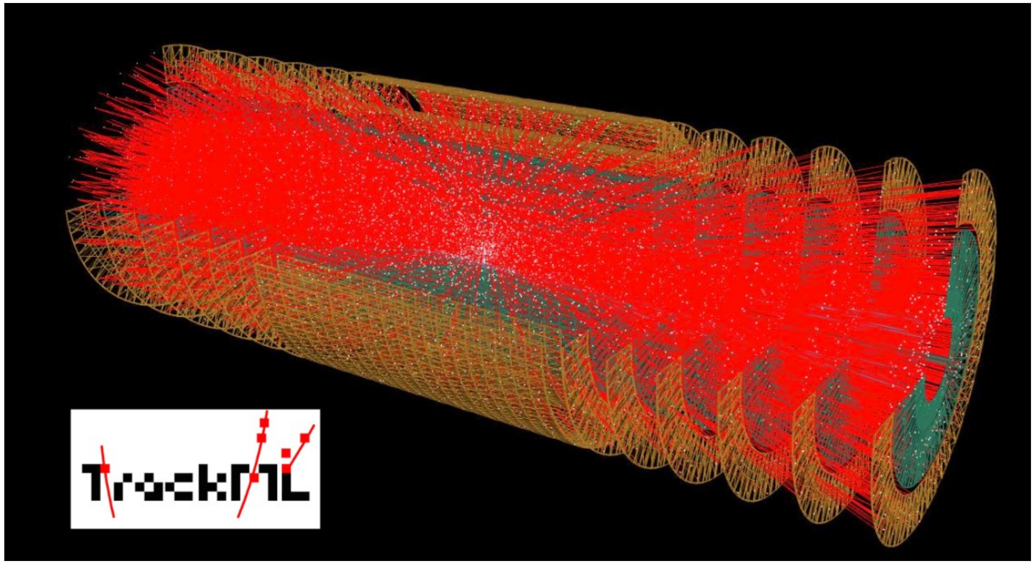
Conclusions from TrackML the HEP Tracking Machine Learning Challenge



Organization

Jean-Roch Vlimant (Caltech)

Vincenzo Innocente, Andreas Salzburger (CERN)

David Rousseau, Yetkin Yilmaz (LAL-Orsay)

Paolo Calafiura, Steven Farrell, Heather Gray (LBNL)

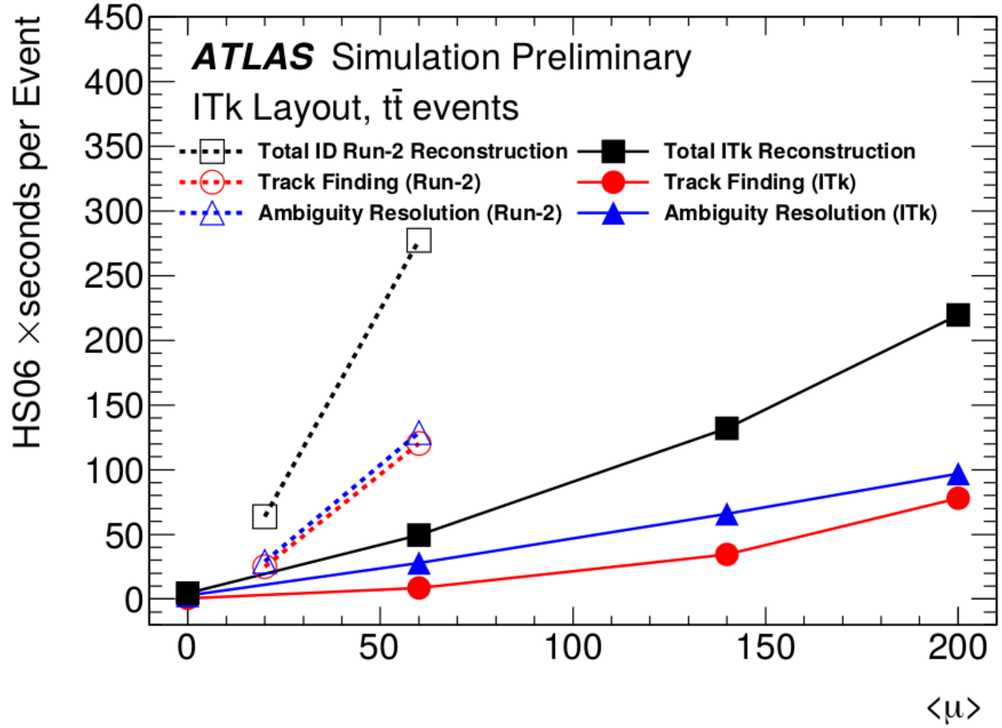
Vladimir Gligorov (LPNHE-Paris)

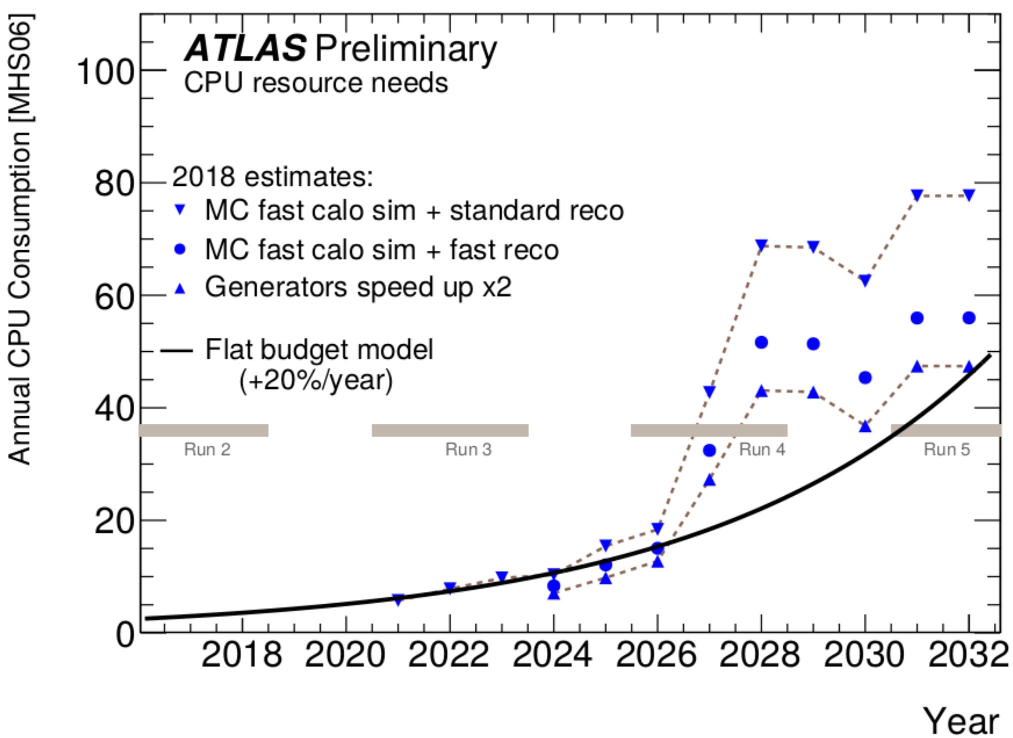
Sabrina Amrouche, Tobias Golling, Moritz Kiehn (Geneva University)

Laurent Basara, Cécile Germain, Isabelle Guyon, Victor Estrade (LRI-Orsay)

Edward Moyse (University of Massachussets)

Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex, HSE)



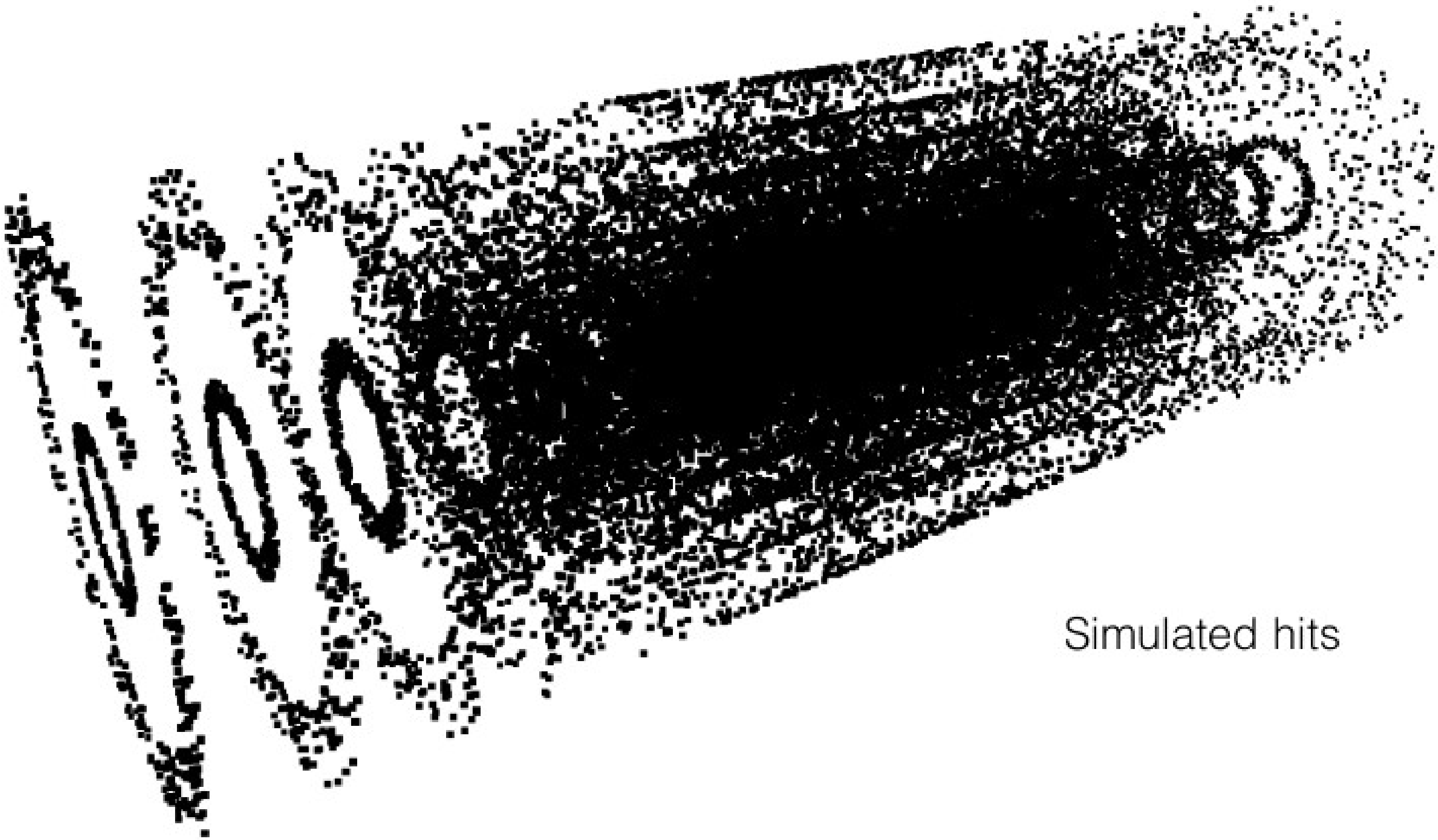


Tracking crisis

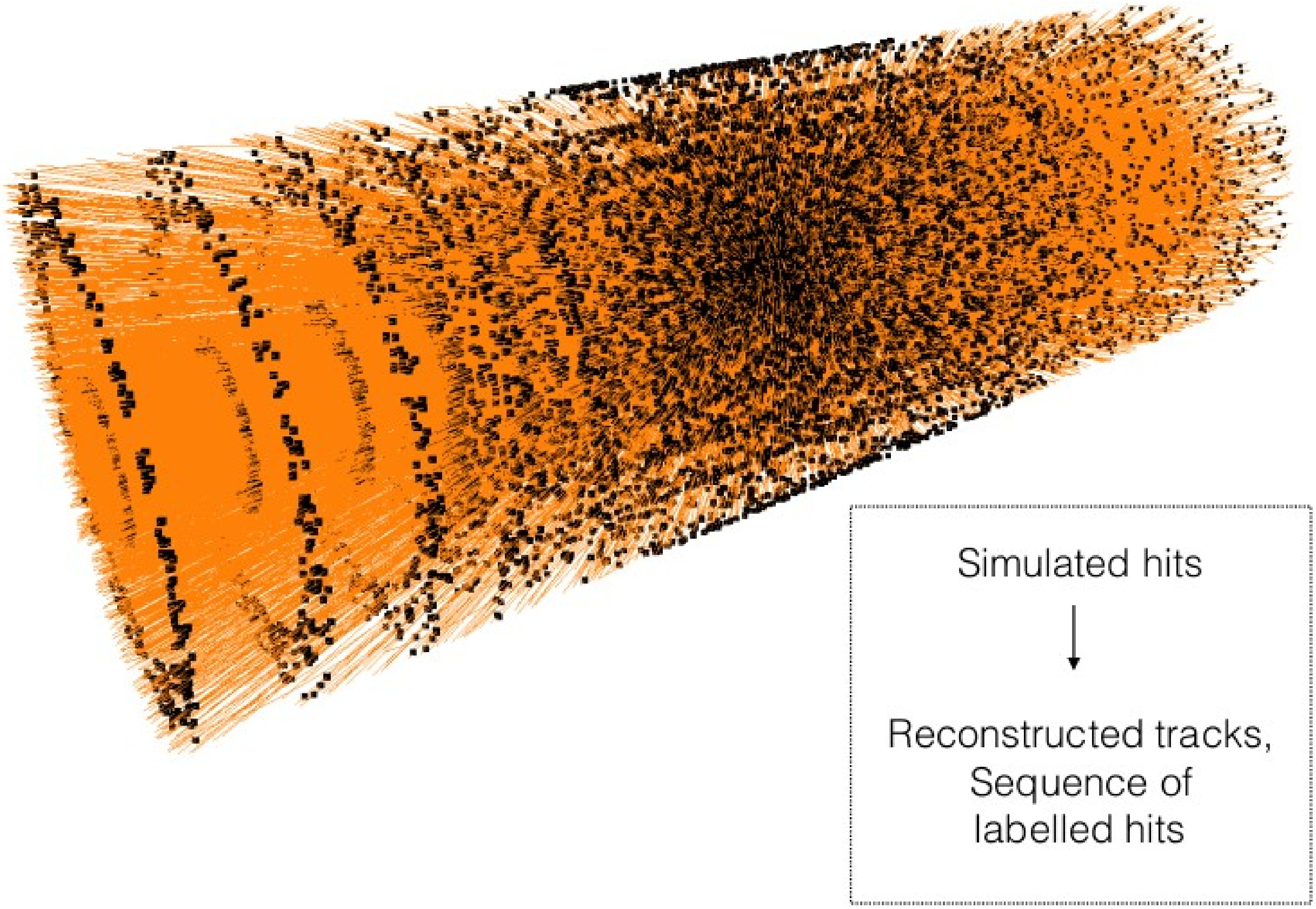
- Tracking dominates reconstruction CPU time
- At best quadratic
- HL-LHC (2025) : unmanageable
- Everything tried?

