Scientific Data Challenges

David Rousseau

LAL-Orsay

rousseau@lal.in2p3.fr @dhpmrou

Institut Pascal Paris-Saclay Advanced Pattern Recognition

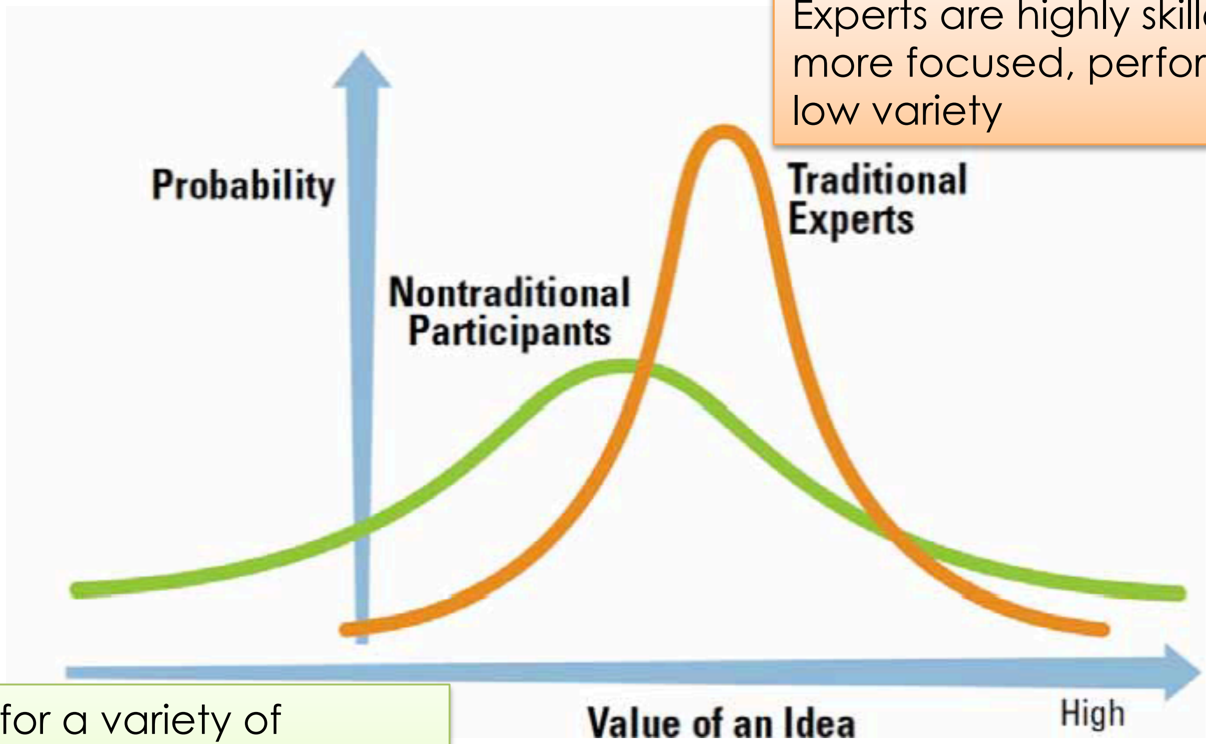
22 Oct 2019

Why challenges work ?

MOTIVATION OF ORGANIZING CONTESTS: EXTREME VALUE

Courtesy : Lakhani 2014

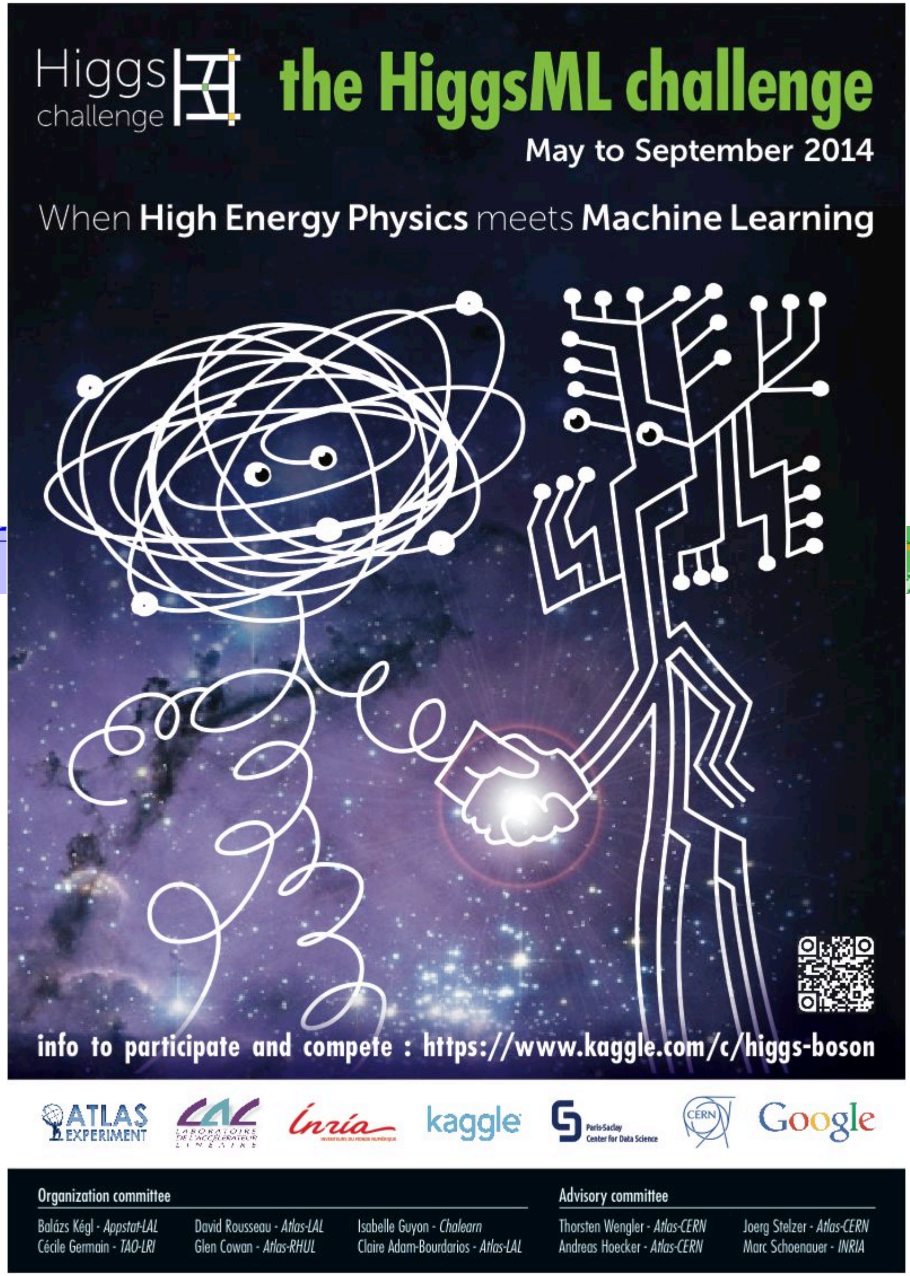
Experts are highly skilled, trained - > more focused, performed solution, low variety



OI is suitable for a variety of nonconvential surprising ideas that are « far » from traditional expertise - > high volatility

