Tracking Machine Learning challenge status

David Rousseau (LAL-Orsay) with Markus Elsing, Vincenzo Innocente, Andreas Salzburger (CERN), Jean-Roch Vlimant (CalTech) Cécile Germain, Balazs Kegl, Yetkin Yilnaz (LAL/ LRI-Orsay) , Isabelle Guyon (Chalearn/LRI-Orsay) Paolo Calafiura, Steve Farell (LBNL), Michael Kagan (SLAC), Davide Costanzo (UCL), Tobias Golling, Moritz Kiehn, Sabrina Amrouche (U Geneva), Amir Farbin (UTA), Mikhail Hushchyn, Andrey Ustyuzhanin (YandexDSA)

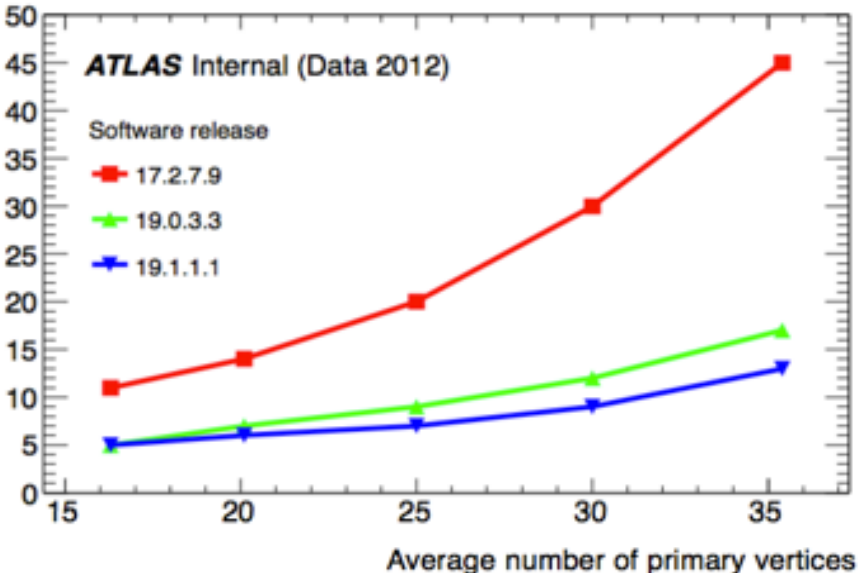
CTDWIT 2017, LAL-Orsay **7th March 2017**