→ TrackML challenge

TrackML sponsors



See also in outreach session talk by D. Rousseau
« TrackML : the roller coaster of organizing a HEP challenge on Kaggle and Codalab »
https://indico.cern.ch/event/577856/contributions/3423422/



See also in outreach session talk by D. Rousseau
« TrackML : the roller coaster of organizing a HEP challenge on Kaggle and Codalab »
https://indico.cern.ch/event/577856/contributions/3423422/

First phase : Accuracy May – August 2018

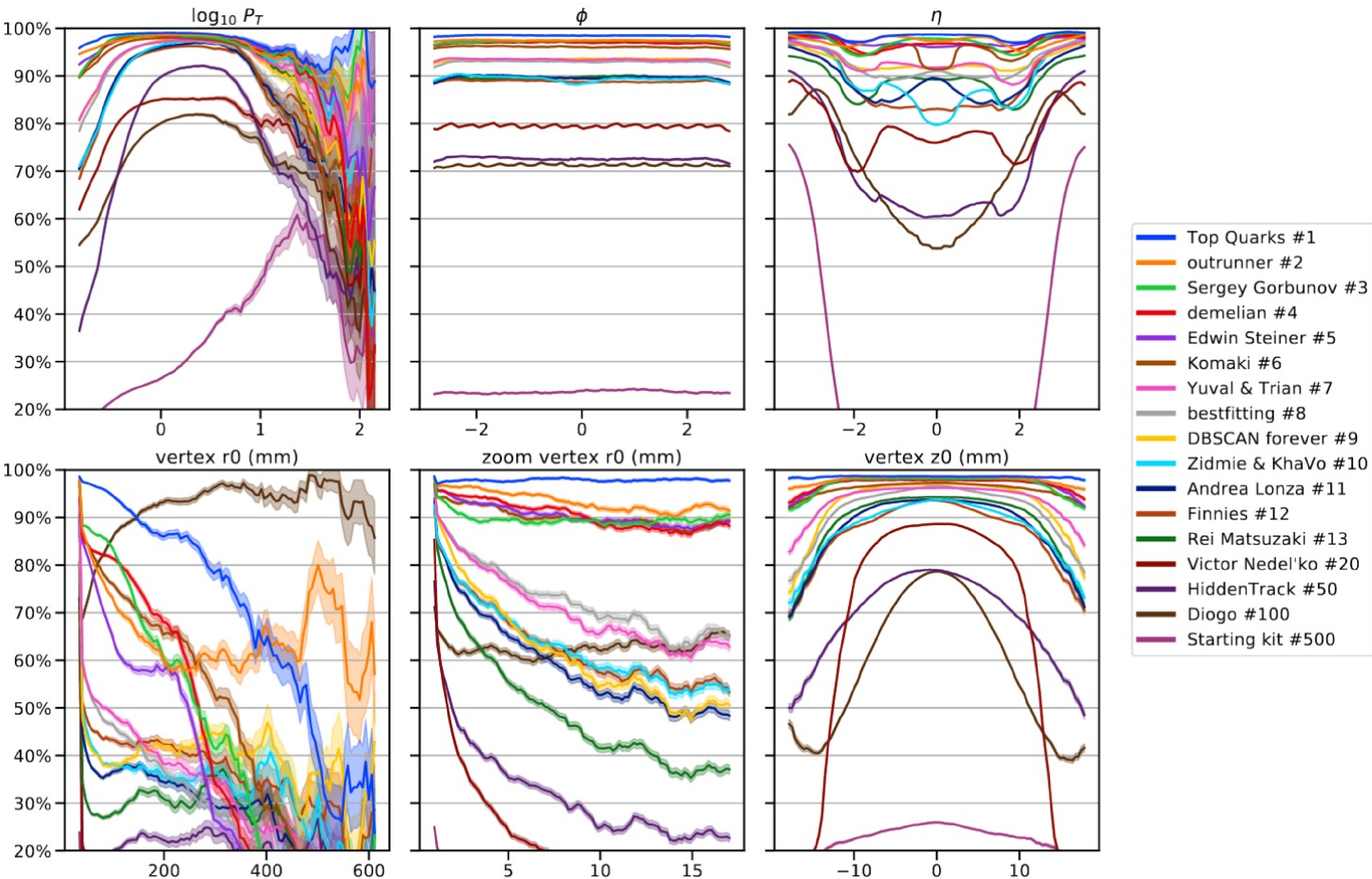
kaggle

Leaderboard scores

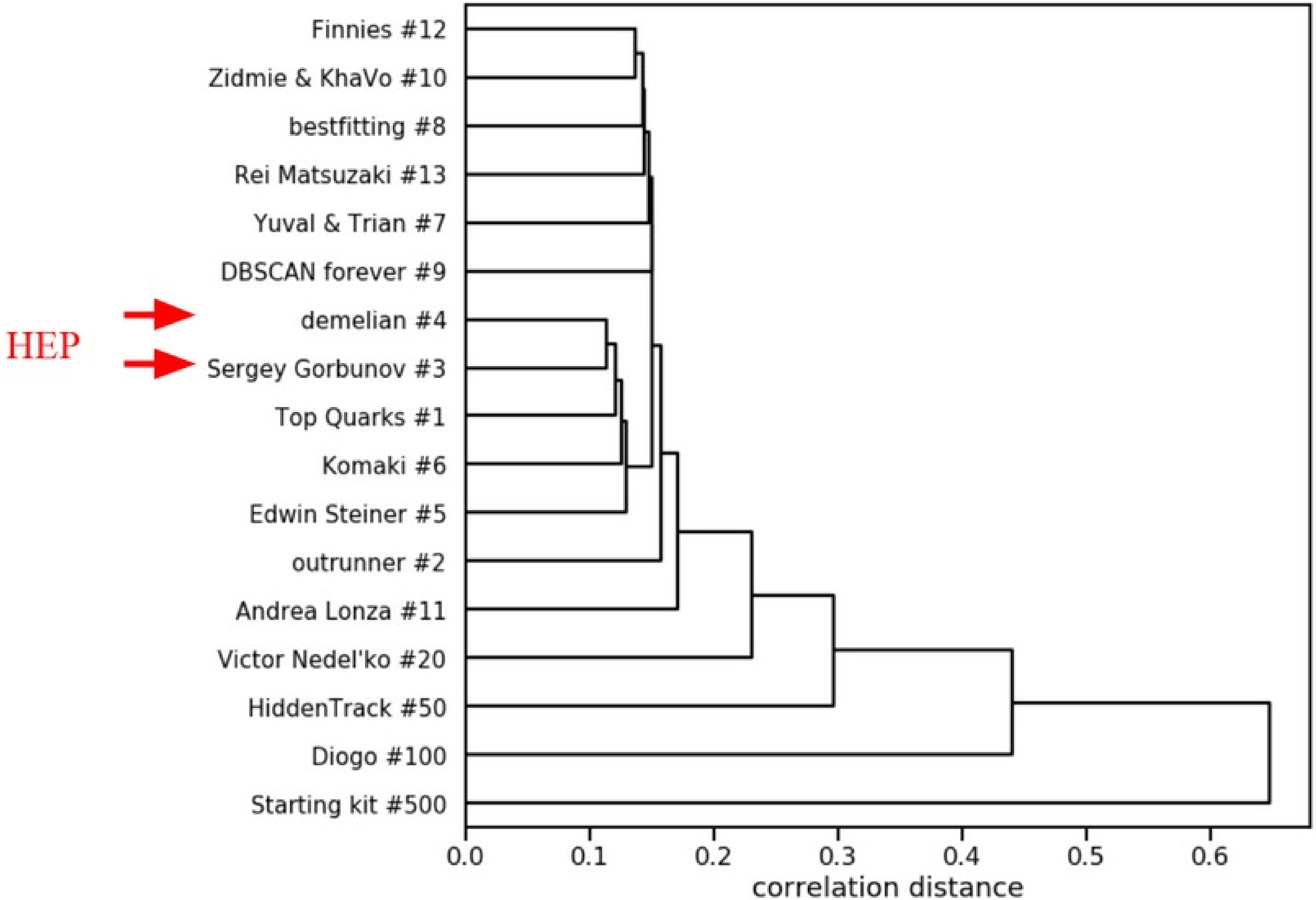
$$Score = \frac{1}{N} \sum_{test\ events} \sum_{good\ hits} weights$$

1	—	Top Quarks	0.92182	10	19d
2	—	outrunner	0.90302	9	18d
3	—	Sergey Gorbunov	0.89353	6	18d
4	—	demelian	0.87079	35	1mo
5	—	Edwin Steiner	0.86395	5	18d
6	—	Komaki	0.83127	22	18d
7	—	Yuval & Trian	0.80414	56	18d
8	—	bestfitting	0.80341	6	18d
9	—	DBSCAN forever	0.80114	23	18d
10	—	Zidmie & KhaVo	0.76320	26	18d
11	—	Andrea Lonza	0.75845	15	18d
12	—	Finnies	0.74827	56	18d
13	—	Rei Matsuzaki	0.74035	12	18d
14	—	Mickey	0.73217	10	2mo
15	—	Vicens Gaitan	0.70429	19	1mo
16	—	Robert	0.69955	3	21d
17	—	Yuval-CPMP tribute band	0.69364	20	20d
18	—	N. Hi. Bouzu	0.67573	9	22d
19	—	Steins;Gate	0.66763	12	19d
20	▲1	Victor Nedel'ko	0.66723	4	2mo