Not just ML, but a general trend: Open Innovation, Open Science

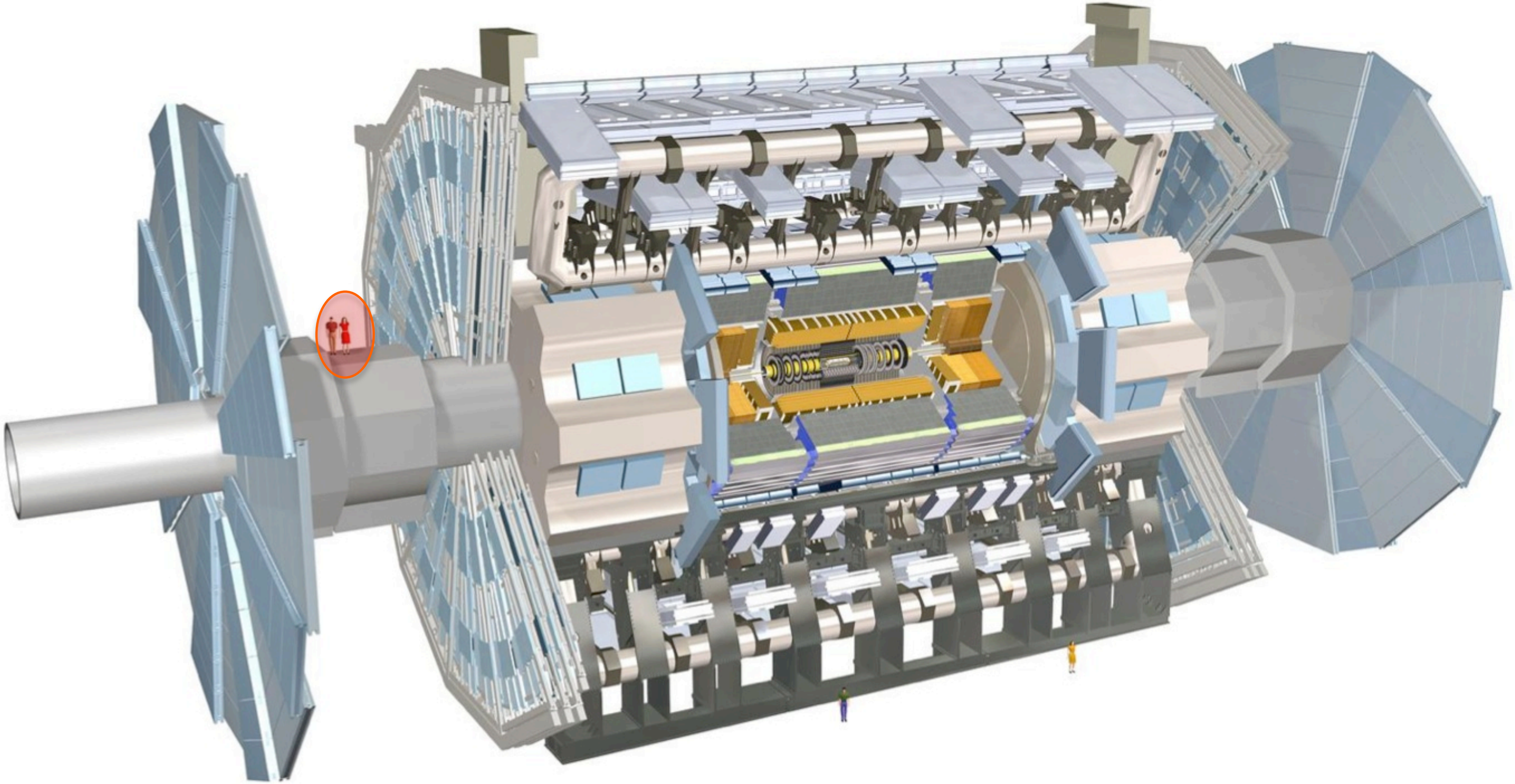
Higgs Machine Learning Challenge



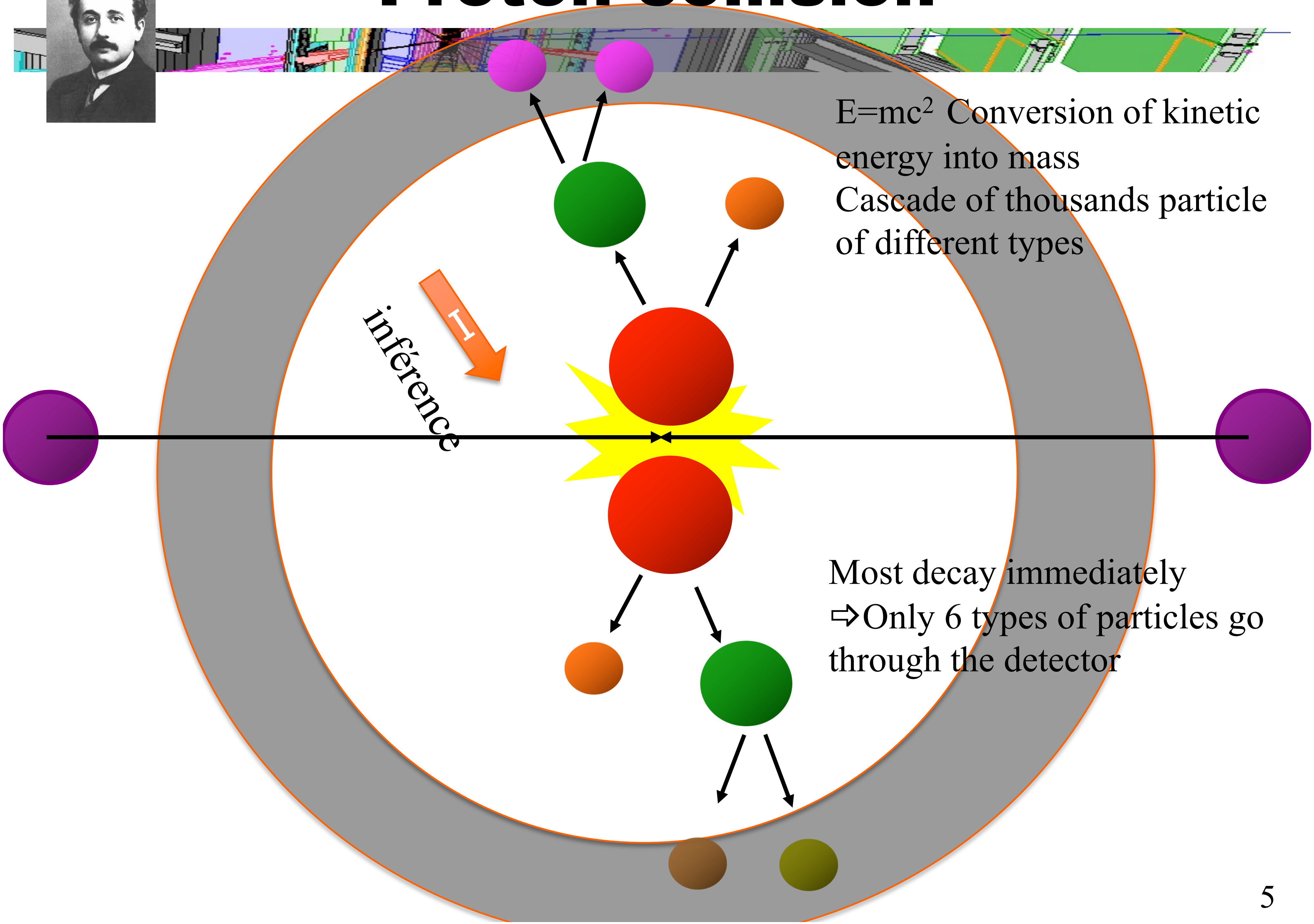
ATLAS detector

Diamètre: 25m
Longueur: 46m
Poids: 7000 tonnes

3000 km de câbles
100 millions de canaux



Proton collision



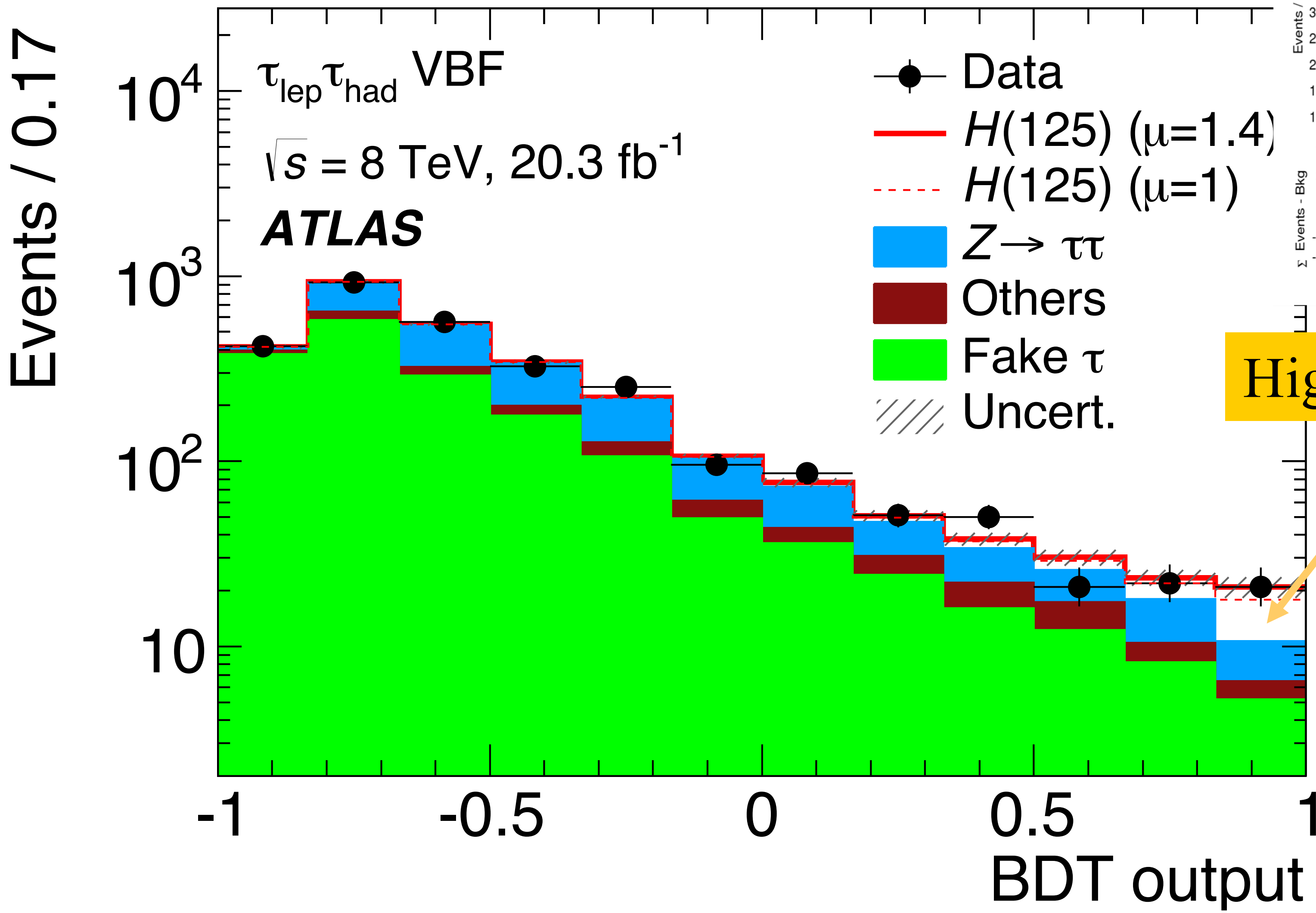
E=mc^2 Conversion of kinetic energy into mass
Cascade of thousands particle of different types

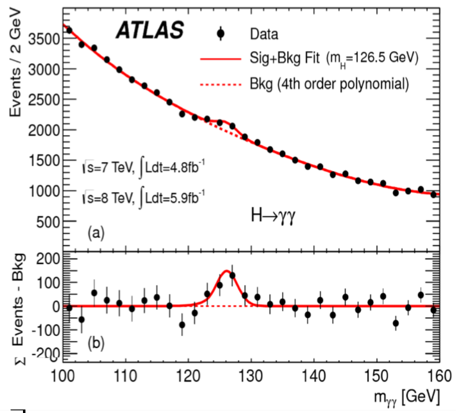
Most decay immediately
⇨Only 6 types of particles go through the detector

Classifier

JHEP 04, 117 (2015) 1501.04943

Boosted Decision Tree using ~a dozen of high level variables





Higgs evidence

Dataset

Permanently available and usable by anyone (also non ATLAS) on CERN Open Data:

http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014

ASCII csv file, with mixture of Higgs to tautau (lephad) signal and corresponding backgrounds, from official GEANT4 ATLAS simulation

800.000 events (250k public, 550k private)

Weight and Label (for training dataset only)

weight (fully normalised)

label : « s » or « b »

Variables (used in ATLAS paper):

DER_mass_MMC
DER_mass_transverse_met_lep
DER_mass_vis
DER_pt_h
DER_deltaeta_jet_jet
DER_mass_jet_jet
DER_prodeta_jet_jet
DER_deltar_tau_lep
DER_pt_tot
DER_sum_pt
DER_pt_ratio_lep_tau
DER_met_phi_centrality
DER_lep_eta_centrality

Primitive 3-vectors allowing to compute the conf note variables (mass neglected),

16 independent variables:

PRI_tau_pt
PRI_tau_eta
PRI_tau_phi
PRI_lep_pt
PRI_lep_eta
PRI_lep_phi
PRI_met
PRI_met_phi
PRI_met_sumet
PRI_jet_num (0,1,2,3, capped at 3)
PRI_jet_leading_pt
PRI_jet_leading_eta
PRI_jet_leading_phi
PRI_jet_subleading_pt
PRI_jet_subleading_eta
PRI_jet_subleading_phi
PRI_jet_all_pt

Real analysis vs challenge

1. Systematics (and data vs MC)
2. 2 categories x n BDT score bins

1. Background estimated from data (embedded, anti tau, control region) and some MC
2. Weights include all corrections. Some negative weights (tt)
3. Potentially use any information from all 2012 data and MC events
4. Few variables fed in two BDT

1. Significance from complete fit with NP etc…
2. MVA with TMVA BDT

1. No systematics
2. No categories, one signal region
3. Straight use of ATLAS G4 MC
4. Weights only include normalisation and pythia weight. Neg. weight events rejected.
5. Only use variables and events preselected by the real analysis
6. All BDT variables + categorisation variables + primitives 3-vector
7. Significance from "regularised Asimov"
8. MVA "no-limit"

Simpler, but not too simple!

Final leaderboard

#	Δrank	Team Name ‡ model uploaded * in the money		Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Gábor Melis ‡ *	7000$	3.80581	110	Sun, 14 Sep 2014 09:10:04 (-0h)
2	↑1	Tim Salimans ‡ *	4000$	3.78913	57	Mon, 15 Sep 2014 23:49:02 (-40.6d)
3	↑1	nhlx5haze ‡ *	2000$	3.78682	254	Mon, 15 Sep 2014 16:50:01 (-76.3d)
4	↑38	ChoKo Team		3.77526	216	Mon, 15 Sep 2014 15:21:36 (-42.1h)
5	↑35	cheng chen		3.77384	21	Mon, 15 Sep 2014 23:29:29 (-0h)
6	↑16	quantify		3.77086	8	Mon, 15 Sep 2014 16:12:48 (-7.3h)
7	↑1	Stanislav Semenov & Co (HSE Yandex)		3.76211	68	Mon, 15 Sep 2014 20:19:03
8	↓7	Luboš Motl's team		3.76050	589	Mon, 15 Sep 2014 08:38:49 (-1.6h)
9	↑8	Roberto-UCIIIM		3.75864	292	Mon, 15 Sep 2014 23:44:42 (-44d)
10	↑2	Davut & Josef		3.75838	161	Mon, 15 Sep 2014 23:24:32 (-4.5d)
45	↑5	crowwork ‡	HEP meets ML award XGBoost authors Free trip to CERN	3.71885	94	Mon, 15 Sep 2014 23:45:00 (-5.1d)
782	↓149	Eckhard	Tuned TMVA	3.49945	29	Mon, 15 Sep 2014 07:26:13 (-46.1h)
991	↑4	Rem.		3.20423	2	Mon, 16 Jun 2014 21:53:43 (-30.4h)
		simple TMVA boosted trees		3.19956		

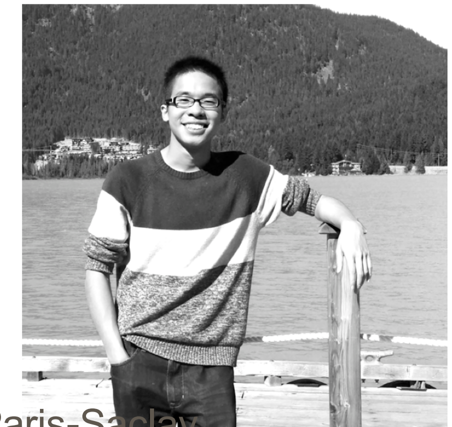
Participation

- Big success !
- 1785 teams (1942 people) have participated (participation=submission of at least one solution)
  - (6517 people have downloaded the data)
  - ➔most popular challenge on the Kaggle platform, ever (Amazon.com employee access challenge 1687 teams, Allstate Purchase Prediction Challenge 1567 teams)
- 35772 solutions uploaded
- 136 forum topics with 1100 posts

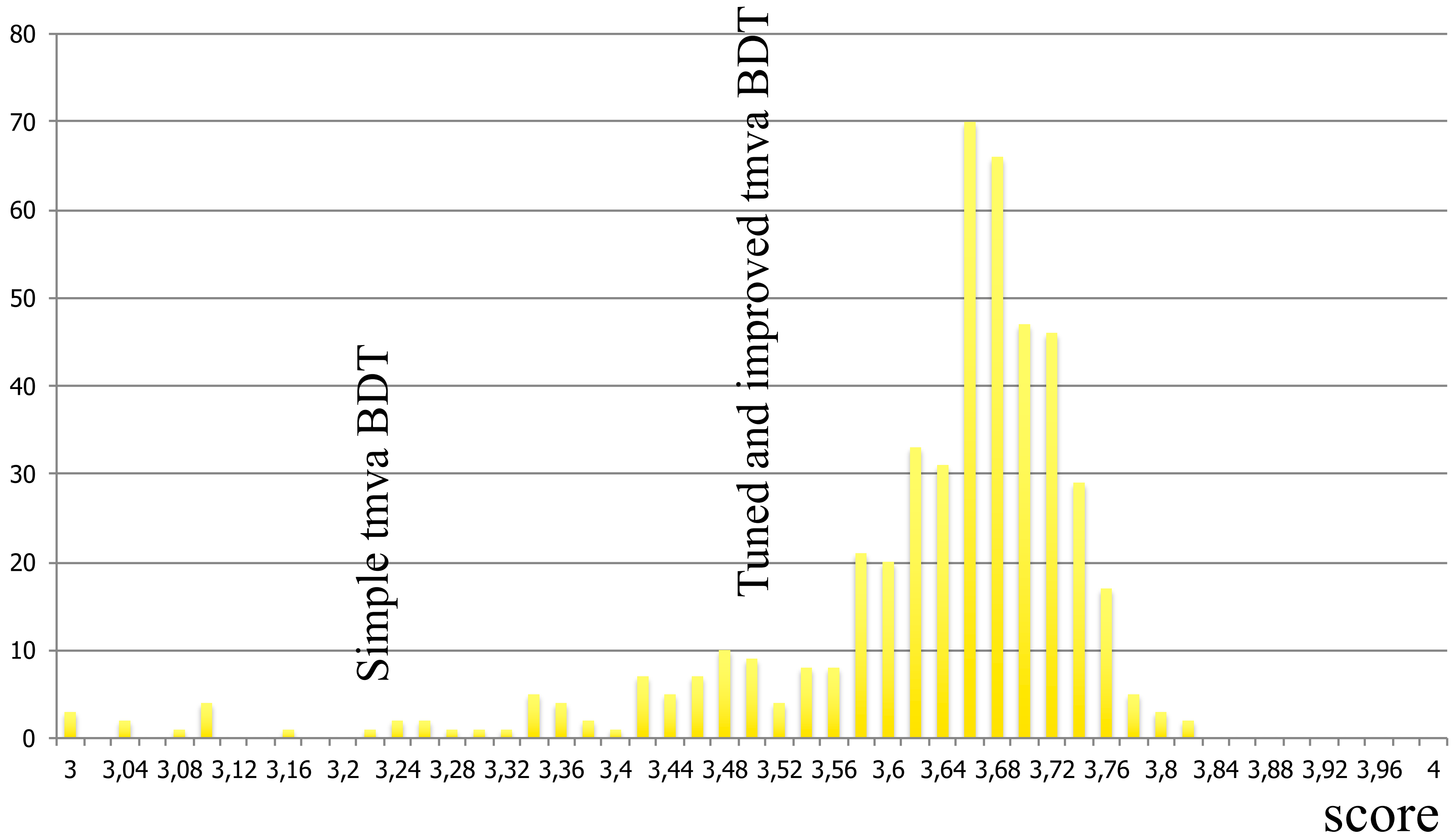
Who are the winners?