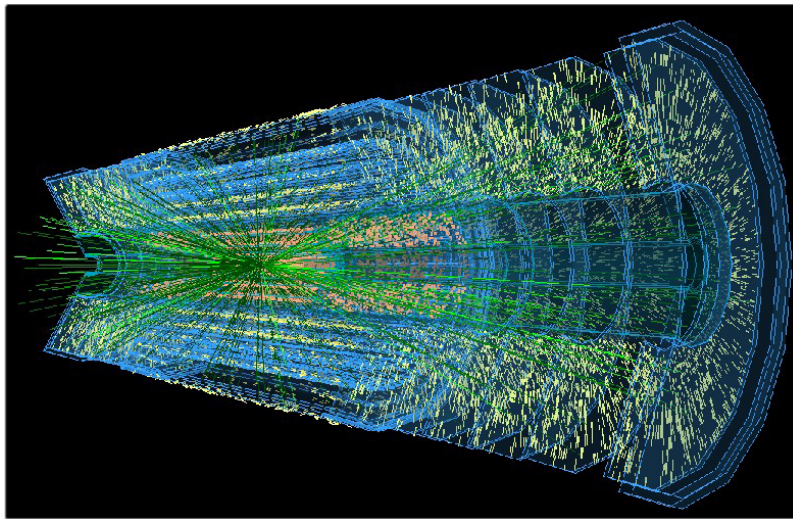
Outline

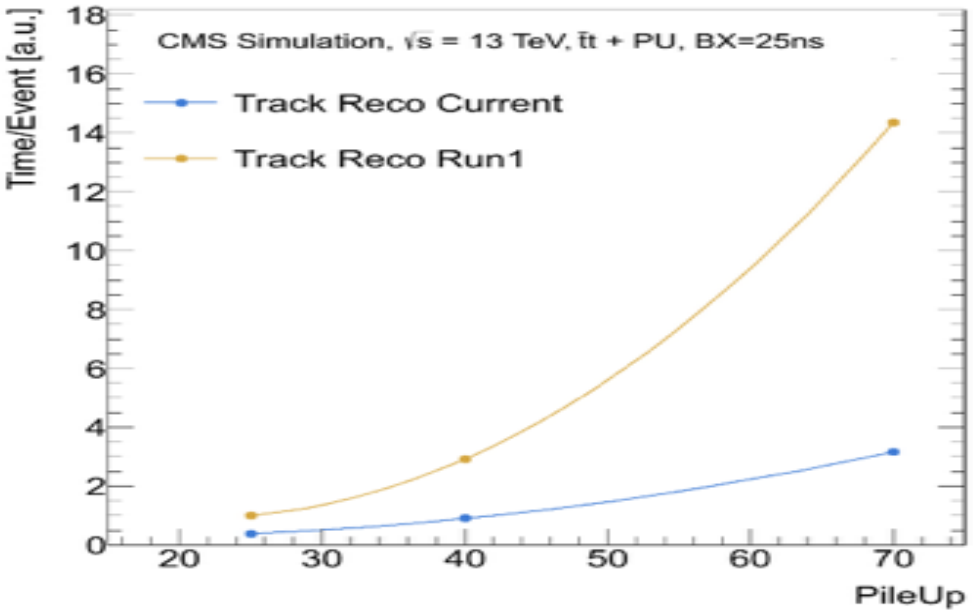
- Tracking Machine Learning challenge suggested at CTD2015 in Berkeley
- …discussed in CTD2016 in Wien
- Status now
  - Motivation
  - Current baseline and open questions
- TrackMLRamp introduction

Motivation 1

- 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): <>~30
  - Phase 2 (2025): <>~150
- CPU time quadratic/exponential extrapolation (difficult to quote any number)







Motivation 2

- LHC experiments future computing budget flat (at best)
- Installed CPU power per $==€==CHF expected increase factor ~10 in 10 years
- Experiments plan on increase of data taking rate ~10 as well (~1kHz to 10kHz)
- ➔HL reconstruction at mu=150 need to be as fast as Run1 reconstruction at mu=20
- ➔requires very significant software improvement, factor 10-100
- Large effort within HEP to optimise software and tackle micro and macro parallelism. Sufficient gains for Run 2 but still a long way for HL-LHC.
- >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 (i.e. Convolutional NN)
- Need to engage a wide community to tackle this problem