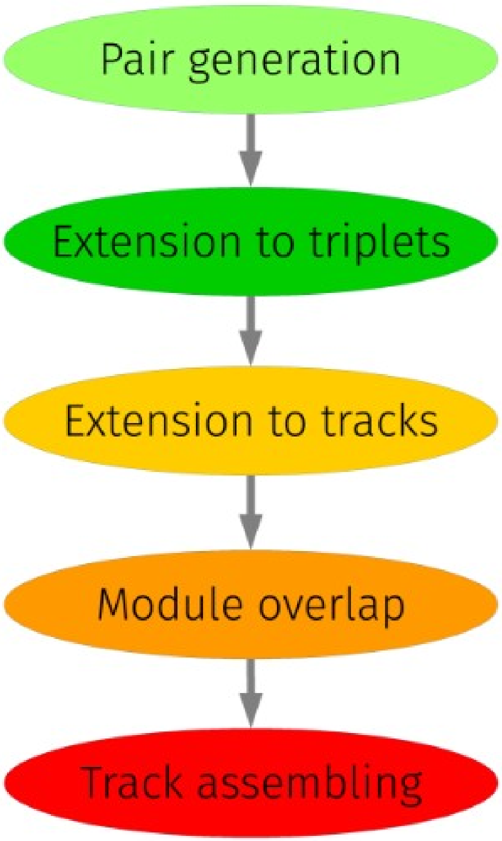
Optimizing score optimizes physics



Participants dendrogram



Phase 1 winner : Top Quarks



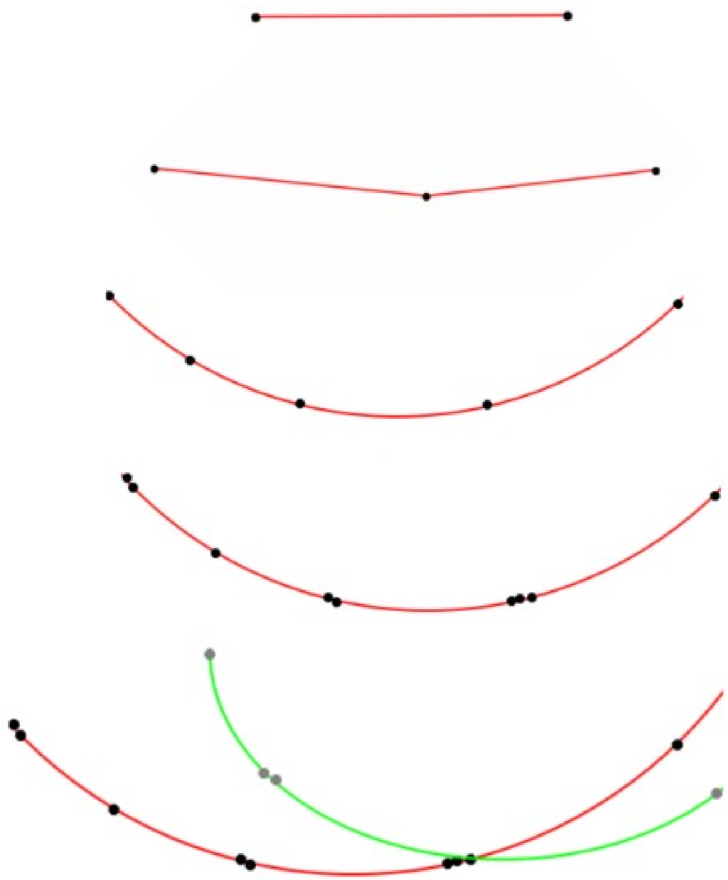
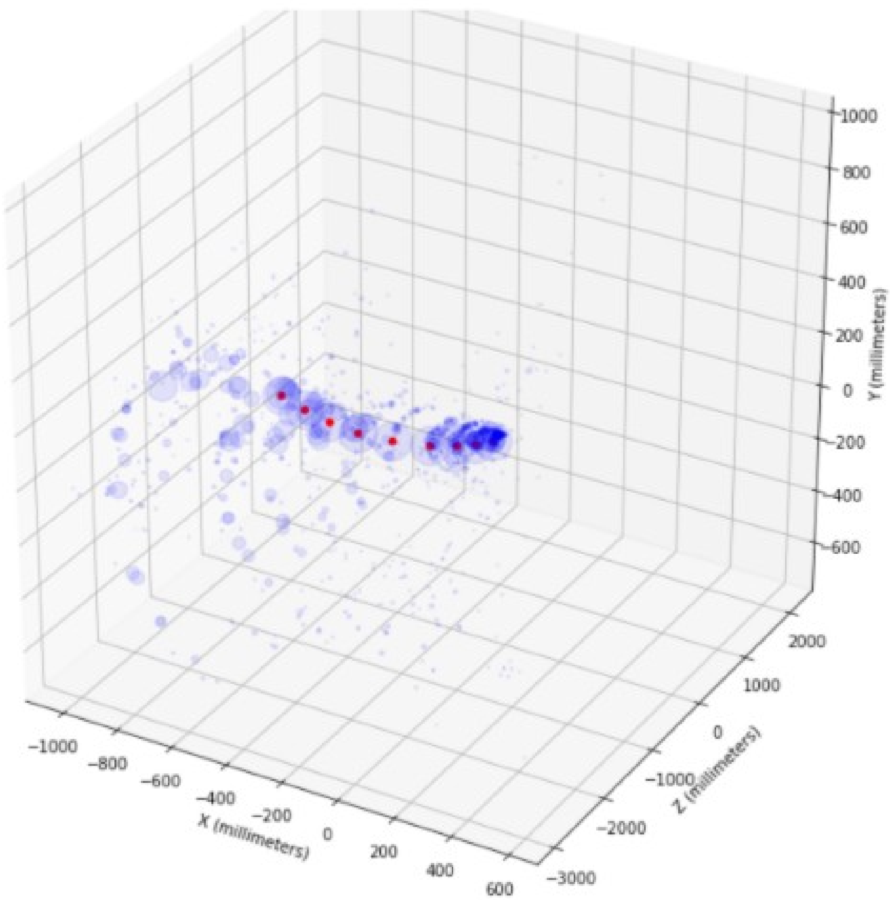


Illustration from J-R. Vlimant

Illustration from J.S. Wind

Phase 1 #2 : outrunner



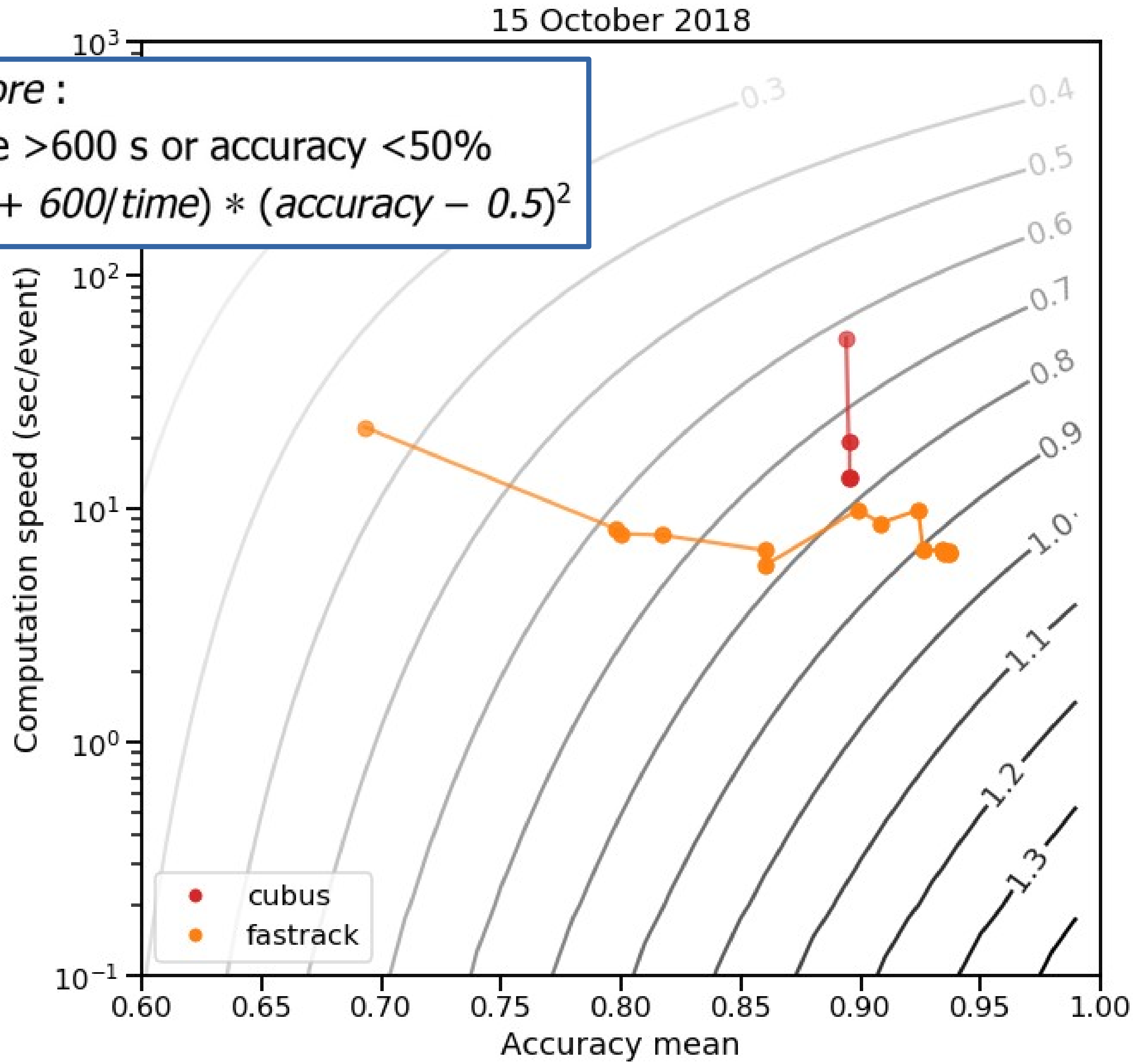
- Train DNN on hit pairs
  - 27 inputs (x,y,z,cells,…)
  - 4k-2k-2k-2k-1k hidden layers
- Compute full hit adjacency matrix:
  - probability P(i,j) that 2 hits match
  - Pick high probability comb
- True Deep Learning Solution
  - No track following
  - No geometric modelling
- 1 Day / event