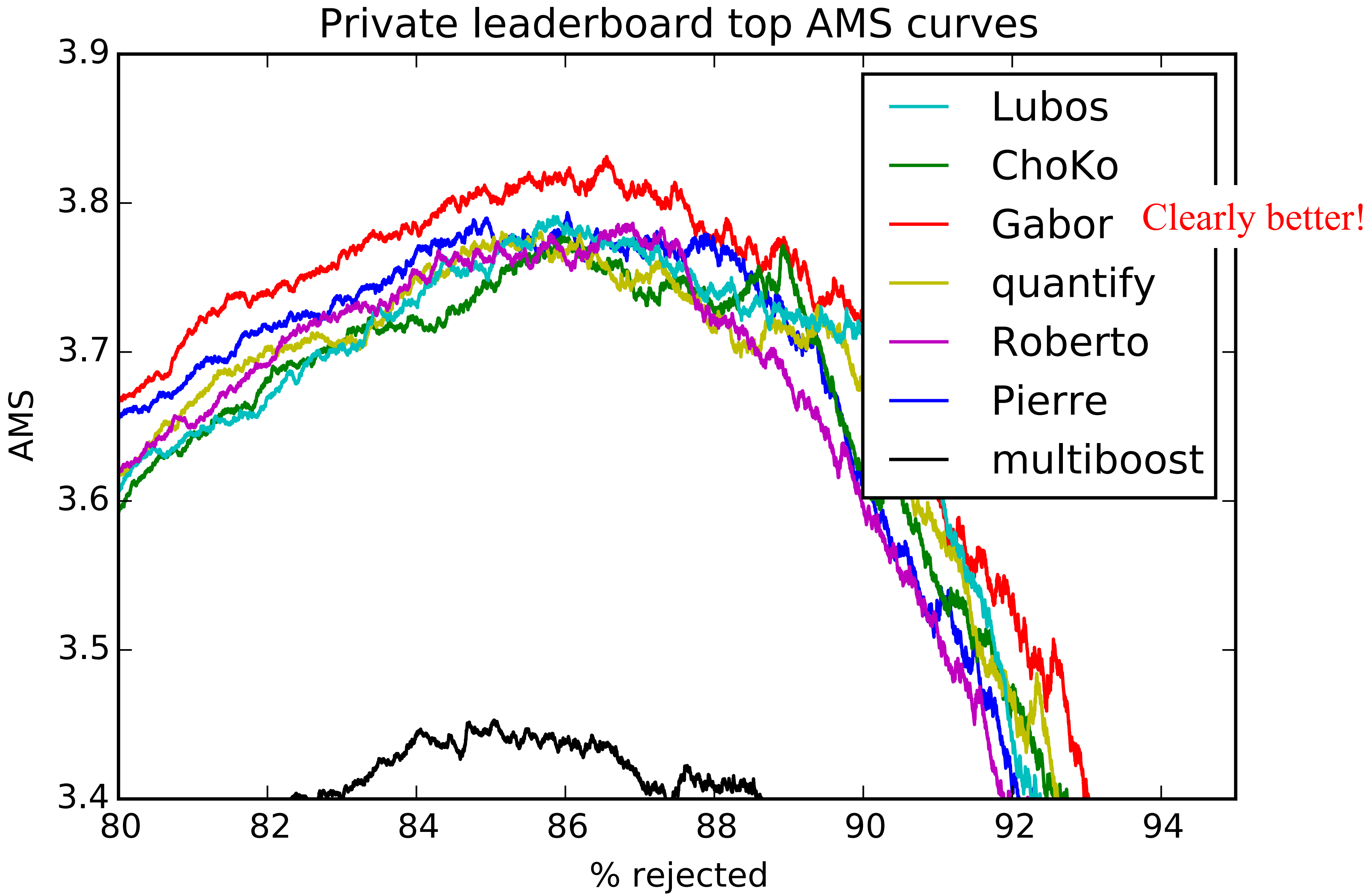
- See http://atlas.ch/news/2014/machine-learning-wins-the-higgs-challenge.html
- 1 : Gabor Melis (Hungary) lisp developer and consultant : wins 7000$
  - Note : hired by DeepMind in summer 2015
- 2 : Tim Salimans (Neitherland) data science consultant: wins 4000$
  - Note : hired by Google Brain
- 3 : Pierre Courtiol (France) ? Data Scientist : wins 2000$
  - Note : hired by OWKIN
- HEP meets ML award: team crowwork, Tianqi Chen and Tong He PhD students in data science at Seattle and Vancouver. Provided XGBoost used by many participants. Win a free trip and visit to CERN in 2015.
  - →longer lasting contribution, XGBoost later became the de facto standard for BDT

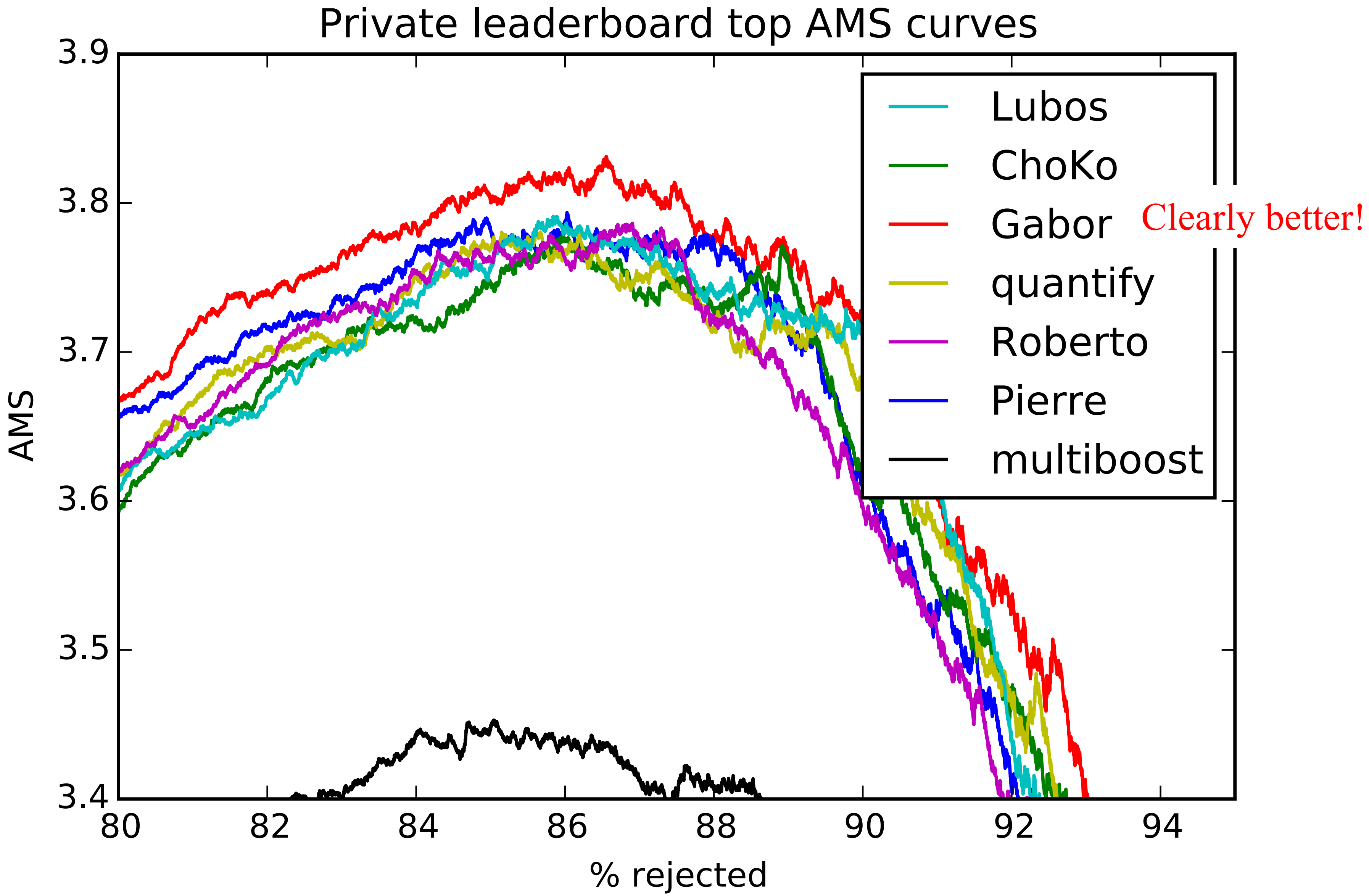
?



Best private scores




Private leaderboard top AMS curves
3.9
3.8
3.7
3.6
3.5
3.4
AMS
80
82
84
86
88
90
92
94
% rejected
Lubos
ChoKo
Gabor
quantify
Roberto
Pierre
multiboost
Clearly better!


Private leaderboard top AMS curves
AMS
% rejected
3.9
3.8
3.7
3.6
3.5
3.4
80
82
84
86
88
90
92
94
Lubos
ChoKo
Gabor
quantify
Roberto
Pierre
multiboost
Clearly better!

HiggsML outcome

- Best significance 20% than with Root-TMVA (equivalent to 40% more data)
- (gradient) BDT algorithm of choice in this case where number variables and number of training events limited
- NN very slightly better (#1 Gabor) but much more difficult to tune, need more training data
  - ➔main hindsight regret on the design of HiggsML : too few data
- XGBoost written for HiggsML, now best BDT on the market
- Wealth of ideas, documented in JMLR proceedings v42
- Raised awareness about ML in HEP

TrackML

Several relevant talks in IPA Advanced Pattern Recognition Workshop : https://indico.cern.ch/event/847626/timetable/?view=standard:

- Moritz Kiehn : TrackML challenge summary
- Andreas Salzburger : ACTS (TrackML simulation engine)
- Sergei Gorbunov : #1 solution
- Marcel Kunze : #3 solution and more
- Lucy Linder : spin-off Quantum Computing
- Basara, Biswas, Ghosh : spin-off Optical Processor Unit
- Sabrina Amrouche : spin-off similarity Hashing
- Tobias Isenberg/Xiyao Wang : spin off data visualisation

Who and How

- Organisation: Jean-Roch Vlimant (Caltech), Vincenzo Innocente, Andreas Salzburger (CERN), 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, Isabelle Guyon, Victor Estrade (LRI-Orsay), Edward Moyse (University of Massachussets), Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex, HSE)

5-6 FTE year

Platforms:

kaggle Phase 1

CodaLab Phase 2

accuracy

throughput

TrackML sponsors

CERN

kaggle

NVIDIA

UNIVERSITÉ DE GENÈVE

Paris-Saclay Center for Data Science

INSTITUTE DATAIA
Data Science, Intelligence & Society

CHA LEARN

iris hep
Institute for Research & Innovation in Software for High Energy Physics

CERN openlab

Inria
INVENTORS FOR THE DIGITAL WORLD

PyTorch

cnrs

erc

Fermilab

COMMON GROUND

inqnet

UNIVERSITÉ PARIS SUD

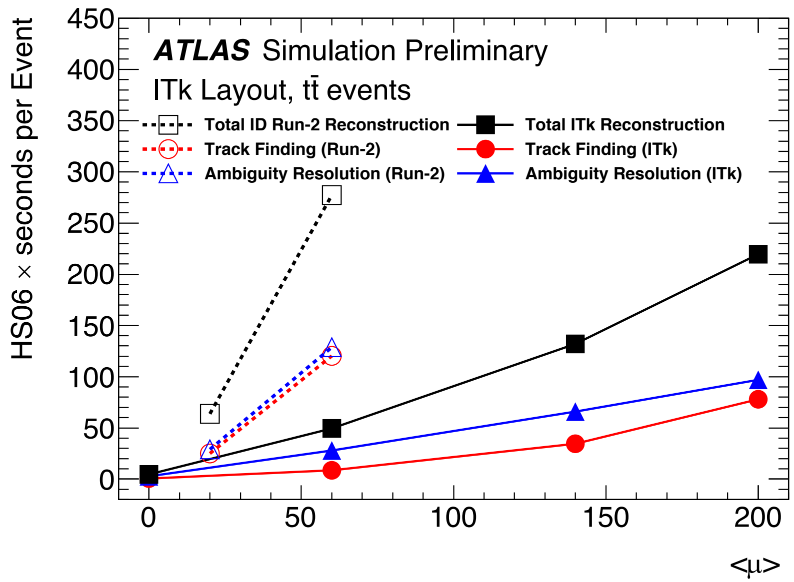
université PARIS-SACLAY

Tracking crisis

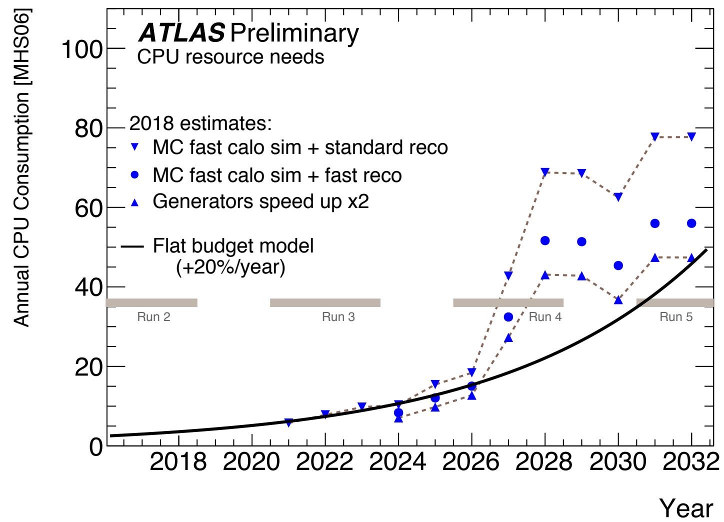
- Tracking (in particular pattern recognition) dominates reconstruction CPU time at LHC
- HL-LHC (phase 2) perspective : increased pileup :Run 1 (2012): <>~20, Run 2 (2015): <>~50,Phase 2 (2025): <>~200
- CPU time quadratic/exponential extrapolation