Interacting with Machine Learners

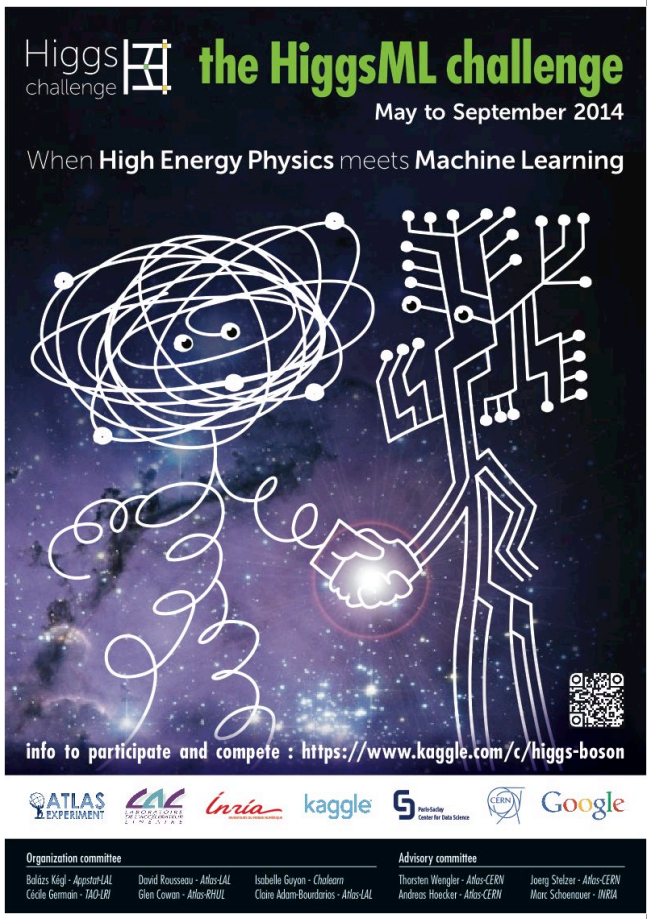
- Suppose we want to improve the tracking of our experiment
- We read the literature, go to workshops, hear/read about an interesting technique (e.g. RANSAC, ConvNets…). Then:
  - Try to figure by ourself what can work, and start coding➔traditional way
  - Find an expert of the new technique, have regular coffee/beer, get confirmation that the new technique might work, and get implementation tips➔better
  - (and in fact we see more and more ML in tracking as clear in this workshop)
- …repeat with each technique...
- Much much better:
  - Release a data set, with a figure of merit, and have the expert do the coding him/herself
  - ➔ he has the software and the know-how so he'll be (much) faster even if he does not know anything about our domain at start
  - ➔engage multiple techniques and experts simultaneously (e.g. 2000 people participated to the Higgs Machine Learning challenge) in a comparable way
  - ➔even better if people can collaborate
  - ➔a challenge is a dataset with a buzz

Motivating Machine Learners

- Why would ML experts spend days/week/months working for free on our problem?
  - Interesting new problem (for them)
  - Potential for publications (beware of experiment policies)
  - Prestige
    - High Energy Physics, "CERN", "Higgs"
    - Kaggle : HiggsML winner was hired by DeepMind, runner-up hired by OpenAI, XGBoost co-author got a US visa and a PhD grant
  - (money)
- The key is the dataset and associated material (figure of merit)
- The challenge is just a way to advertise the dataset, and organise the collaboration between experts
- The learning threshold to participate should be as low as possible (to entice the experts to spend time on our challenge not another):
  - Relatively easy for a classification problem
  - Less so for a tracking challenge (no on the shelf solution)
- In particular, things should be presented with ML vocabulary e.g. "classifier" instead of MVA, "feature" instead of "variable", "false positive" instead of "accepted background", etc….

Higgs Machine learning challenge

- See talk DR CTD2015 Berkeley
- An ATLAS Higgs signal vs background classification problem, optimising statistical significance
- Ran in summer 2014
- 2000 participants (largest on Kaggle at that time)
- Outcome
  - Best significance 20% than with TMVA
  - BDT algorithm of choice in this case where number variables and number of training events limited (NN very slightly better but much more difficult to tune)
  - XGBoost best BDT on the market (quite wide spread nowadays)
  - Wealth of ideas, documented in JMLR proceedings v42
  - Still working on what works in real life what does not
  - Raised awareness about ML in HEP