Second phase : Throughput Oct 2018 – March 2019

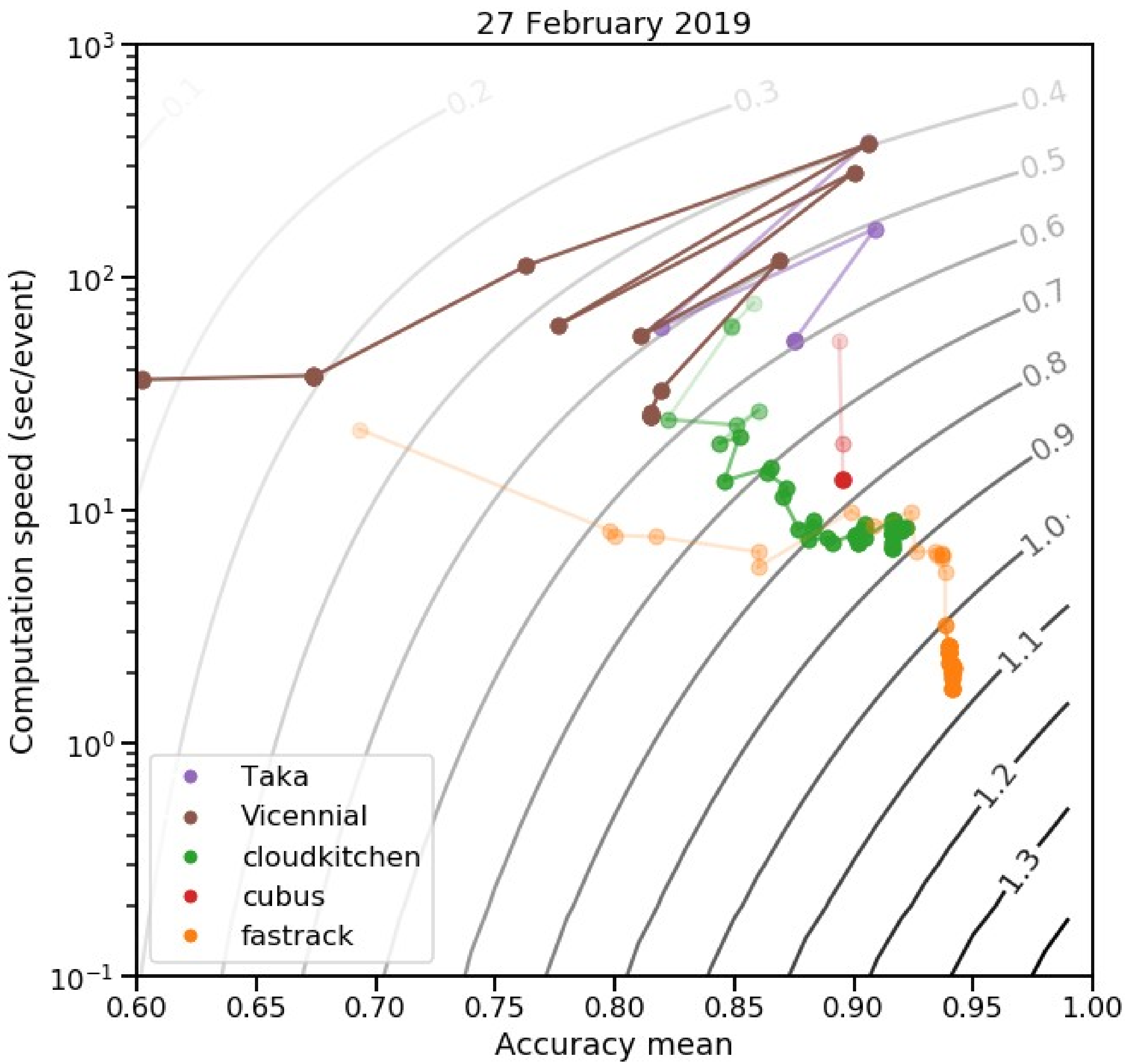
CodaLab

Leaderboard evolution

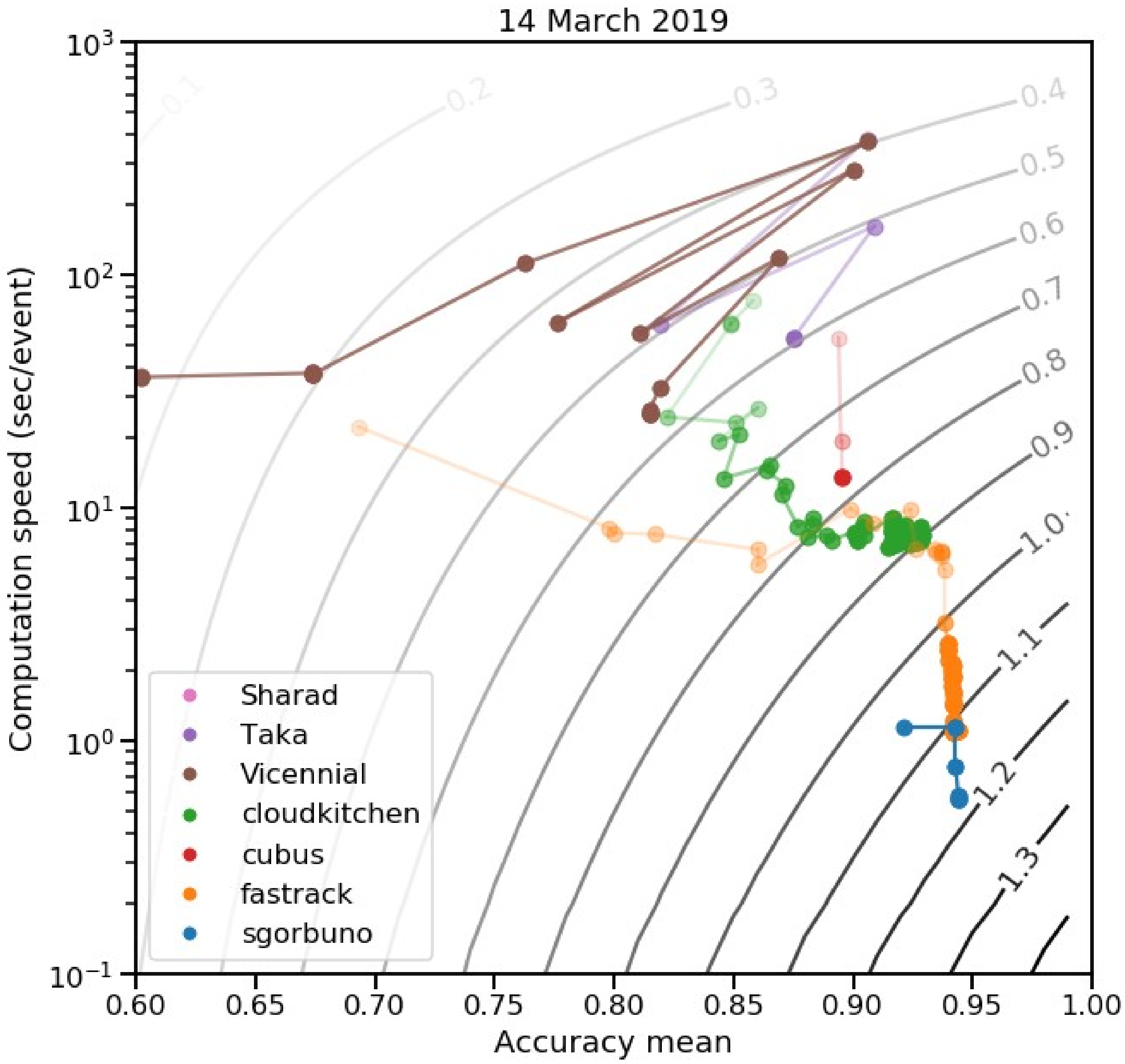
- *Ranking score* :
  - 0 if time >600 s or accuracy <50%
  - $\sqrt{log(1 + 600/time)} * (accuracy - 0.5)^2$



Leaderboard evolution



Leaderboard evolution

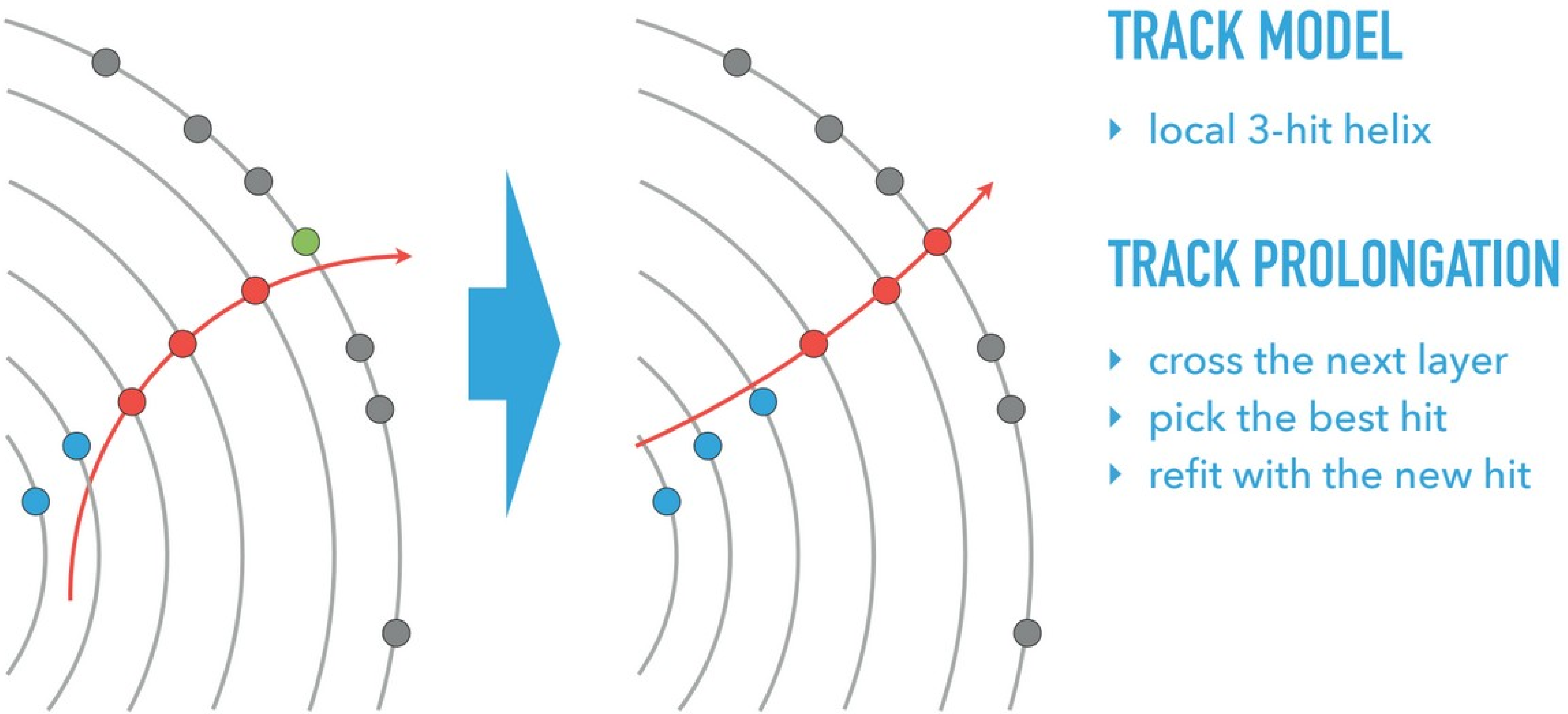


Final leaderboard

RESULTS

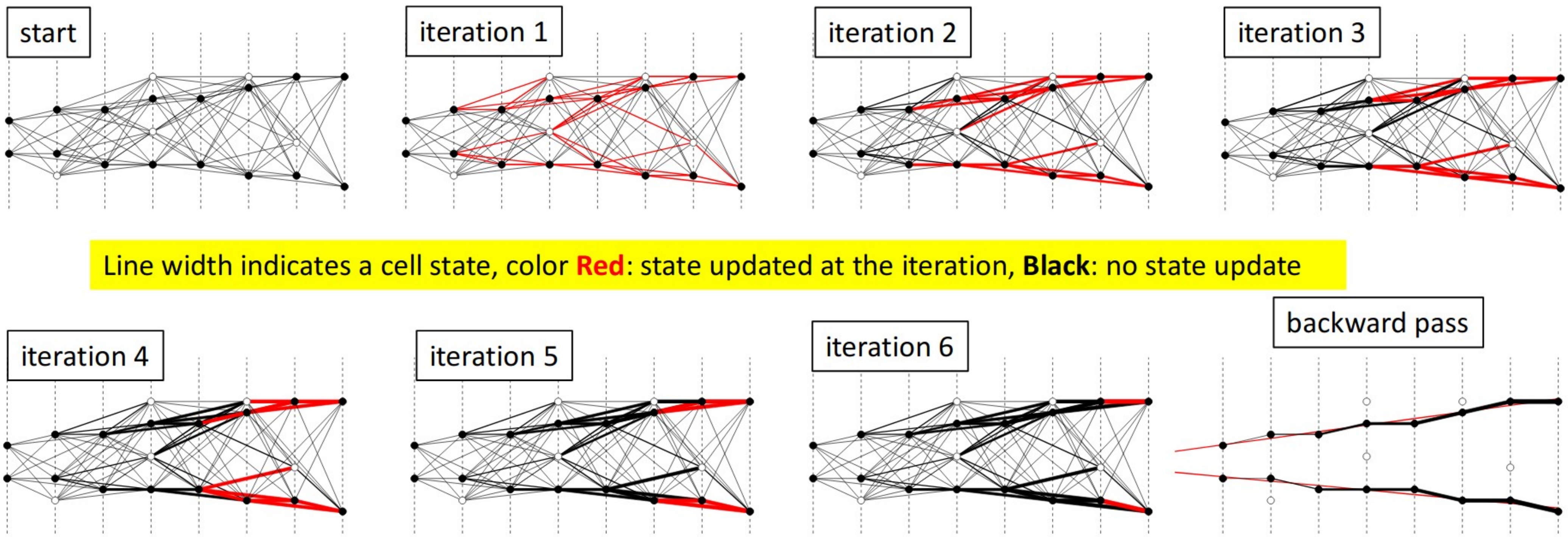
#	User	Entries	Date of Last Entry	score ▲	accuracy_mean ▲	accuracy_std ▲	computation time (sec) ▲	computation speed (sec/event) ▲	Duration ▲
1	sgorbuno HEP	9	03/12/19	1.1727 (1)	0.944 (2)	0.00 (14)	28.06 (1)	0.56 (1)	64.00 (1)
2	fastrack HEP	53	03/12/19	1.1145 (2)	0.944 (1)	0.00 (15)	55.51 (16)	1.11 (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)

#1 S. Gorbunov: « fast combinatorial »



https://indico.cern.ch/event/813759/contributions/3479705/attachments/1871795/3080436/trackML_Gorbunov-CERN2019.pdf

#2 FASTrack: Graph of neighbours, cellular automata and Kalman filter



https://indico.cern.ch/event/813759/contributions/3479706/attachments/1870758/3078234/TheTrackML_workshop_talk.pdf

Phase 2 cloudkitchen

Author: **Marcel Kunze**

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

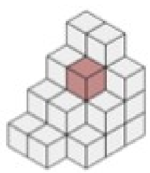
partly based on top quarks Phase 1 solution

1	—	**Top Quarks**	0.92182	10	2mo

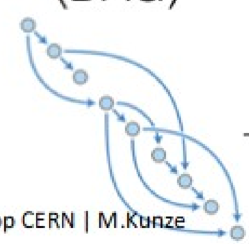
Algorithm outline

hits

sorted in voxels

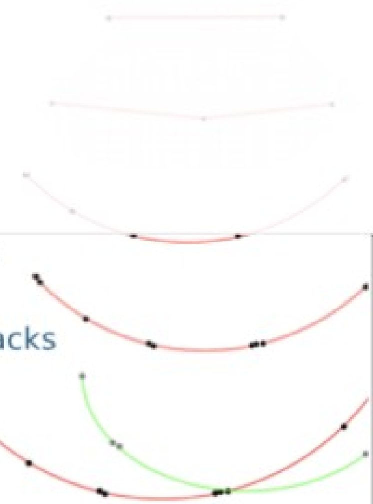


organised in
direct acyclic graphs
(DAG)



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



Triplet finder

NN3 — ca. 300k, 97.2%

doublet finder

NN1 NN2 — ca. 500k, 99.4%

DAGs are pre-trained on ~25 events ground truth

DAGs are used to fast navigate through voxel space → Disc section / Tube section (Threaded) — ca. 2 Mio.

https://indico.cern.ch/event/813759/contributions/3479709/attachments/1871944/3080707/TrackML.Kunze.pdf