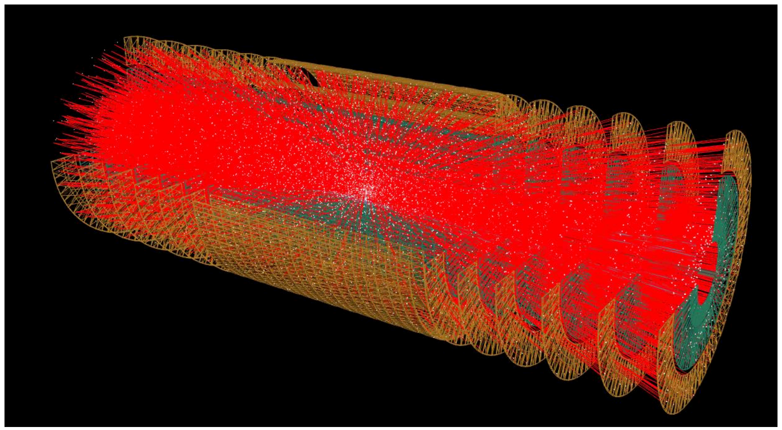
- On-going Large effort within HEP to optimise software and tackle micro and macro parallelism.
- >20 years of LHC tracking development. Everything has been tried?
  - Maybe yes, but maybe algorithm slower at low lumi but with a better scaling have been dismissed ?
  - Maybe no, brand new ideas from ML

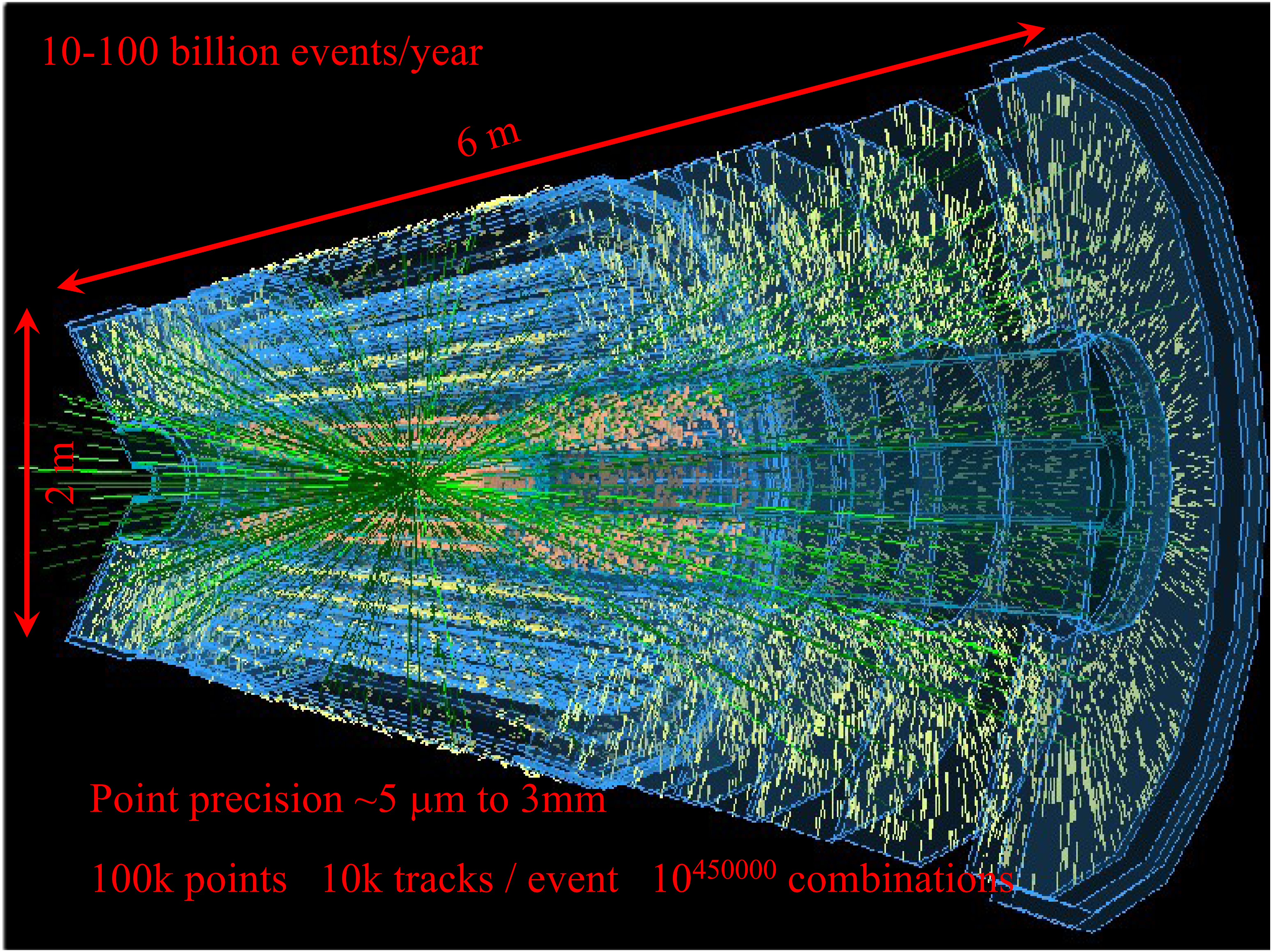
- →challenge !!



Similar plots from CMS

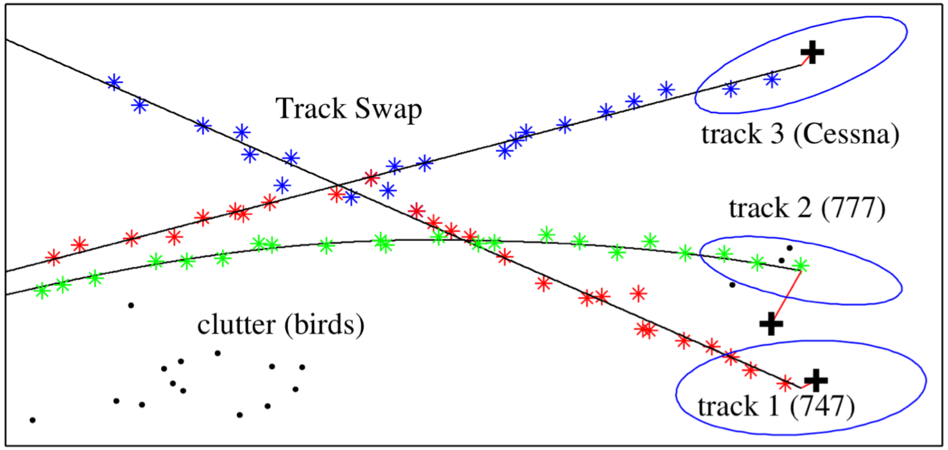


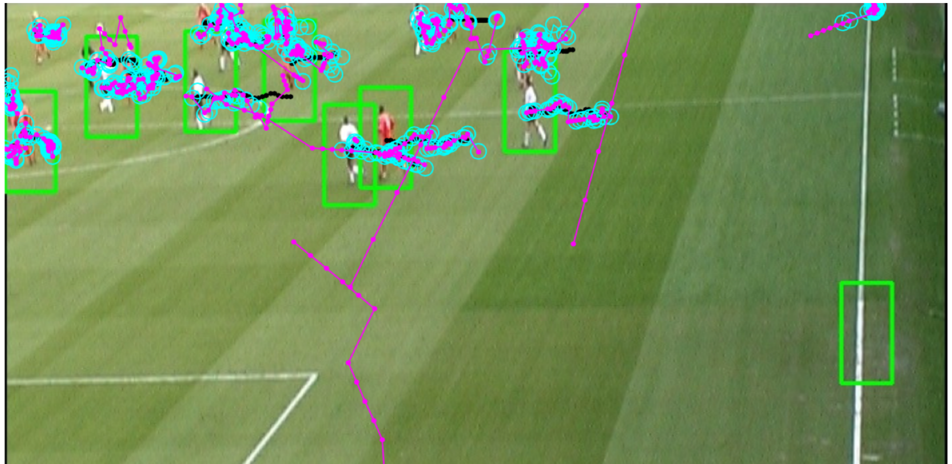


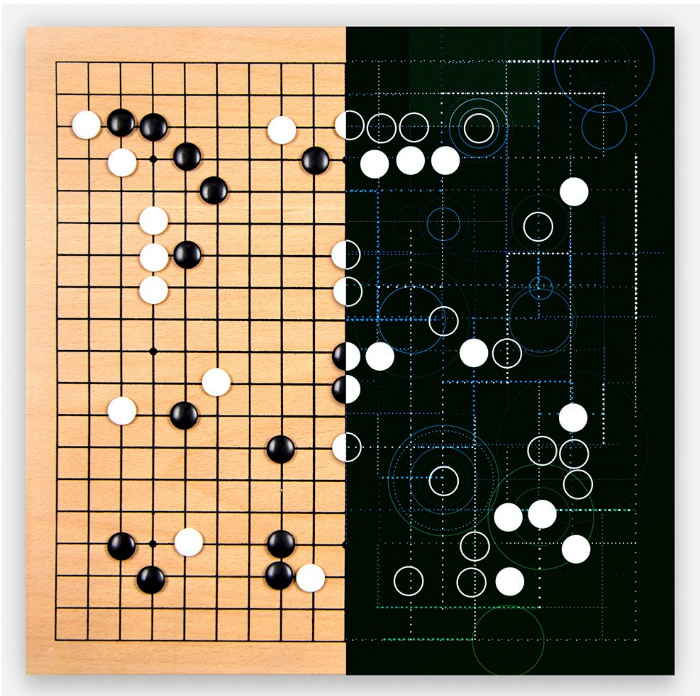

10-100 billion events/year
6 m
2 m
Point precision ~5 μm to 3mm
100k points 10k tracks / event 10450000 combinations