Final leaderboard

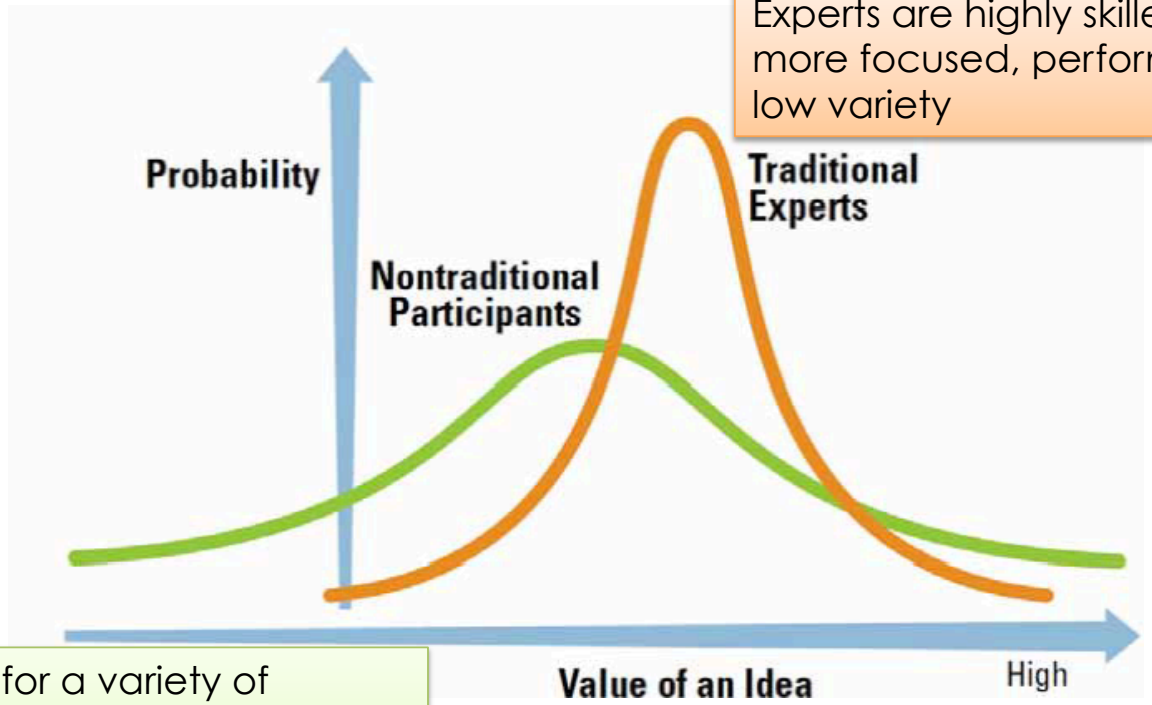
#	Δrank	Team Name ‡ model uploaded * in the money		Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑1	Gábor Melis ‡ *	7000$	3.80581	110	Sun, 14 Sep 2014 09:10:04 (-0h)
2	↑1	Tim Salimans ‡ *	4000$	3.78913	57	Mon, 15 Sep 2014 23:49:02 (-40.6d)
3	↑1	nhlx5haze ‡ *	2000$	3.78682	254	Mon, 15 Sep 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	Best physicist	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	TMVA expert, with TMVA improvements	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		

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 nonconventional surprising ideas that are « far » from traditional expertise - > high volatility

Not just ML, but a general trend: Open Innovation

From domain to challenge and back

Domain e.g. HEP

Challenge organisation

Challenge

Problem

simplify

Problem

Domain experts solve the domain problem

4 months?