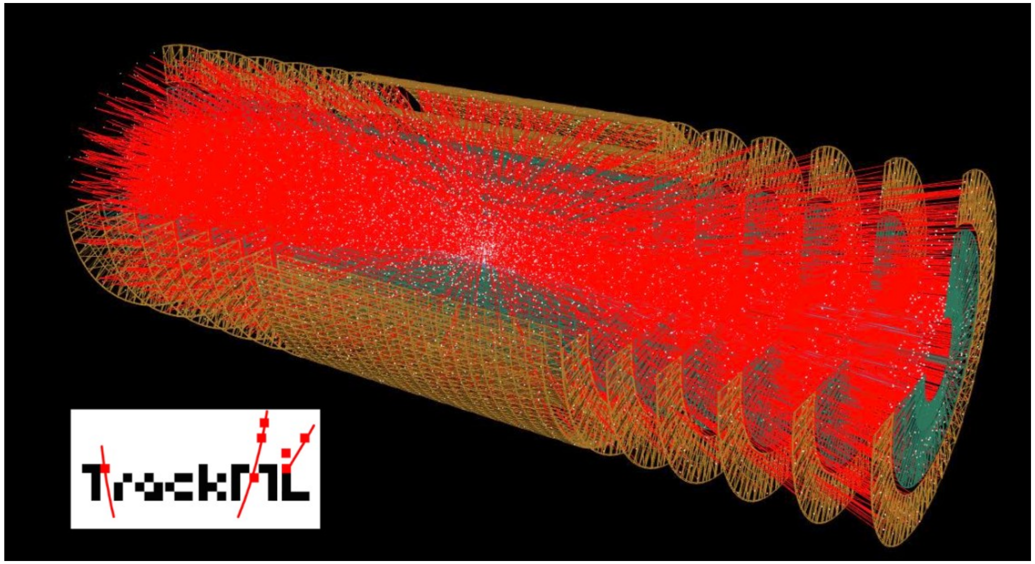
Conclusions

- Open tracking competition organised
  to reach out to CS and ML communities
- Winner and runner-up HEP tracking experts…
- Retained solution will be blend
  from HEP expertise and new ideas
- Dataset released on CERN Open Data Portal
  to serve as benchmark
- Ongoing work

Contacts

- Contact : trackml.contact@gmail.com
- https://sites.google.com/site/trackmlparticle
- Twitter : @trackmllhc
- Accuracy phase @ Kaggle: https://www.kaggle.com/c/trackml-particle-identification
    - Chapter in the NeurIPS2018 Competition book arXiv:1904.06778
- Throughput phase @ Codalab: https://competitions.codalab.org/competitions/20112
    - Write-up to be finalised
- TrackML challenge Grand Finale: https://indico.cern.ch/event/813759/

Conclusions from TrackML the HEP Tracking Machine Learning Challenge



Organization

Jean-Roch Vlimant (Caltech)

Vincenzo Innocente, Andreas Salzburger (CERN)

David Rousseau, Yetkin Yilmaz (LAL-Orsay)

Paolo Calafiura, Steven Farrell, Heather Gray (LBNL)

Vladimir Gligorov (LPNHE-Paris)

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

Laurent Basara, Cécile Germain, Isabelle Guyon, Victor Estrade (LRI-Orsay)

Edward Moyse (University of Massachussets)

Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex, HSE)

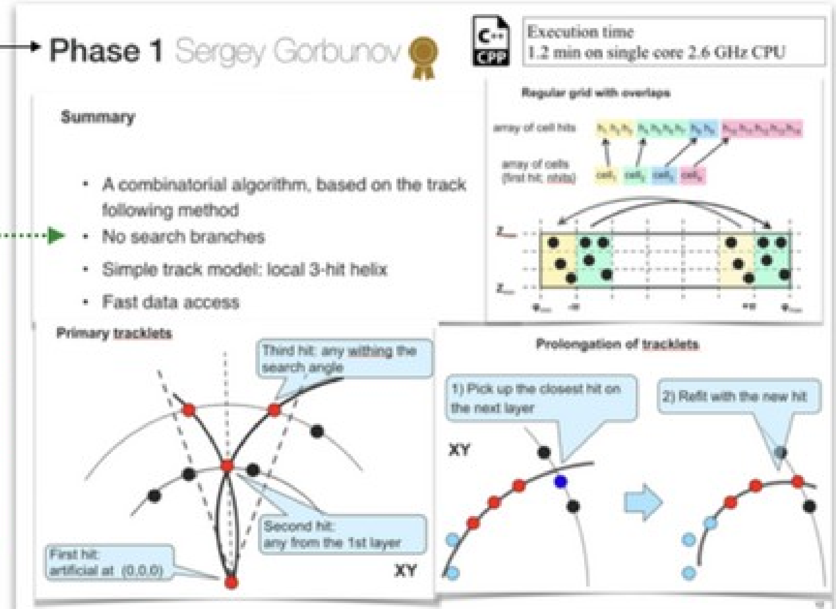
Phase 2 *Mikado*

Author: Sergey Gorbunov

C++ CPP

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 recently down to
Memory: 0.6 Gb

first runner-up to podium in Phase-1

4	—	demelian		0.87079	35	2mo

Phase-1 w/o measurement shapes

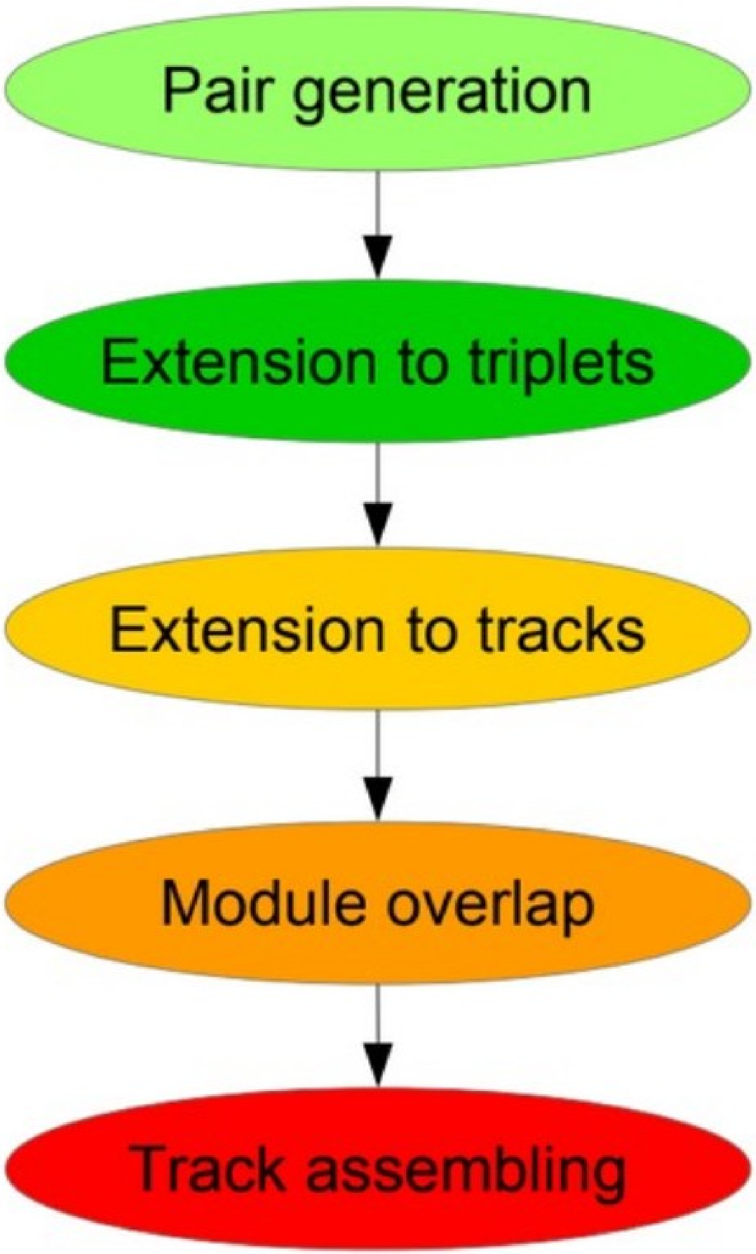
Algorithm outline

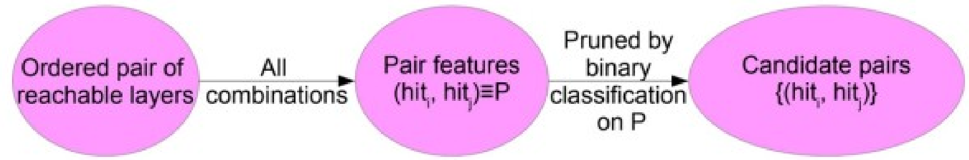
- using measurement shapes to predict intervals of track inclination
- segment based track following network with embedded Kalman Filter
  - **connection graph** pre-build (&compiled) from `Detector.csv` file
  - run with a **Cellular Automaton (CA), parallelised** with **OpenMP**
  - **candidate building:** 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

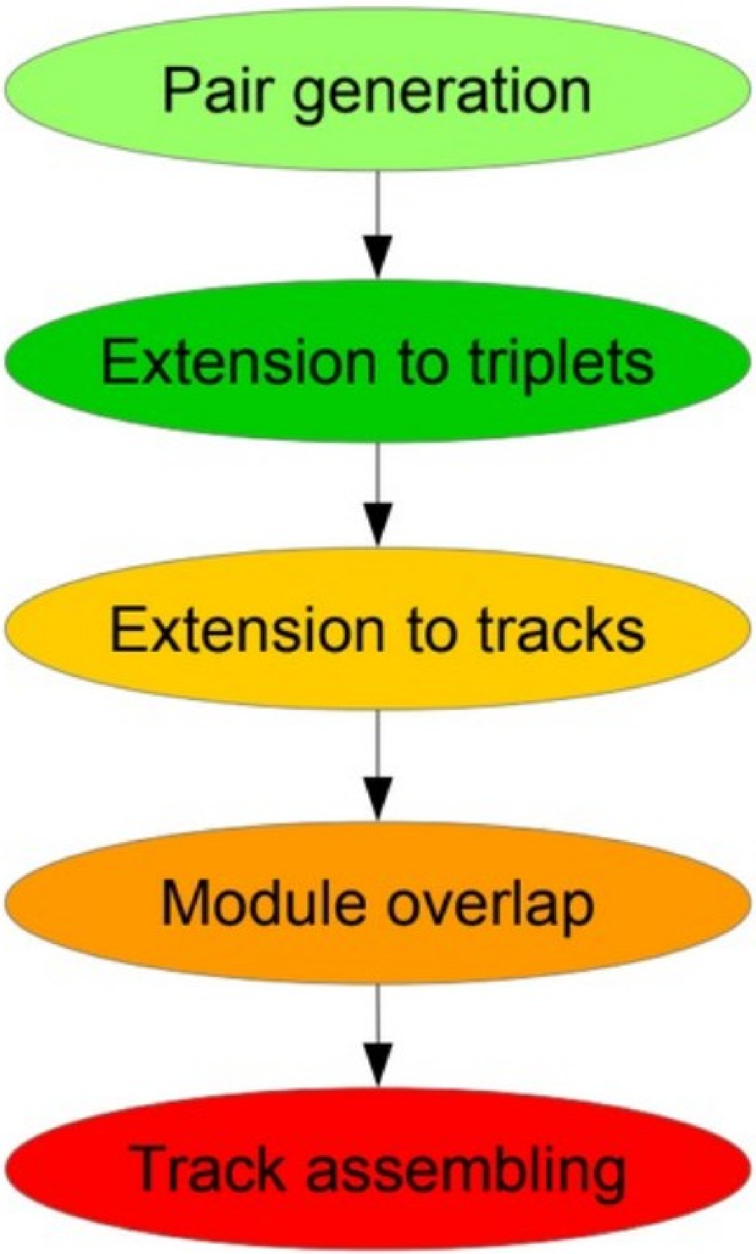
TrackML, David Rousseau, IML, 16th April 2019