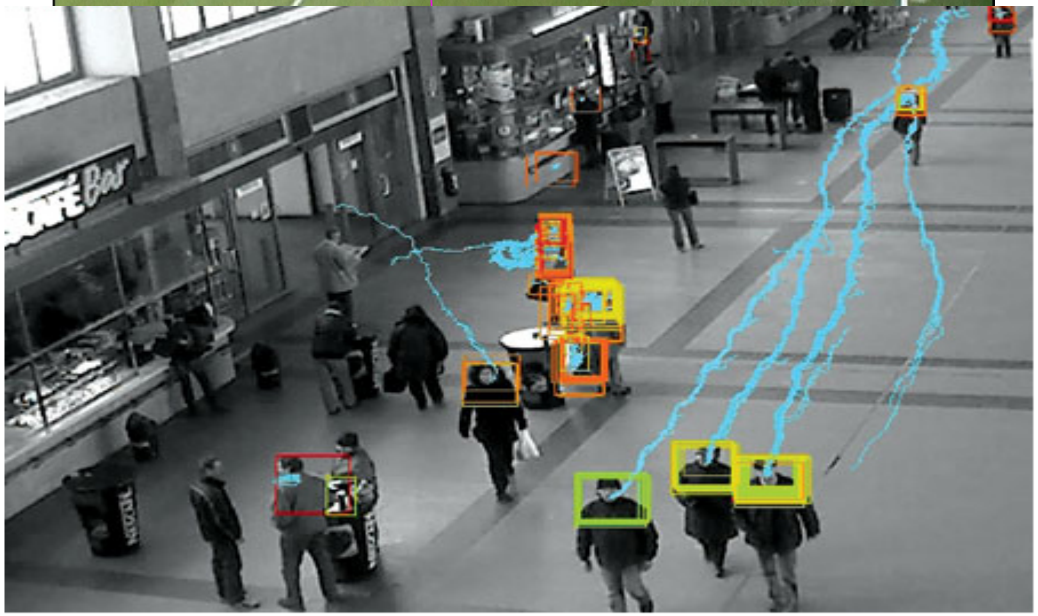
Tracking outside HEP

- …is very different









An early attempt

Computer Physics Communications 49 (1988) 429–448
North-Holland, Amsterdam

NEURAL NETWORKS AND CELLULAR AUTOMATA IN EXPERIMENTAL HIGH ENERGY PHYSICS

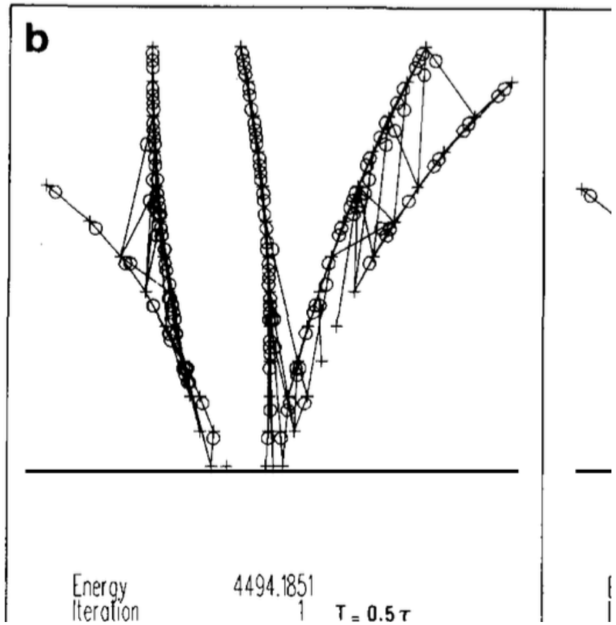
B. DENBY

Laboratoire de l'Accélérateur Linéaire, Orsay, France

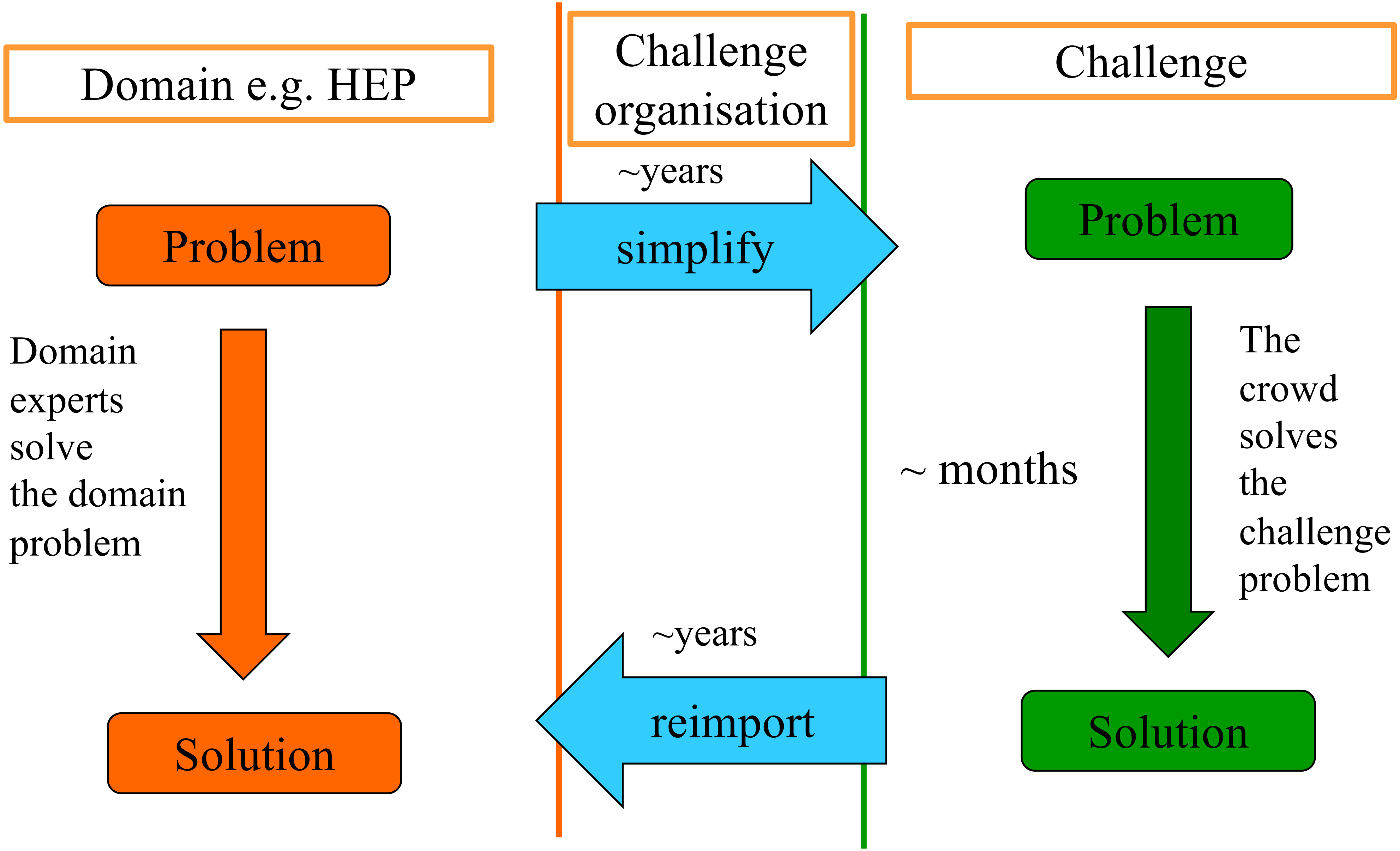
Received 20 September 1987; in revised form 28 December 1987

Bruce Denby

- 1987 Very first Neural Net in HEP paper known
- NN for tracking and calo clustering
- B. Denby then moved from Delphi at LEP to CDF at Tevatron. He still active outside HEP: 2017 analysis of ultrasonic image of the tongue
- 1992 JetNet Carsten Peterson, Thorsteinn Rognvaldsson (Lund U.) , Leif Lonnblad (CERN) (~500 citations) really started NN use in HEP



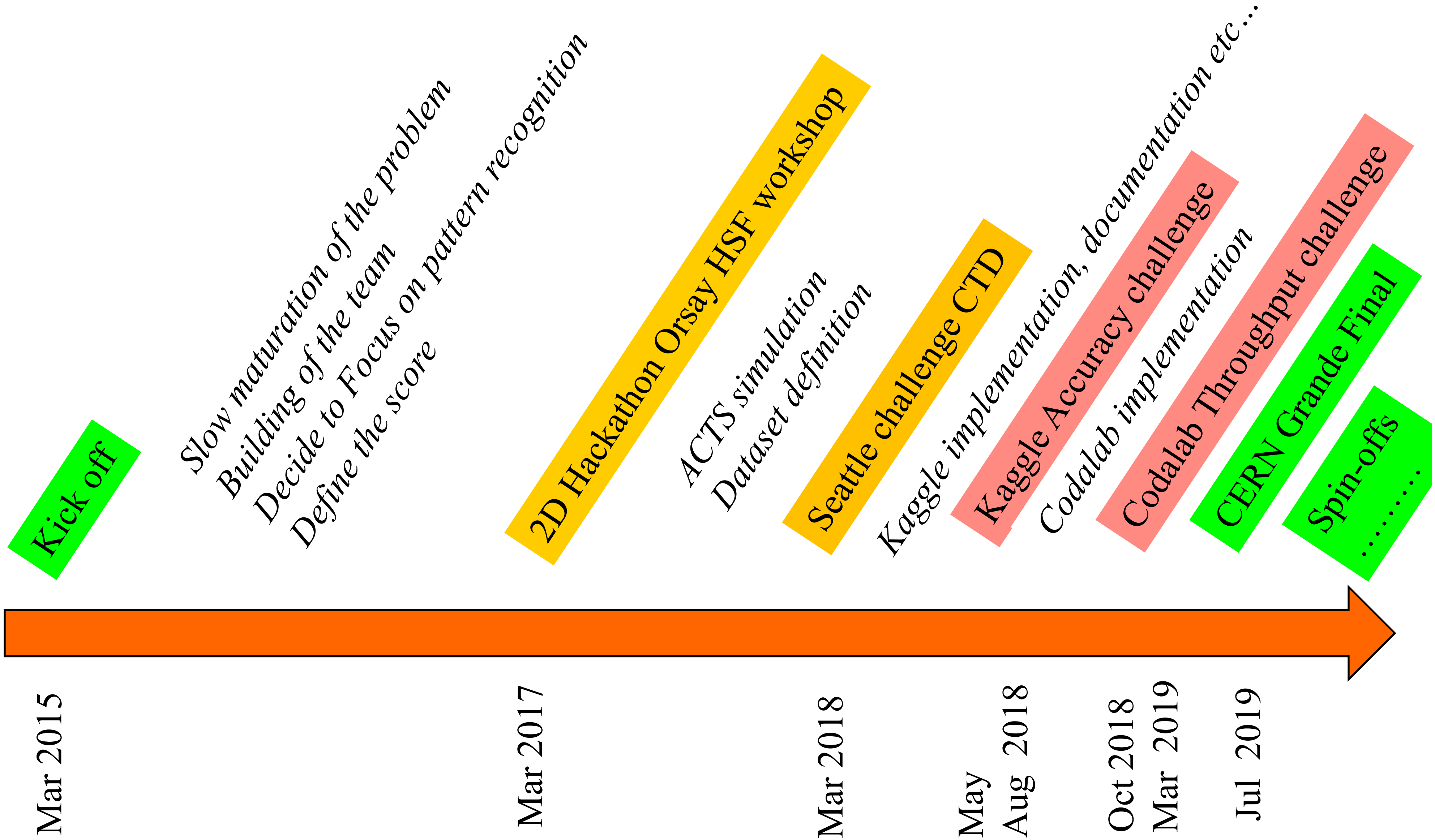
From domain to challenge and back



TrackML in a nutshell

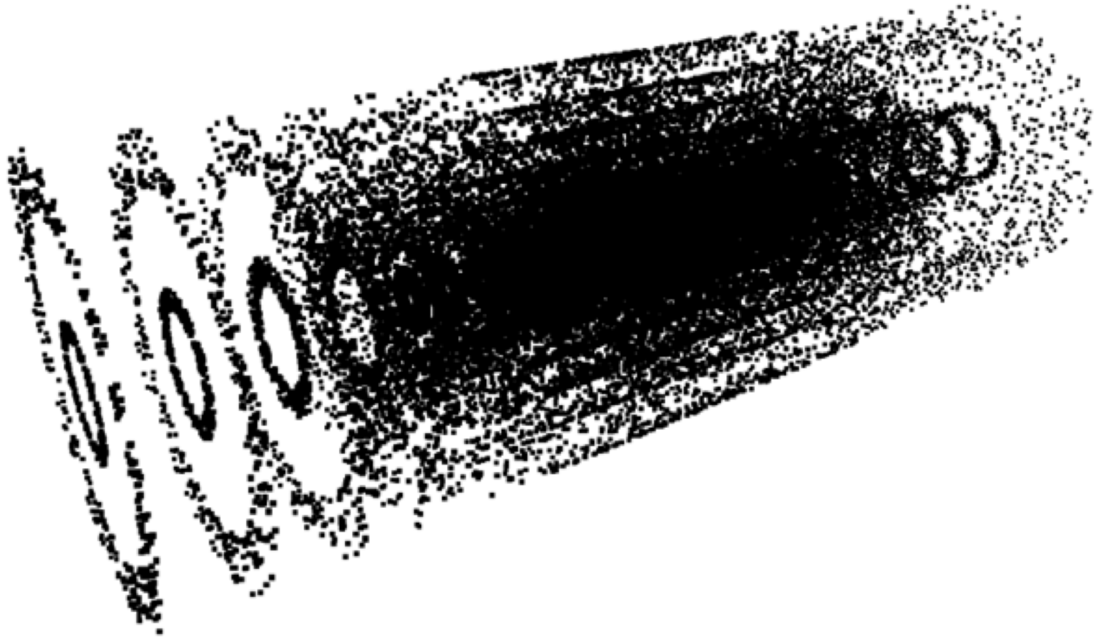
- Accurate simulation engine (ACTS https://gitlab.cern.ch/acts/acts-core) to produce realistic events
  - Ttbar events with 200 pileup
  - Silicon detector with barrels and disks (simplified HL-LHC ATLAS or CMS Si detector)
  - One file with list of 3D points
  - Ground truth : one file with point to particle association
  - Ground truth auxiliary : true particle parameter (origin, direction, curvature)
  - Typical events with ~200 parasitic collisions (~10.000 tracks/event)
- Large training sample 10k events, 0.1 billion tracks, 1 billion points, ~100GByte
- Accuracy phase (May to August 2018) on Kaggle
  - Participants are given the test sample (with usual split for public and private leaderboard) and run the evaluation to find the tracks
  - They should upload the tracks they have found
  - A track is a list of 3D points
  - Score : fraction of points correctly grouped together
  - Evaluation on test sample with per-mille precision on 100 event
- Throughput phase Sep to Mar 2019 on Codalab
  - Participants submit their code to solve the same probmem
  - Strong CPU incentive