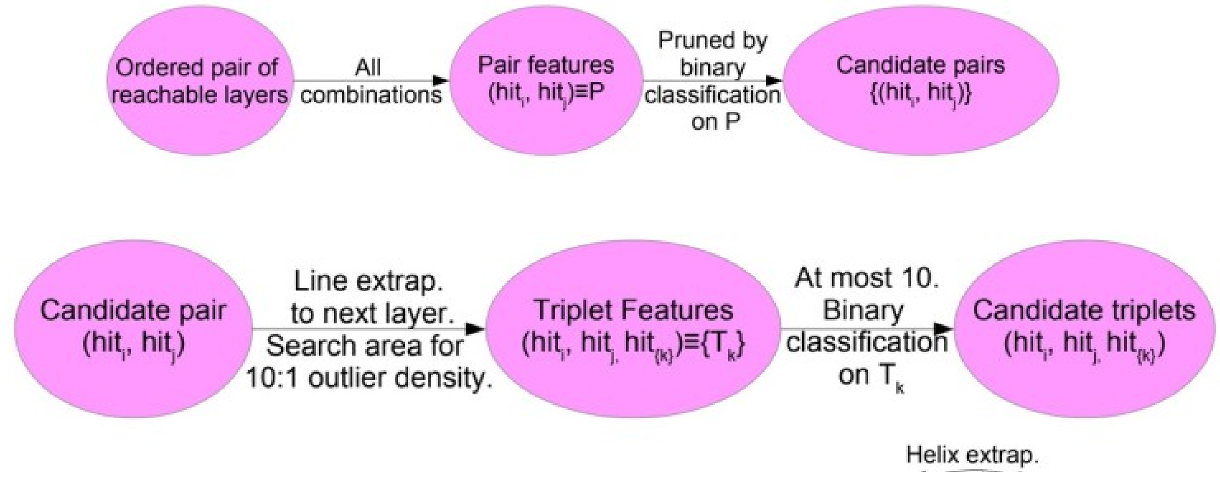
Winning solution → not ML



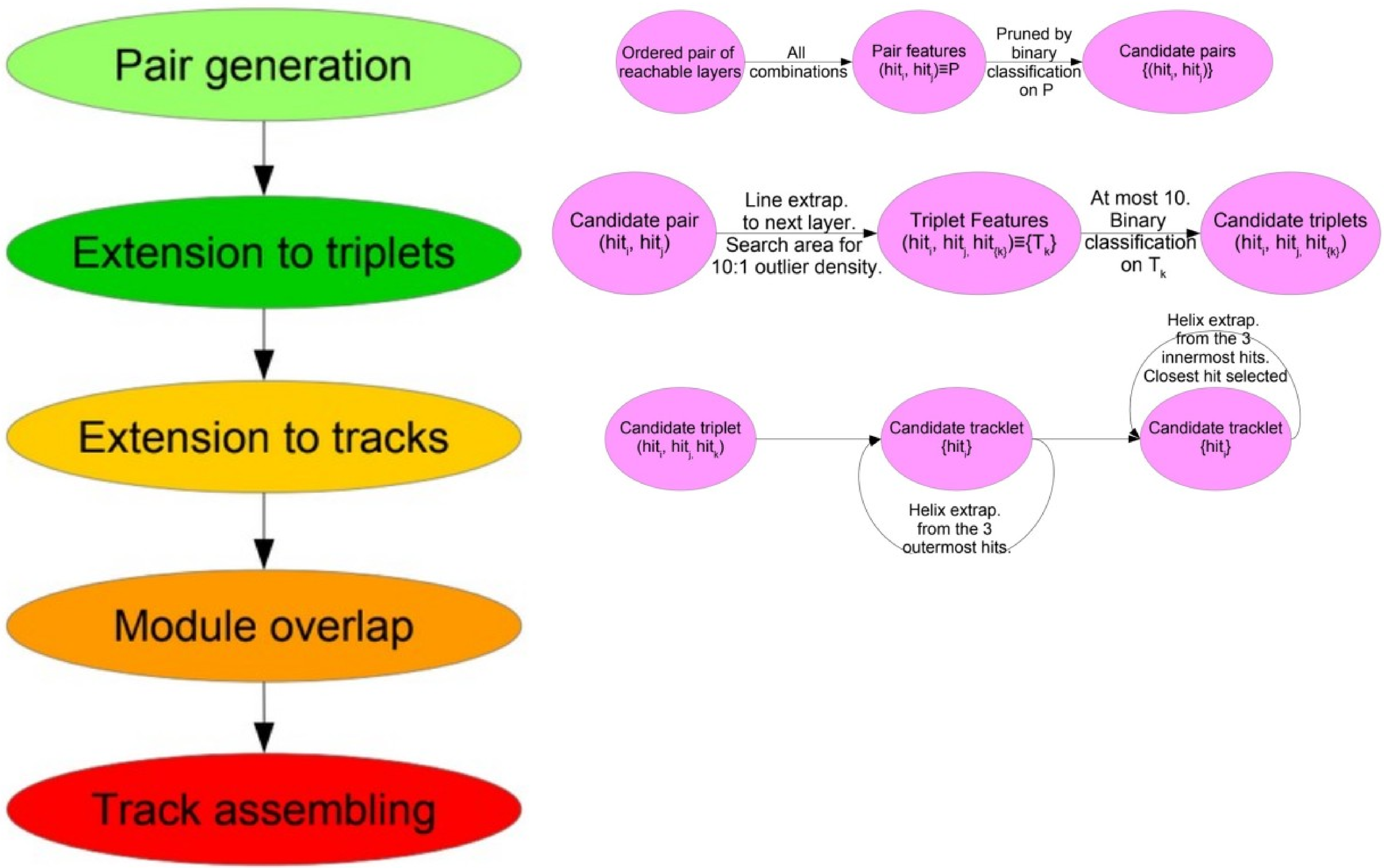


Winning solution → not ML

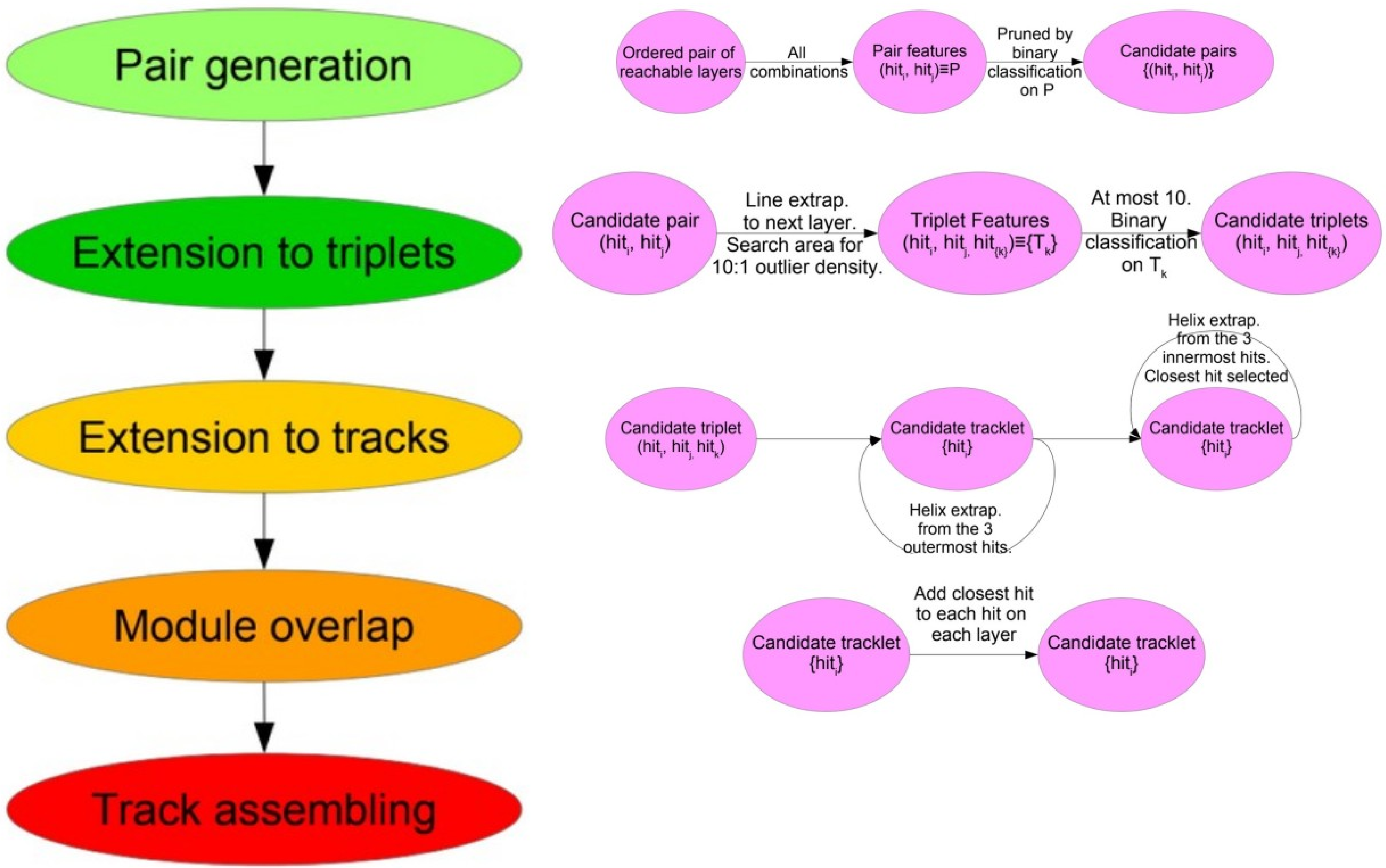




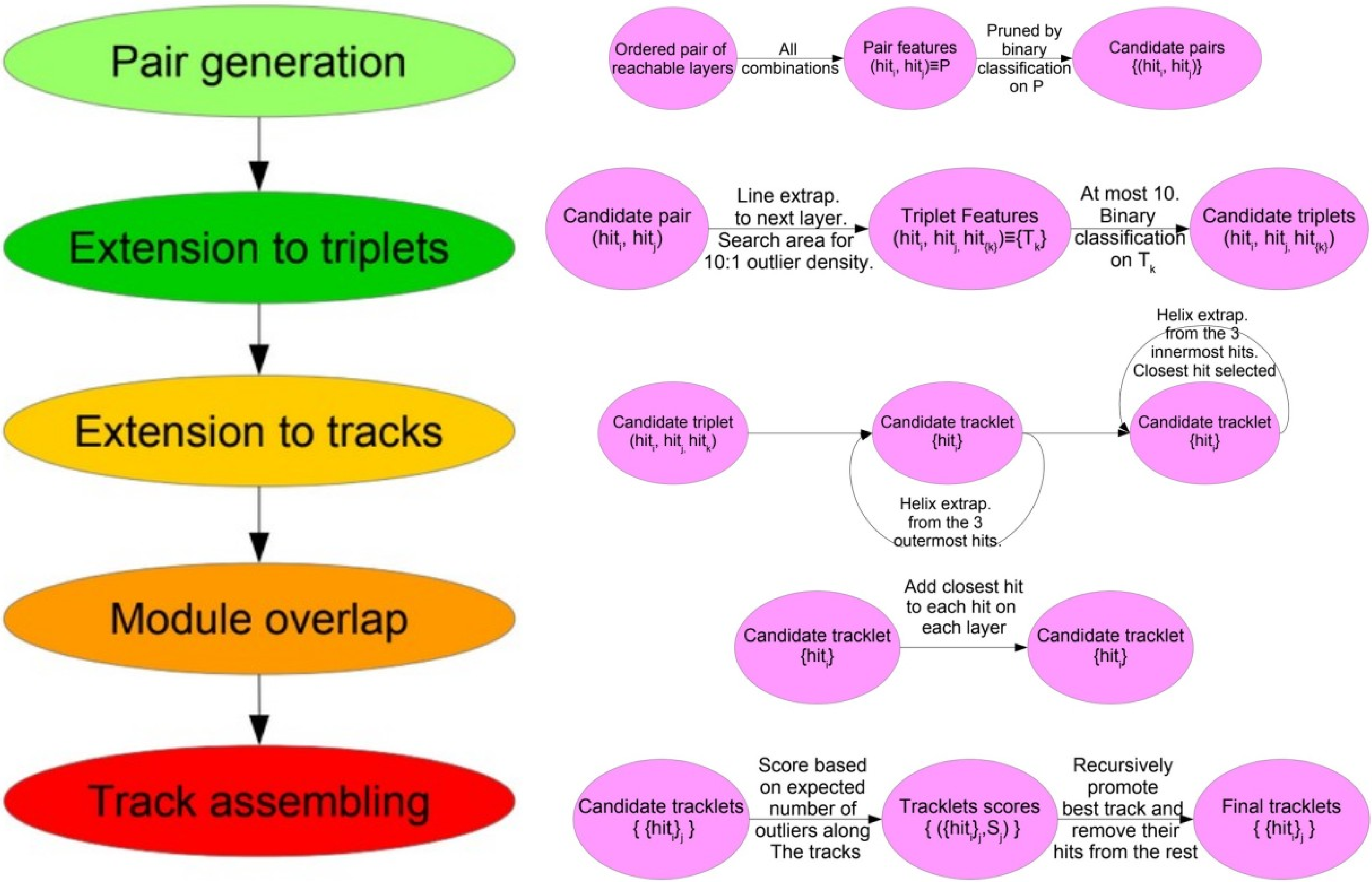
Winning solution → not ML



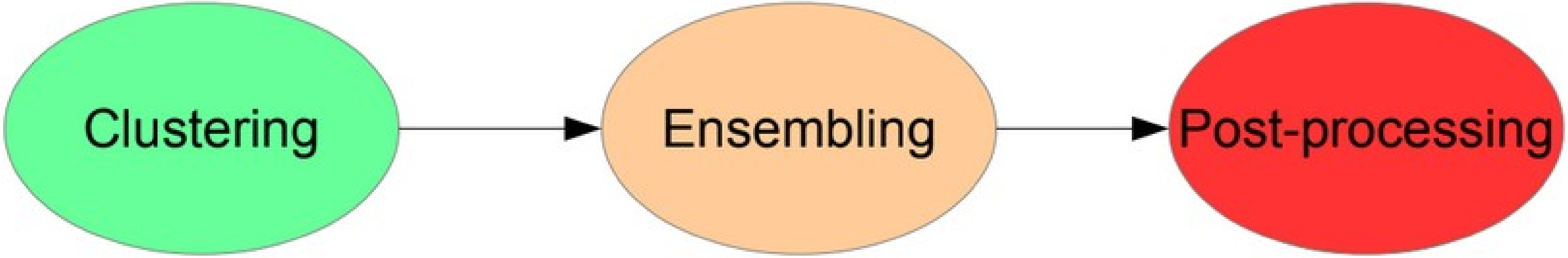
Winning solution → not ML

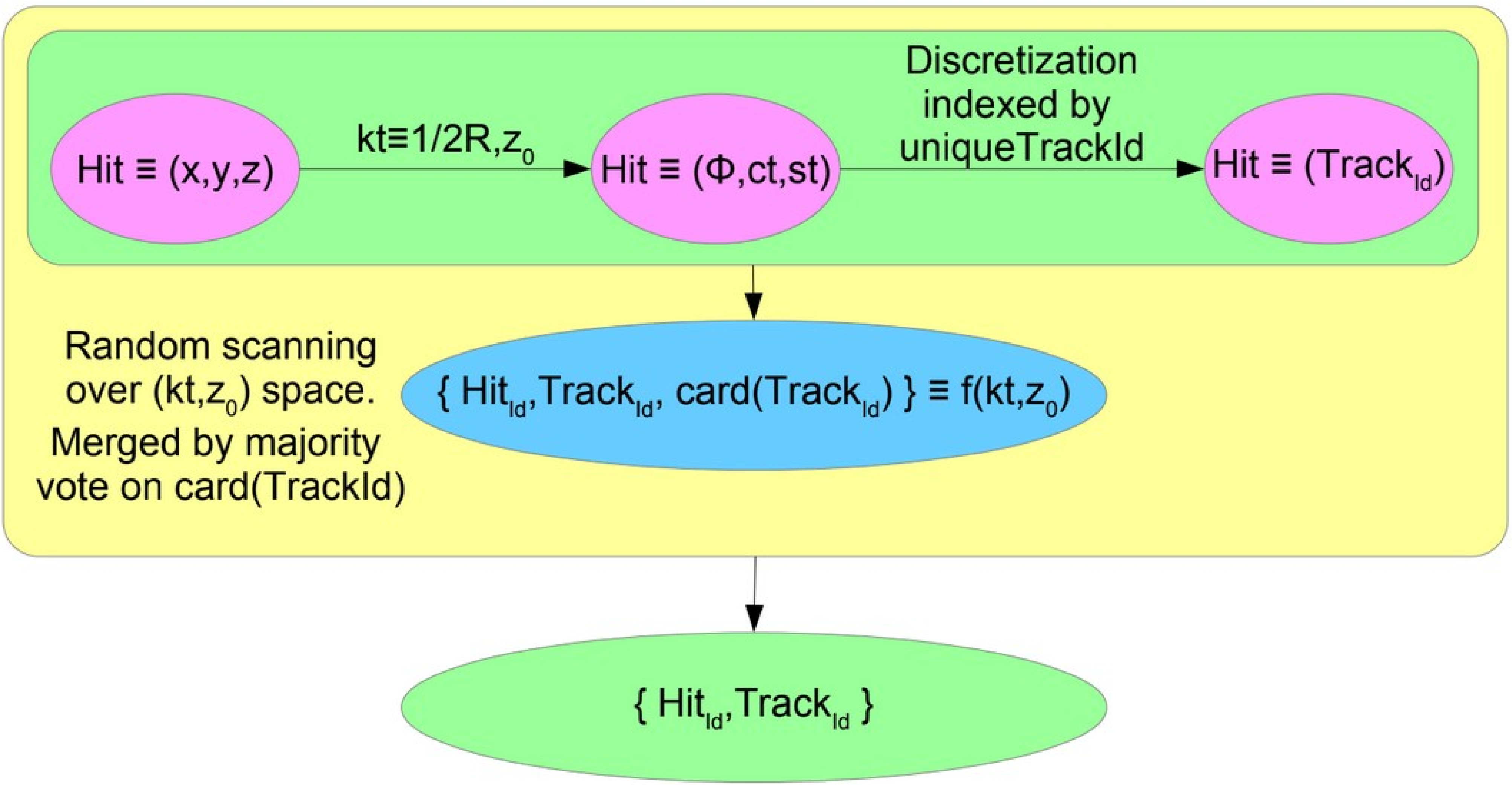


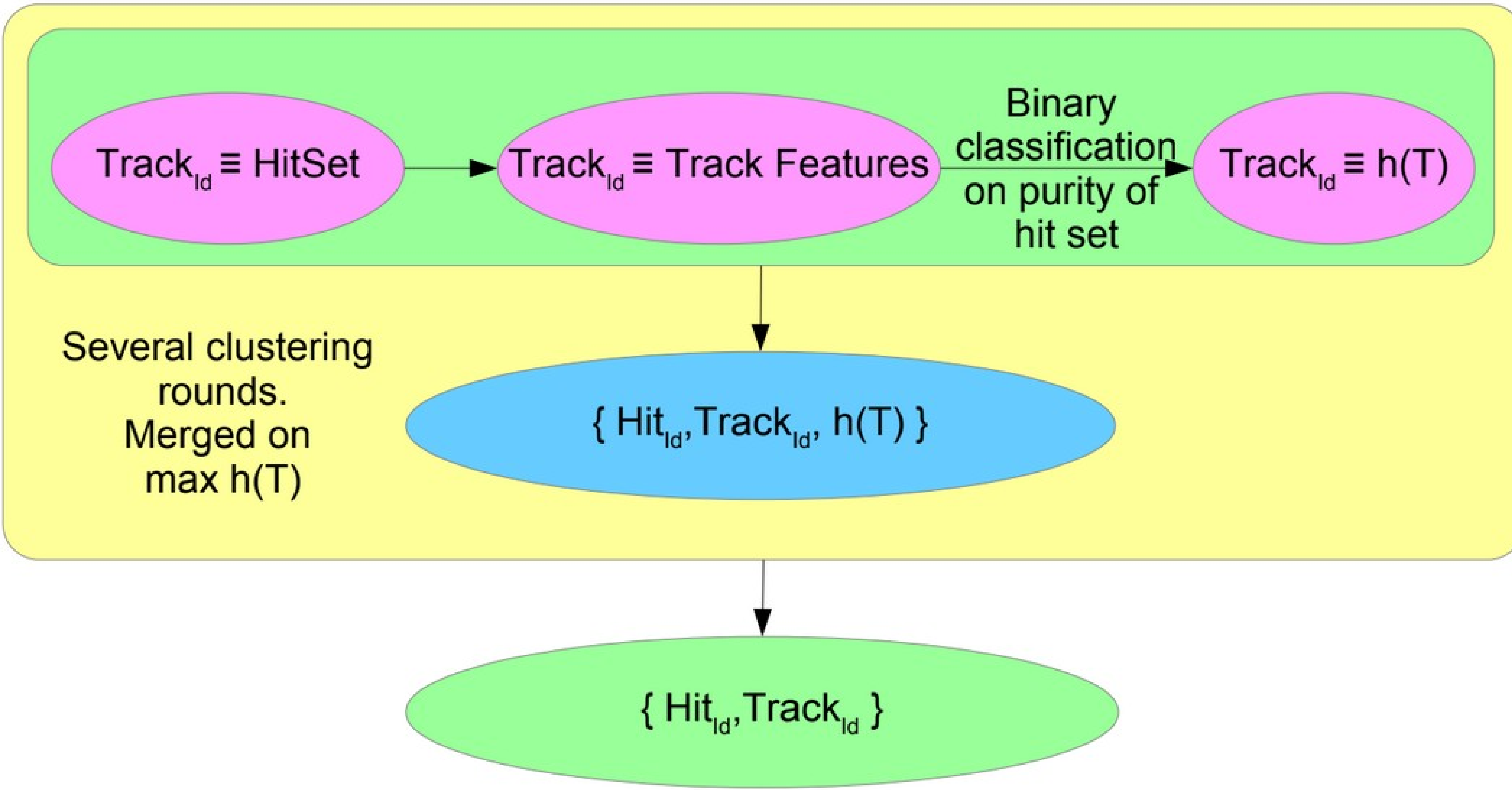
Winning solution → not ML

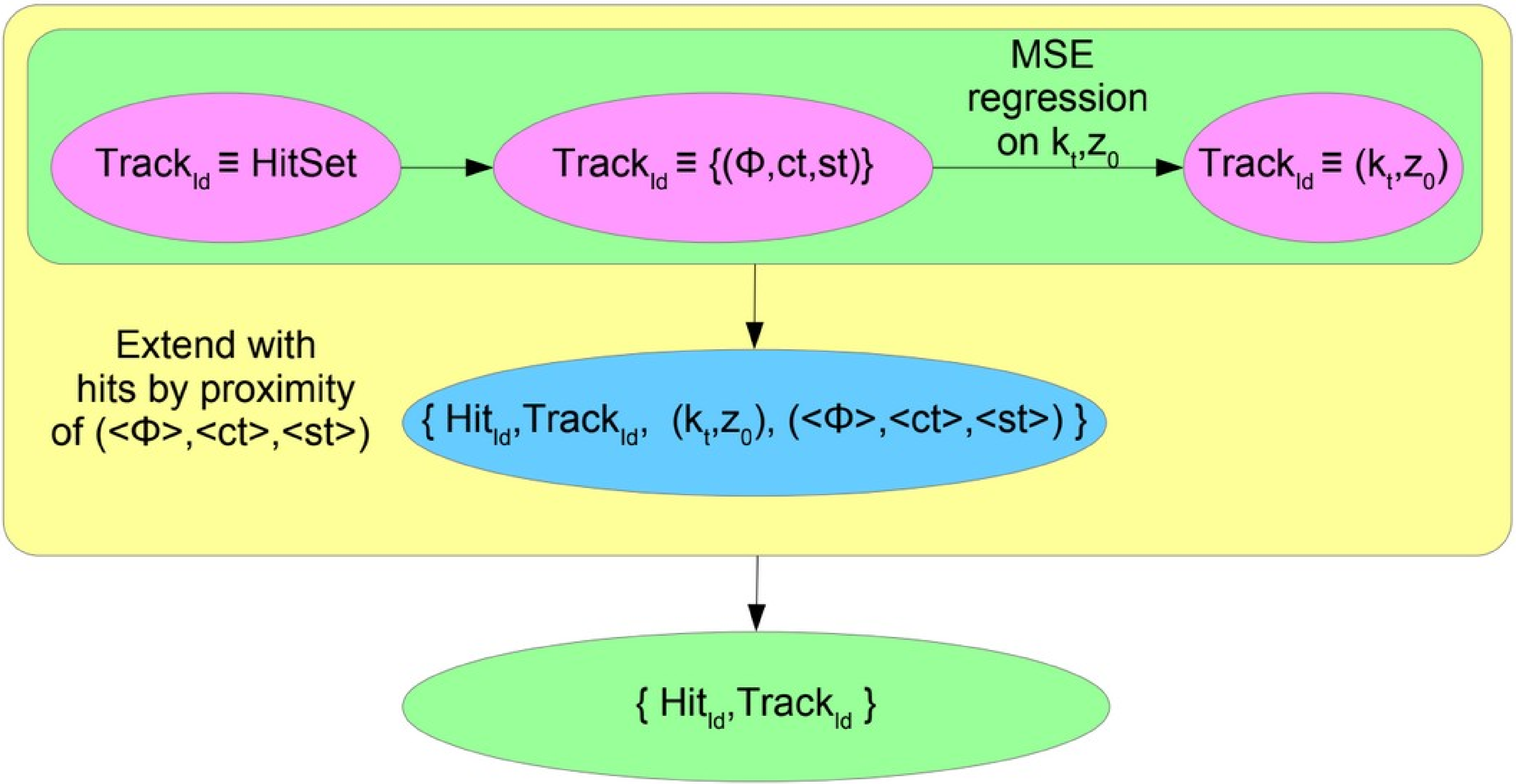


Yuval & Trian (#7)




Hit ≡ (x,y,z)
kt≡1/2R,z_0
Hit ≡ (Φ,ct,st)
Discretization indexed by uniqueTrackId
Hit ≡ ($Track_{Id}$)
Random scanning over (kt,z_0) space.
Merged by majority vote on card(TrackId)
{ Hit_{Id},$Track_{Id}$, card($Track_{Id}$) } ≡ f(kt,z_0)
{ Hit_{Id},$Track_{Id}$ }


Track$_{Id}$ ≡ HitSet
Track$_{Id}$ ≡ Track Features
Binary classification on purity of hit set
Track$_{Id}$ ≡ h(T)
Several clustering rounds. Merged on max h(T)
{ Hit$_{Id}$,Track$_{Id}$, h(T) }
{ Hit$_{Id}$,Track$_{Id}$ }