TrackML timeline



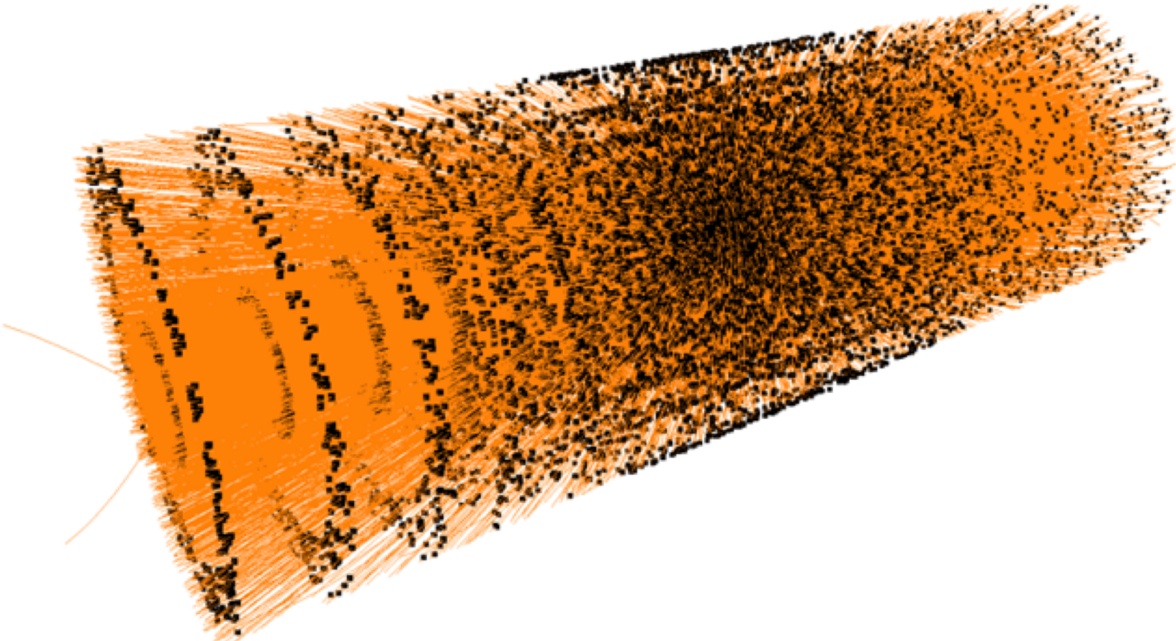
Dataset

3D points



Dataset

3D points ➔ tracks



Datasets

❑ Particle file

		origin vertex (mm)			momentum (GeV)			charge
	particle_id	**vx**	**vy**	**vz**	**px**	**py**	**pz**	**q**
0	4503805785800704	-0.021389	-0.012618	-0.624757	38.907001	-16.146099	-84.311096	-1
1	4504011944230912	-0.021389	-0.012618	-0.624757	-0.661993	0.118267	249.181000	1
2	4504080663707648	-0.021389	-0.012618	-0.624757	0.821614	0.954217	0.948994	-1
3	4504149383184384	-0.021389	-0.012618	-0.624757	0.300791	0.080450	2.656530	1
4	4504218102661120	-0.021389	-0.012618	-0.624757	-0.552250	-0.481988	-0.888733	1

(note : we do not ask participant to reconstruct these track parameters but these could be useful latent variables)

❑ (static)Detector file

				center position (mm)			3x3 rotation matrix			
	volume_id	**layer_id**	**module_id**	**cx**	**cy**	**cz**	**rot_xu**	**rot_xv**	**rot_xw**	**ro**
0	6	2	1	-65.7965	-5.17830	-1502.5	0.078459	-0.996917	0.0	-0.99
1	6	2	2	-139.8510	-6.46568	-1502.0	0.046183	-0.998933	0.0	-0.99
2	6	2	3	-138.6570	-19.34190	-1498.0	0.138156	-0.990410	0.0	-0.99
3	6	2	4	-64.1764	-15.40740	-1498.0	0.233445	-0.972370	0.0	-0.97

Datasets

❑ Hit file

(measured position mm) (pixel location and charge)

	hit_id	volume_id	layer_id	module_id	x	y	z	ncells	pixels
0	1	7	2	1	-63.9659	-3.70513	-1502.5	1	[[141, 605, 0.297491]]
1	2	7	2	1	-40.2738	2.82386	-1502.5	1	[[48, 176, 0.291861]]
2	3	7	2	1	-88.1049	-11.72380	-1502.5	1	[[263, 1044, 0.327308]]
3	4	7	2	1	-39.7041	-8.71702	-1502.5	1	[[279, 182, 0.327097]]
4	5	7	2	1	-30.4918	-8.19262	-1502.5	1	[[283, 18, 0.258165]]

❑ Truth file

(true position mm particle momentum GeV)

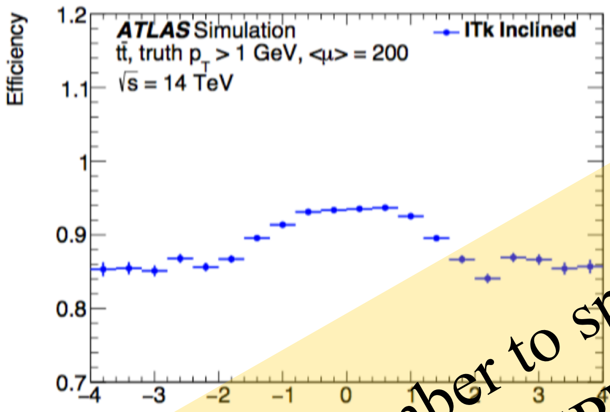
	hit_id	particle_id	tx	ty	tz	tpx	tpy	tpz	weight
0	1	585626006354657280	-63.972698	-3.72889	-1502.5	-0.342366	-0.001899	-7.83544	0.018565
1	2	1035829975879516160	-40.287201	2.84328	-1502.5	-0.366049	0.013878	-13.55470	0.035088
2	3	1080880403243335680	-88.089600	-11.72360	-1502.5	-0.550128	-0.041929	-9.22279	0.018542
3	4	1080909265423564800	-39.712601	-8.71581	-1502.5	-0.363936	-0.094646	-14.01150	0.035088
4	5	1081035022065991680	-30.470400	-8.18647	-1502.5	-0.413489	-0.123403	-20.65790	0.000000

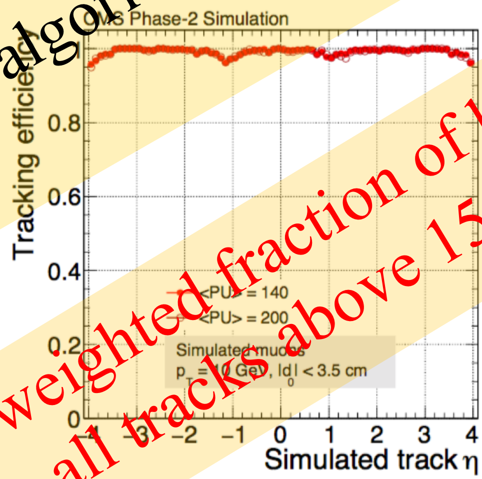
Score

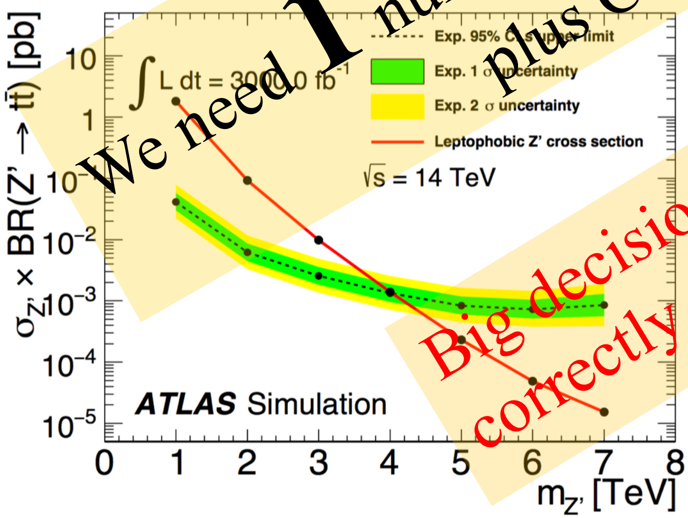
- 2017 CMS tracker Technical Design Report : Chapter 6 expected performance 31 pages 58 figures
- ATLAS Si strip Technical Design Report Chapter 4 ITk Performance and Physics Benchmark Studies 54 pages 80 figures

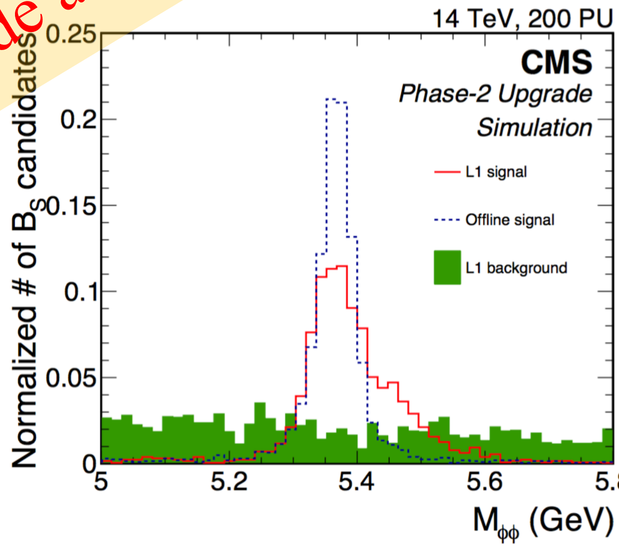
We need **1** number to specify how good an algorithm is!
plus CPU time

Big decision : score is ~ « the weighted fraction of hits correctly associated ». Include all tracks above 150MeV









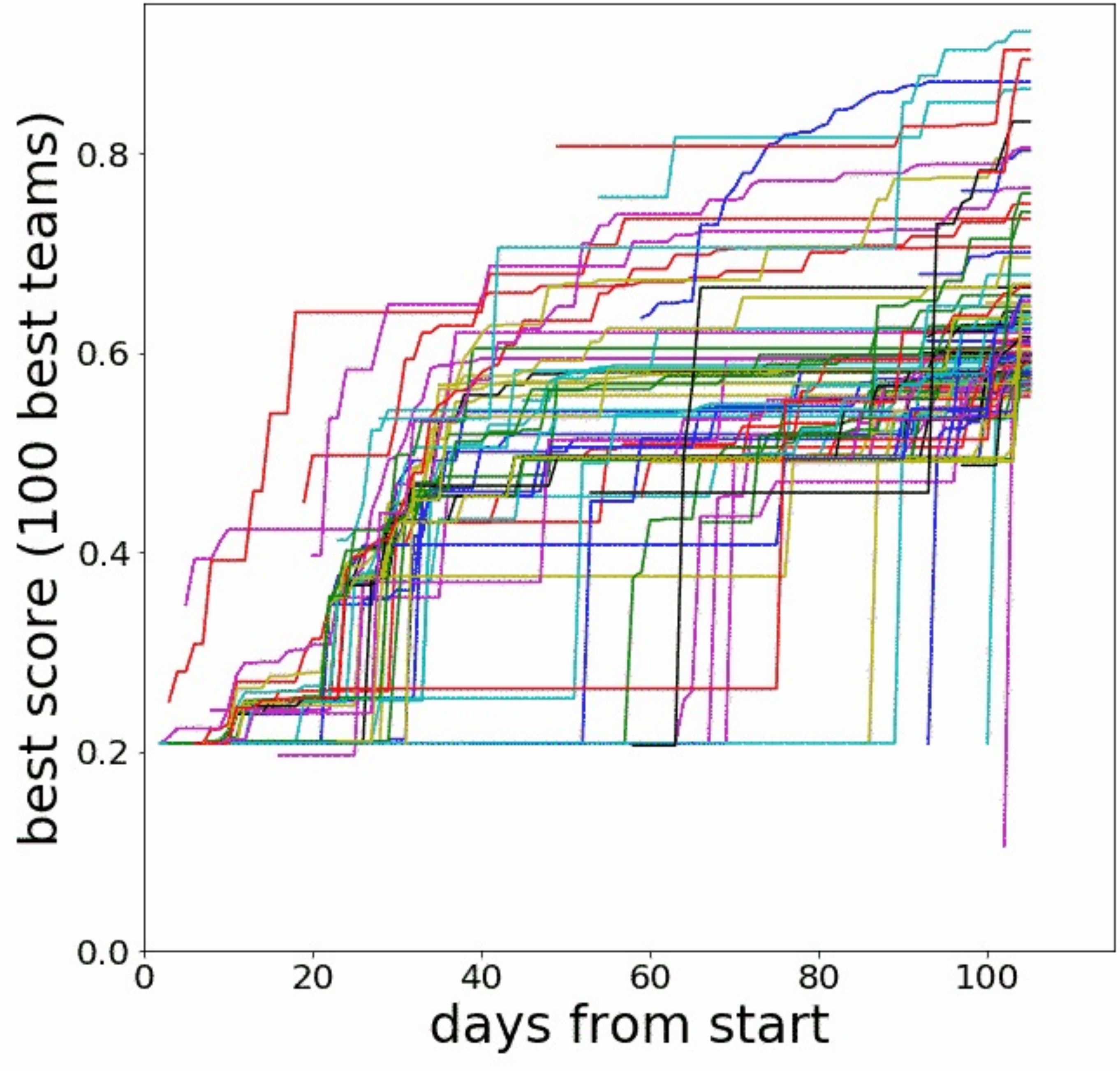
Real life vs challenge

1. Wide type of physics events
2. Full detailed Geant 4 / data
3. Detailed dead matter description
4. Complex geometry (tilted modules, double layers, misalignments…)
5. Hit merging
6. Allow shared hits
7. Output is hit clustering, track parameter and covariance matrix
8. Multiple metrics (see TDR's)

1. One event type (ttbar)
2. ACTS (MS, energy loss, hadronic interaction, solenoidal magnetic field, inefficiency)
3. Cylinders and slabs
4. Simple, ideal, geometry (cylinders and disks)
5. No hit merging
6. Disallow shared hits
7. Output is hit clustering
8. Single number metrics

Simpler, but not too simple!

Evolution of leaderboard



Final Leaderboard

Only public LB to private LB rank change

#	Δ	Team	Score	Entries	Last
1	—	Top Quarks	0.92182	10	19d
2	—	outrunner	0.90302	9	18d
3	HEP —	Sergey Gorbunov	0.89353	6	18d
4	HEP —	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

Experience with first phase

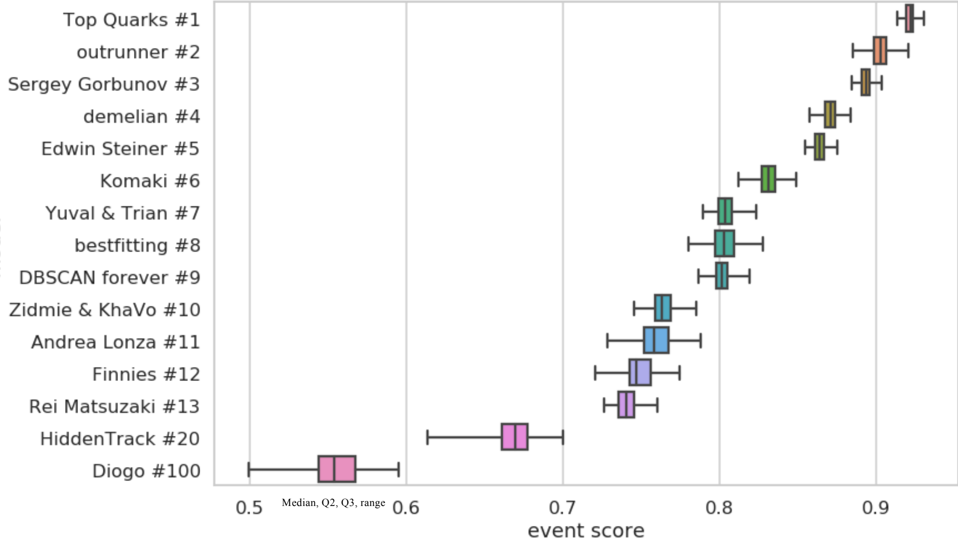
- 630 participants
- Some only downloaded provided solutions, but >100 did provide original code (or tuning of existing code)
- Lots of exchange on the forum
  - People googling courses on HEP tracking…
  - Exchanging ideas, and even code…
  - …up to a certain point (score <=50%)
- A variety of algorithms with various role for ML

A few competitors

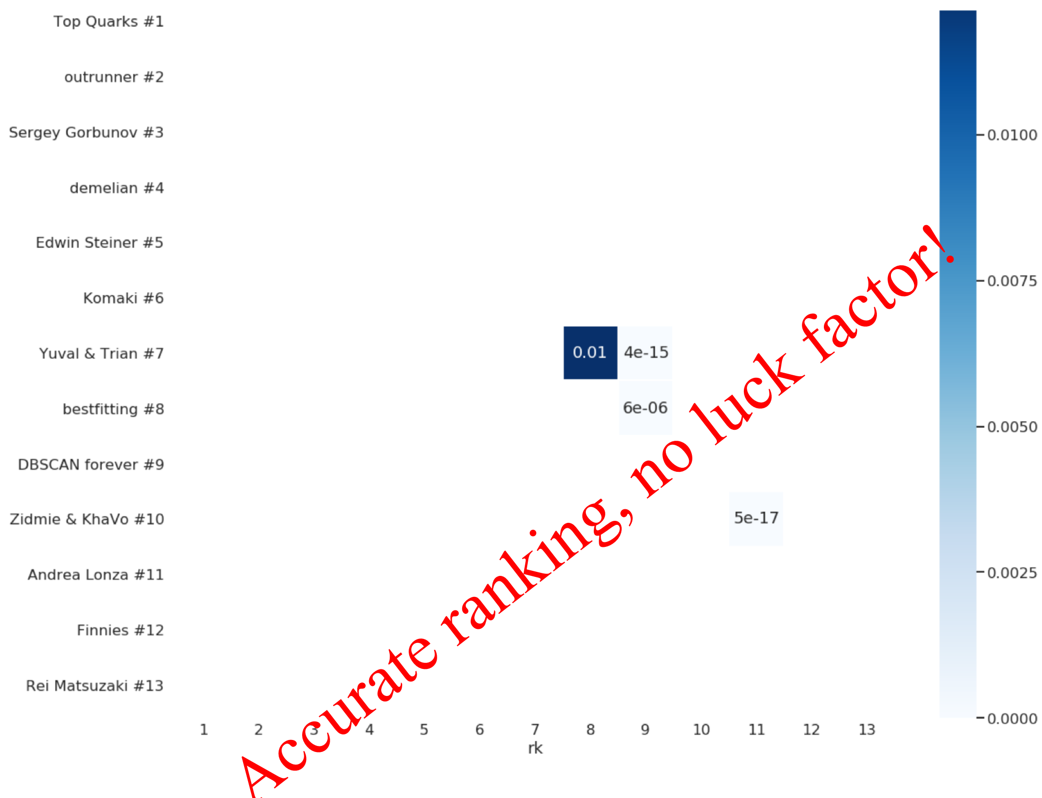
- icecube (Oslo CS master student) #1 92.2 % : combinatorial approach+ML
- outrunner (Taïwan sw engineer) #2 90.3% Deep Learning approach
  - Evaluates probability of any two points to be on same track
  - Very innovative!
  - However takes one full day per event !
- Sergey Gorbunov #3 89.4% demelian #4 87.1% : HEP tracking trigger experts
- Yuval (israëli entrepreneur) & Trian (greek sw engineer) #7 80.4% : unsupervised clustering with randomized projections
- CPMP (french IBM engineer) #9 80.1% : DBSCAN unsupervised clustering algorithm
  - we gave DBSCAN in starting kit, with a 20% score, because in only required a few lines

Ranking Accuracy

Score distributions on 125 test events



Wilcoxon Signed-Rank test



e.g. Participant Data Analysis

We provided a data visualisation notebook:
but
participants did much better within two days:

See link


Data Exploration:

Data

Comments

The (hit_id , x) and (hit_id , y) pairs show us how different volumes are layered.

This figure idea is taken from Joshua Bonatt's notebook.

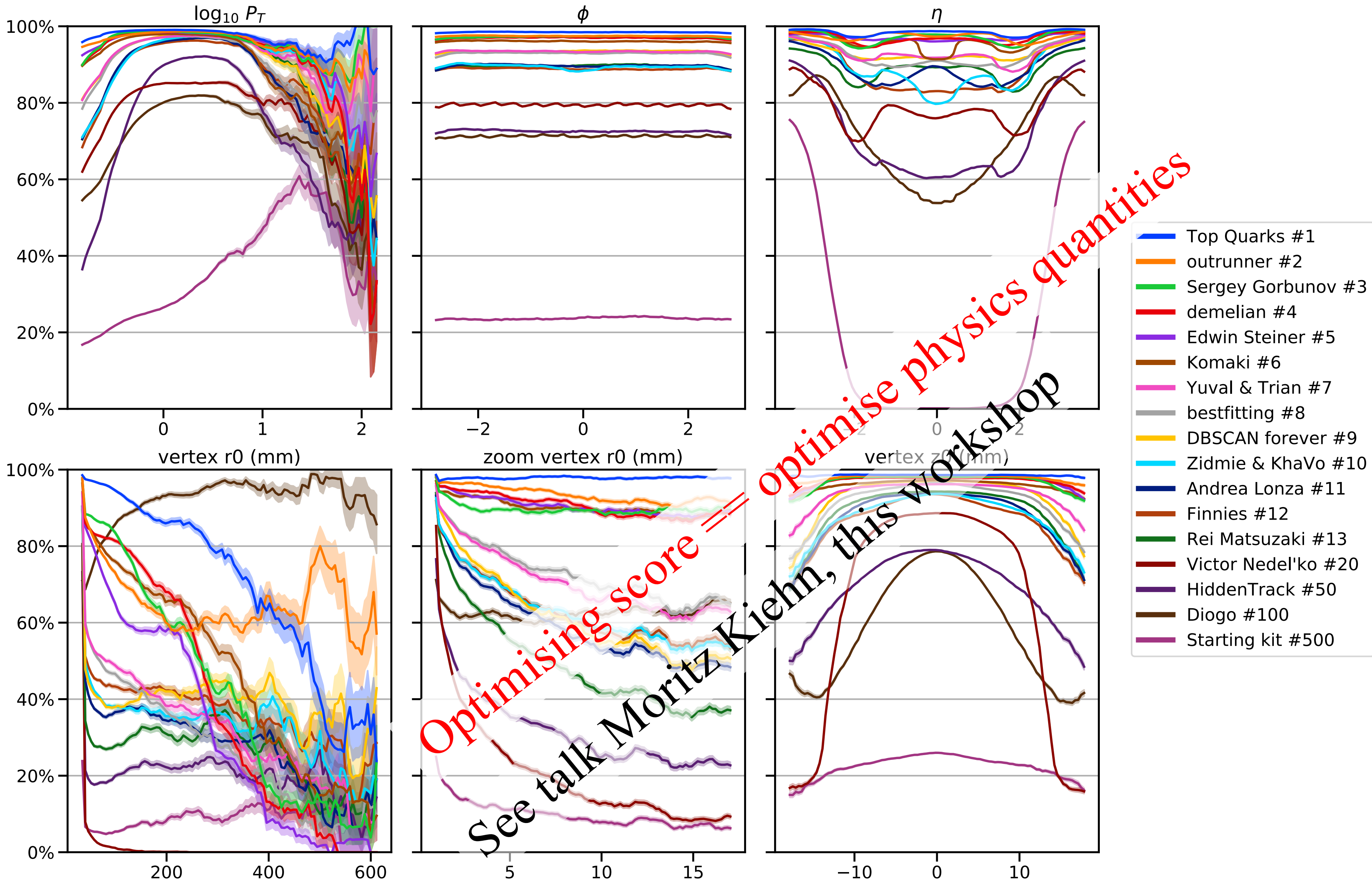
Code

Particle Data

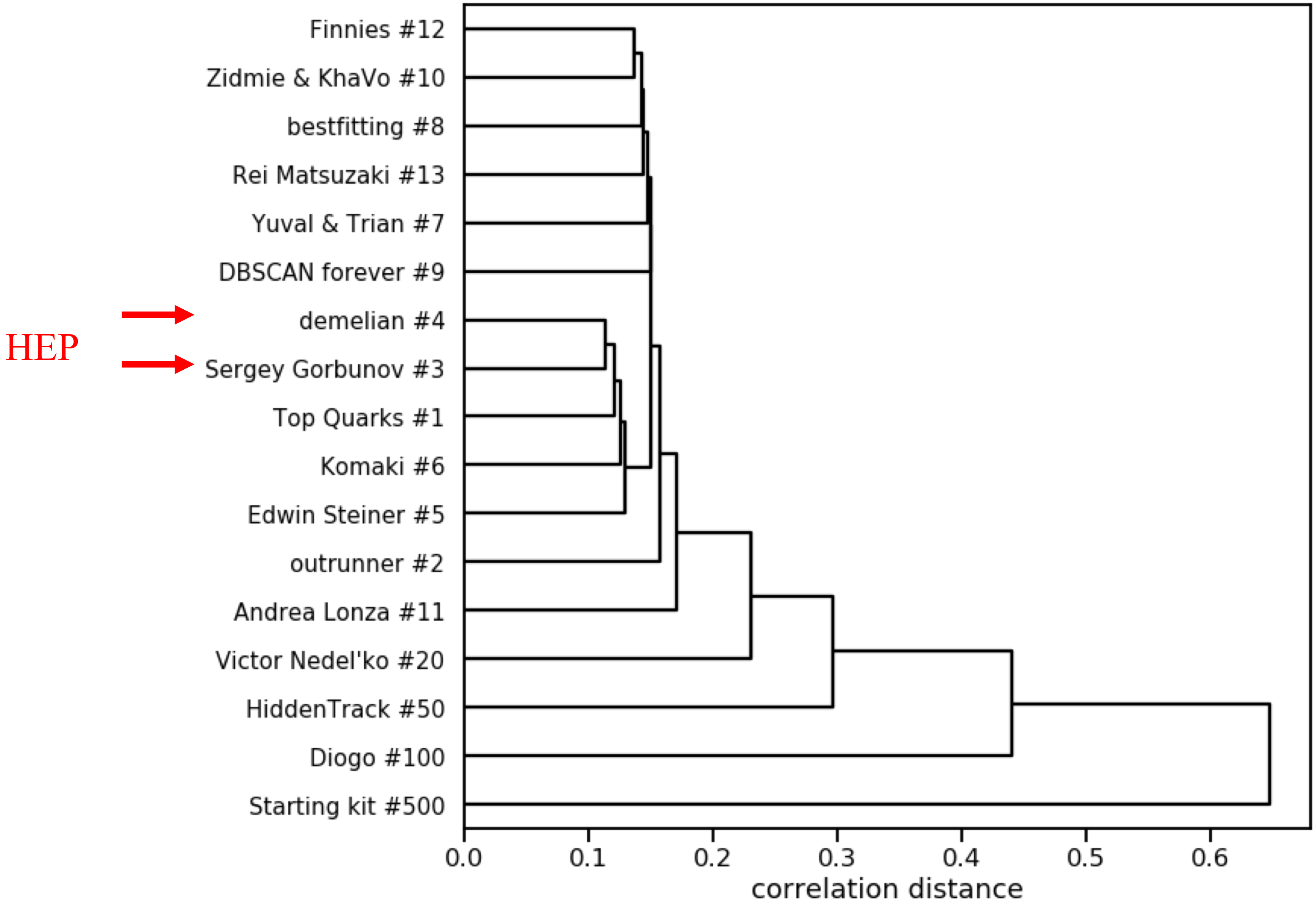
The particle data help us understand each particle's initial position, momentum, and charge, which we can join with the event truth data set to get the particle's final position and momentum. This is needed to identify the tracks that each particle generated.