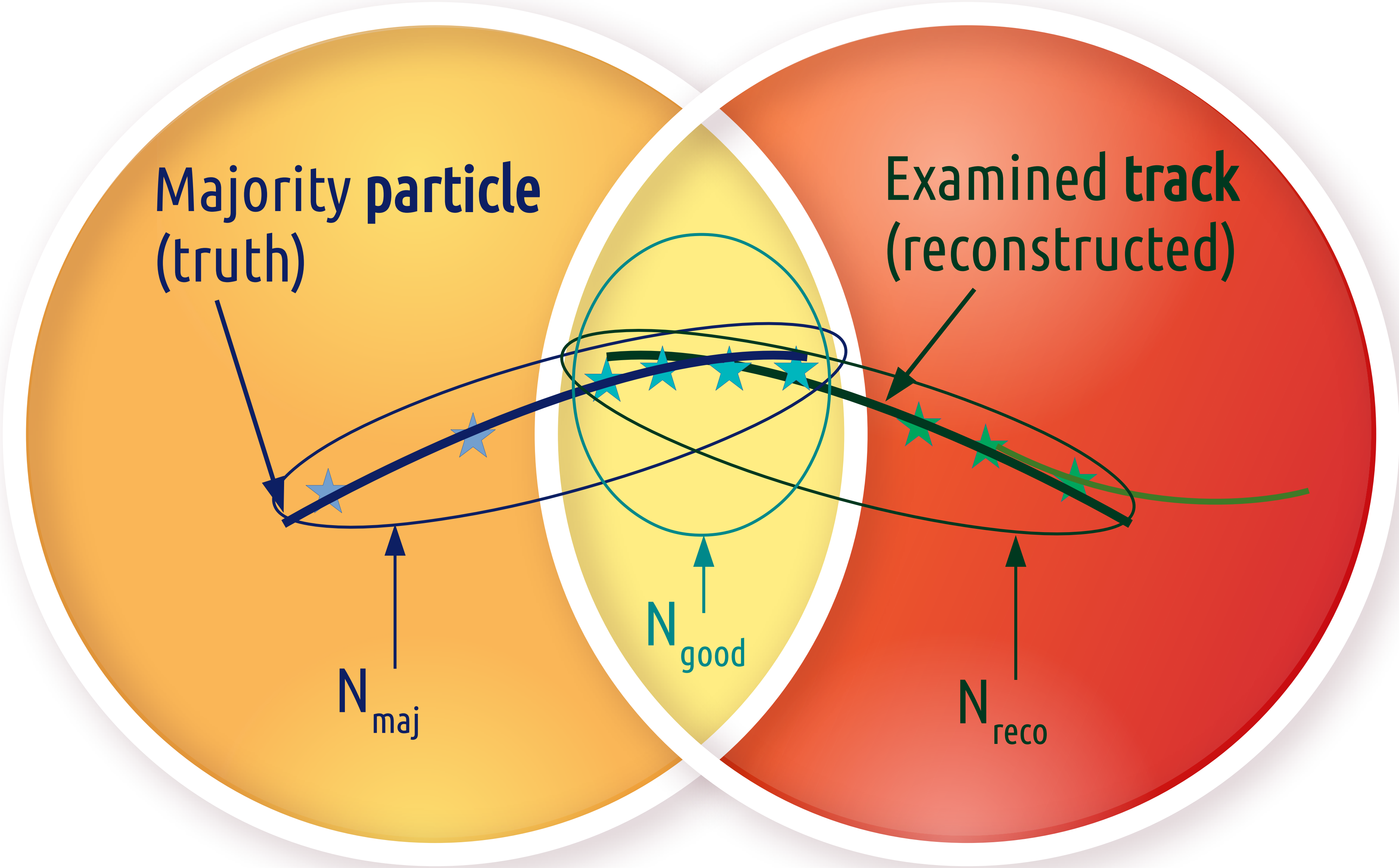


MSE regression on k_t,z_0
$Track_{Id} \equiv HitSet$
$Track_{Id} \equiv \{(\Phi,ct,st)\}$
$Track_{Id} \equiv (k_t,z_0)$
Extend with hits by proximity of (<Φ>,<ct>,<st>)
$\{ Hit_{Id}, Track_{Id}, (k_t,z_0), (\langle\Phi\rangle,\langle ct\rangle,\langle st\rangle) \}$
$\{ Hit_{Id}, Track_{Id} \}$

Other contestants

- #7 : Yuval & Trian
  - « Binned randomized Hough transform » for clustering
  - ML (LightGBM) to merge tracks
- # 9 : CPMP - « DBSCAN Forever »
  - DBSCAN on transformed space including deviation from helix
  - On each iteration clusters = new candidate tracks, merged
- #12 : The Finnies
  - DBSCAN variants → 5 hit track seeding
  - LSTM → estimate 10 hits
  - KNN for Track fitting

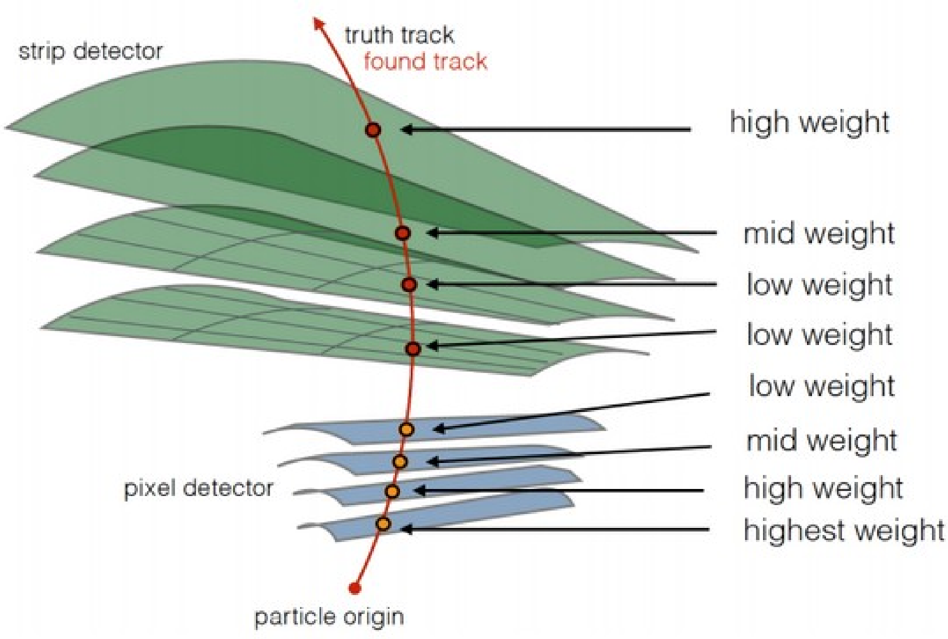
Scoring



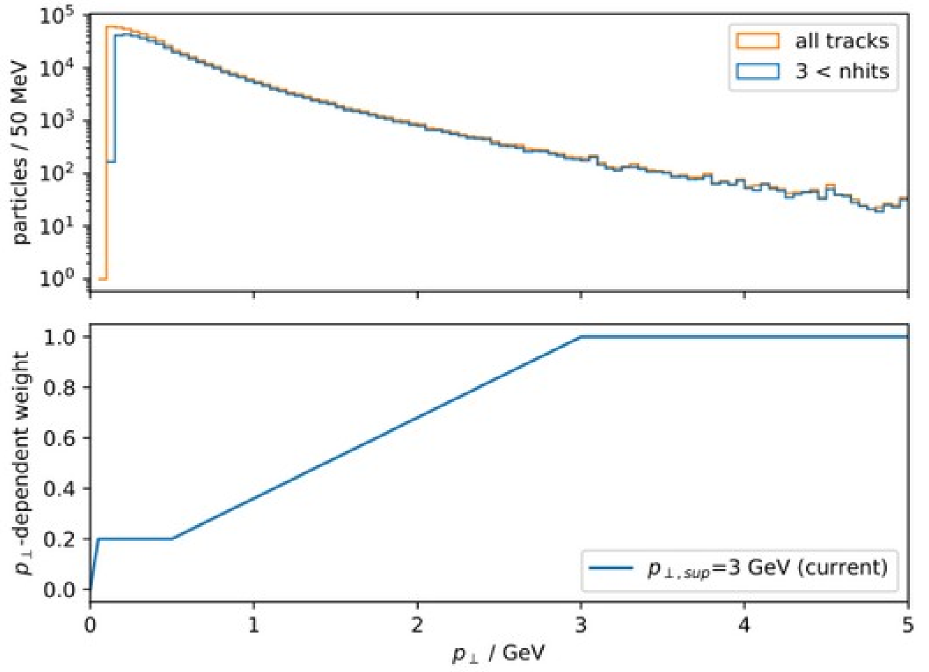
$Purity_{maj} = N_{good} / N_{maj}$

$Purity_{reco} = N_{good} / N_{reco}$

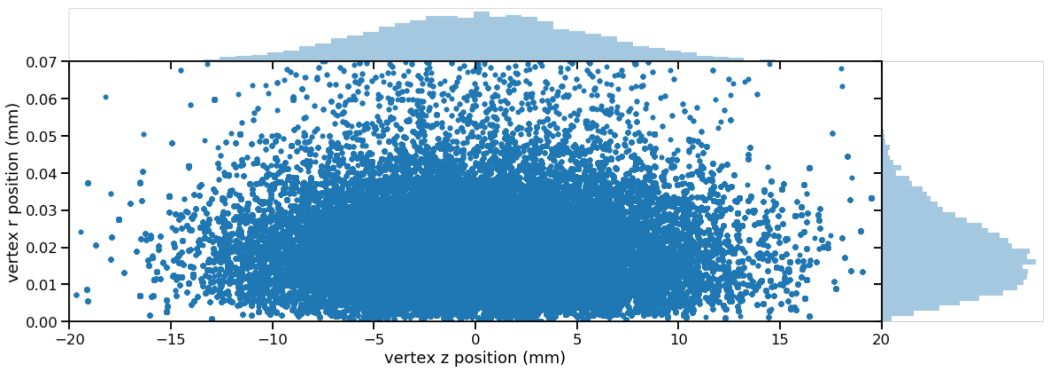
Hit order



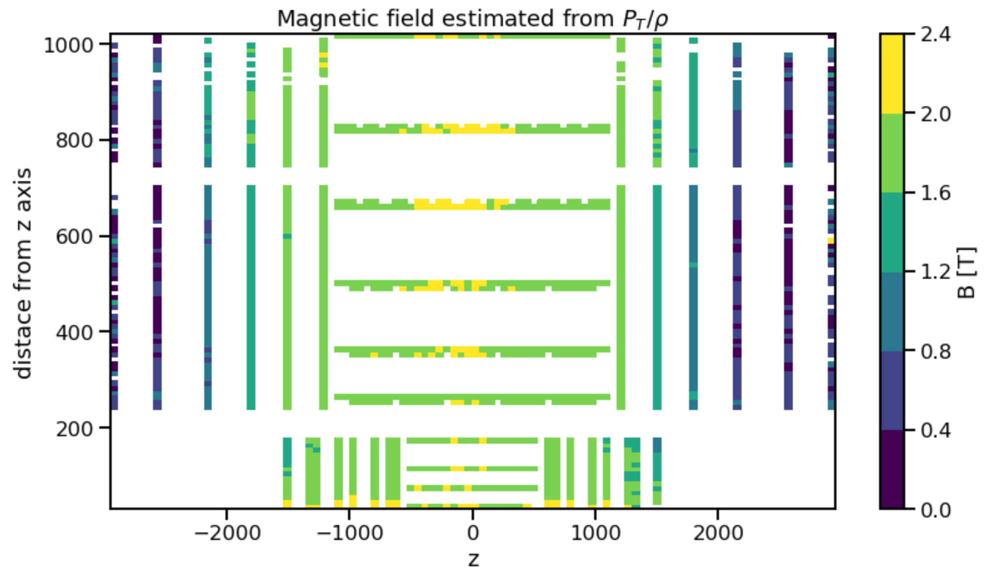
Particle $p_\perp$



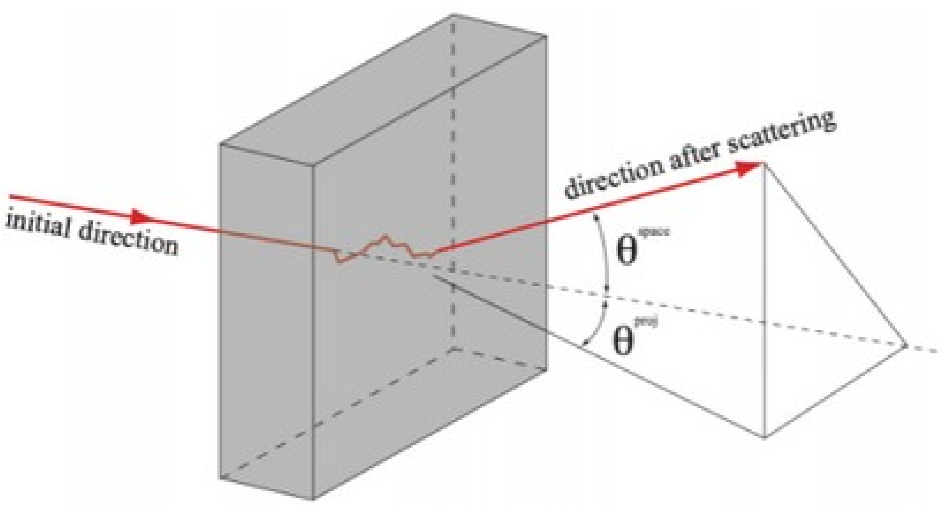
« passing through origin »



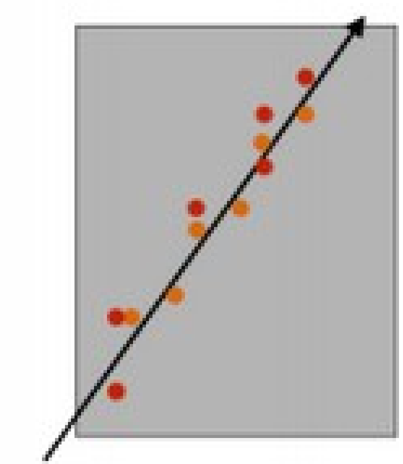
« homogeneous magnetic field »



Scattering in detector



Energy (hence momentum) loss



LHC / HL-LHC Plan