The data look like this:

Efficiency all



Participants dendrogram



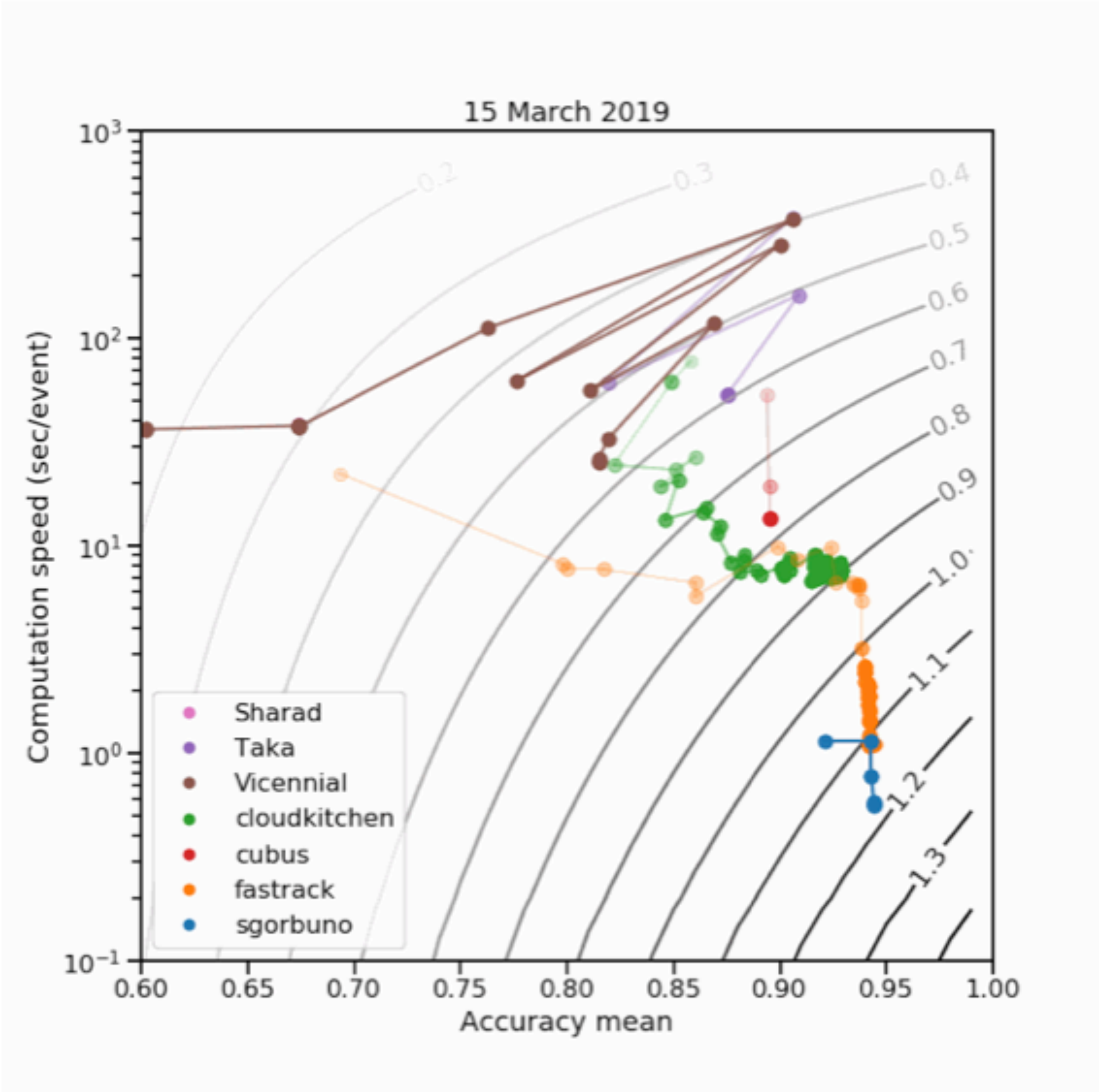
Throughput Phase

Now participants submit their software…
… and are evaluated on accuracy AND speed !

Launched 6th Sep 2018 until 12th March 2019 on Codalab

Throughput platform

- Kaggle initially told us they would also provide the speed estimate…
- …but they suddenly declined.
- …so we did it ourself on Codalab, with U Paris-Sud resources.
- Specific difficulties:
  - Speed measurement reproducibility no better than 3% (even on dedicated machine)
  - Many hacks anticipated (e.g. dumping the data in the log file…)
  - More hacks for sure…
- ➔decision : remeasure speed at the end of the competition many times on a dedicated machine
  - ➔it worked
- Providing for competition with accurate online time measurement is an open problem (Kaggle is working on it, given the demand, see e.g. « the Airbus Ship Detection challenge »)



Accuracy mean

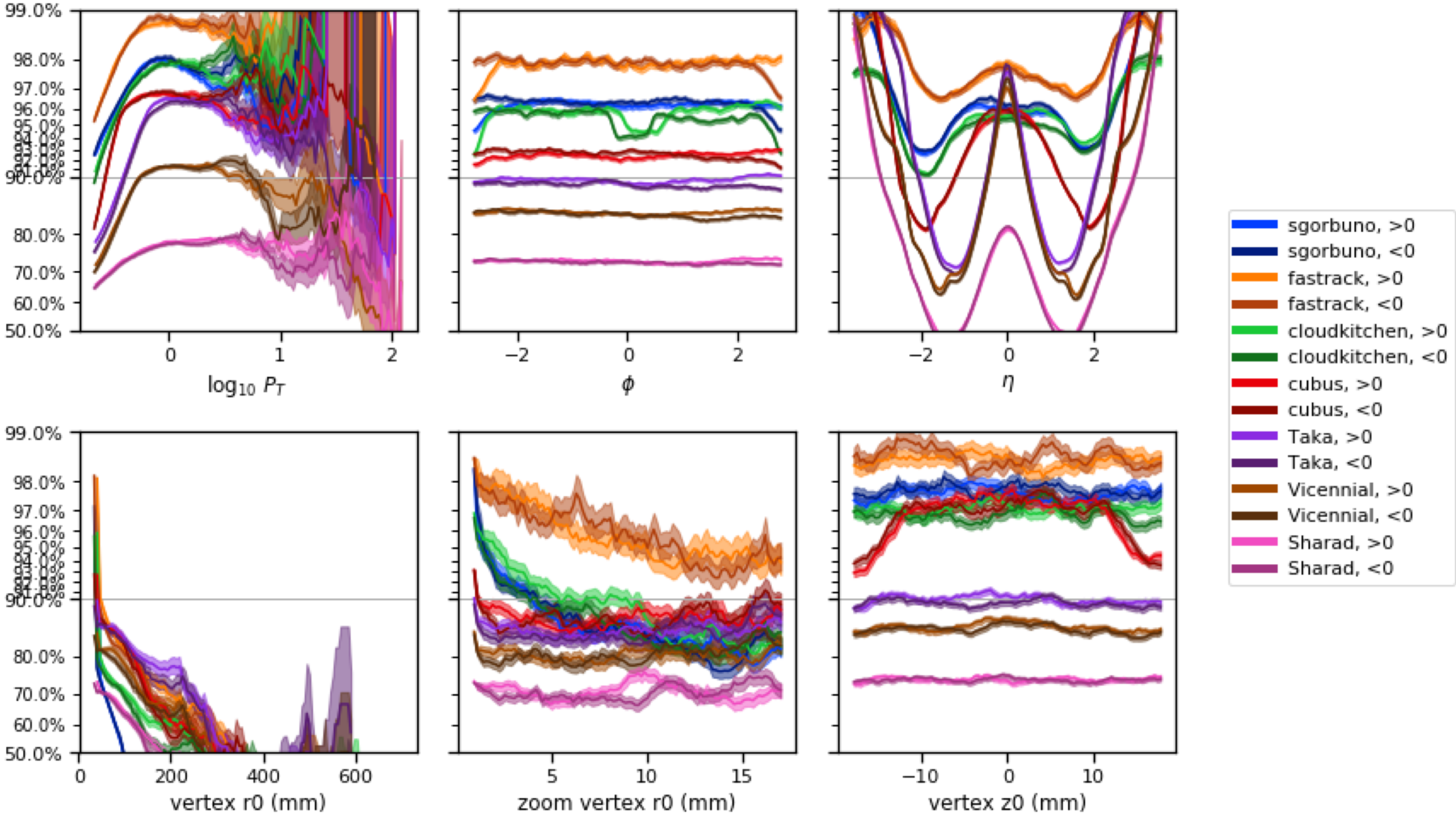
Throughput phase LB

RESULTS

Private leaderboard

#	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) 1.16	0.944 (2) 0.943	0.00 (14)	28.06 (1)	0.56 (1) 0.60	64.00 (1)
2	fastrack HEP	53	03/12/19	1.1145 (2) 1.12	0.944 (1) 0.944	0.00 (15)	55.51 (16)	1.11 (16) 1.00	91.00 (6)
3	cloudkitchen HEP	73	03/12/19	0.9007 (3) 0.897	0.928 (3) 0.927	0.00 (13)	364.00 (18)	7.28 (18) 7.41	407.00 (8)
4	cubus	8	09/13/18	0.7719 (4) 0.770	0.895 (4) 0.895	0.01 (9)	675.35 (19)	13.51 (19) 13.7	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					

Throughput phase Efficiency



Not quite as good as for accuracy phase: do not spend time when not worth it

Where did ML people go ?

- 100 participants registered on Codalab but only 10 submitted non trivial code. Why ? Our guesses:
  - Kaggle visibility vs Codalab visibility.
  - On Kaggle people win points across competition, can access « Grand master status », etc… very valuable on their CV
  - « Professional » kagglers move from one challenge to the next. No interest in long term involvement
    - (still we had some praises like « most interesting challenge I had ever done »)
  - Codalab is a research platform
  - No GPU (while ML code « naturally » run on GPU)
  - C++ vs python : python was allowed but people realise they had to write in C++ for speed. Many ML people do not know C++
  - Not completely trivial effort to properly wrap code for submission

HEP wins at the end

- The podium are HEP experts. Was it worth it ?
- Definitely : best solutions in <1 s to be compared to >10 s for ATLAS or CMS (order of magnitude comparison)
- HEP people liked the gamification of the problem.
  - Also one is ALICE, one is ATLAS, one is Computing Center management.
- The dataset will be released on CERN Open Data portal for future development
  - Already used in research papers e.g. tracking with quantum computing (see talk in CERN Grand Finale workshop)
- On going work to integrate the best ideas (of both phases) in future algorithms for ATLAS and CMS (many discussion in that workshop)

TrackML Conference talks

- Connecting The Dots 2015 Seattle
- Connecting The Dots 2016 Vienna
- CHEP 2016 Okinawa
- Connecting The Dots / Intelligent Trackers 2017 Orsay
- NeurIPS 2017 Los Angeles CiML workshop
- Connecting The Dots 2018 Seattle
- CHEP 2018 Sofia
- WCCI 2018 Rio de Janeiro
- ICHEP 2018 Seoul
- IEEE NSSMIC 2018 Sidney
- IEEE eScience 2018 Amsterdam
- NeurIPS 2018 Montreal Competition workshop
- ACAT 2019 Saas-Fe
- Connecting The Dots 2019 Valencia
- EPS 2019 Ghent
- CHEP 2019 Adelaïde
- …and much more workshops and seminars….

Useful links

- See also Laurent Basara's talk in Detector and Data Handling session Friday 12:45, about the algorithms exposed
- 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 NeurIPS 2018 Competition book arXiv:1904.06778 final version just released
- Throughput phase @ Codalab : https://competitions.codalab.org/competitions/20112
  - Write-up being finalized
- CERN Grand Finale workshop 1-2 Jul 2019 : https://indico.cern.ch/event/813759/

Wrapping-up

How to design a scientific data challenge ?

No recipe!


프랑수아 클루제
오마
Le Républicain Lorrain
상위 1% 귀족남과
하위 1% 무일푼이 만났
2100만명을 무장해제시킨 코믹감동실화
언터처블
1%의 우정
UNTOUCHABLE
Written and directed by Eric Toledano and Olivier Nakache
3월 22일, 전대미문 흥행 센세이션을 만난다

Still some guidelines

- An interesting scientific problem
  - For science (e.g. physics)
  - For Computer Science
- Personpower to create the challenge (2-3 FTE year for HiggsML 5-6 for TrackML)
- Focus on one problem, and simplify
- A realistic (high statistics) dataset
- One figure of merit
  - Should match the scientific expertise
  - Easily understandable for non experts
- A challenge platform
  - Kaggle very popular, can fund prize money
  - RAMP or Codalab more flexible
- A starting kit to kick-start people (not too directive)
- A competitive market ! Another challenge is one click away
  - Simple explanation "for the dummies"
  - Submission of trivial solution should be possible in a few minutes
  - Communication !
- Monitor competition, be responsive on forum
- Take care of most active participants
- Most important is what happens after the competition :
  - long lasting collaboration with participants
  - long lasting use of dataset