The crowd solves the challenge problem

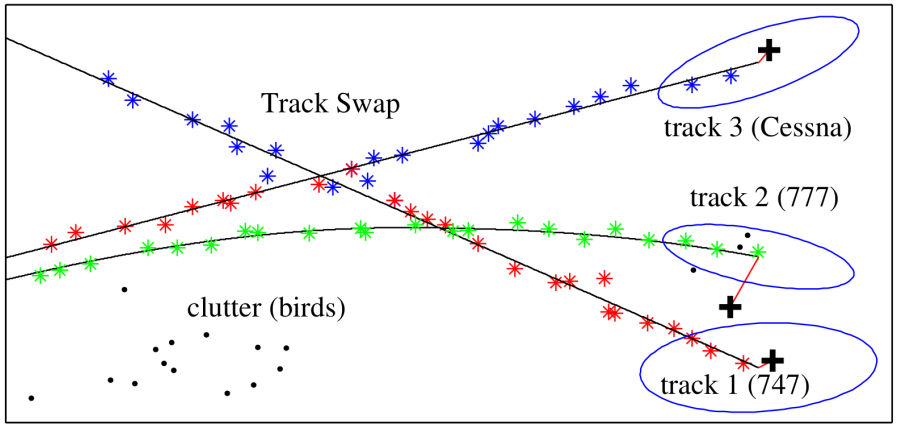
reimport

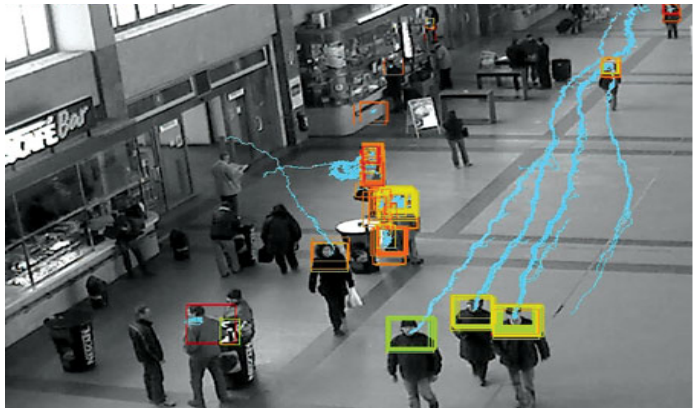
Solution

Solution

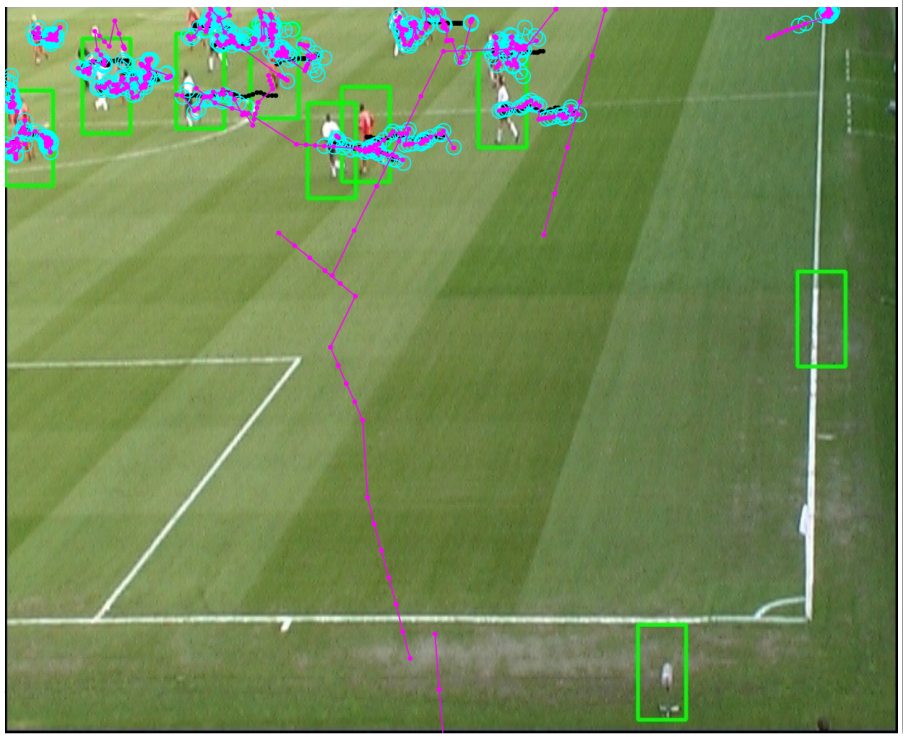
Tracking in ML

- Tracking outside HEP means very different things (see earlier talks)
- These applications also have CPU issues (online)
- Big difference compared to HiggsML classification problem: no off-the-shelf algorithm

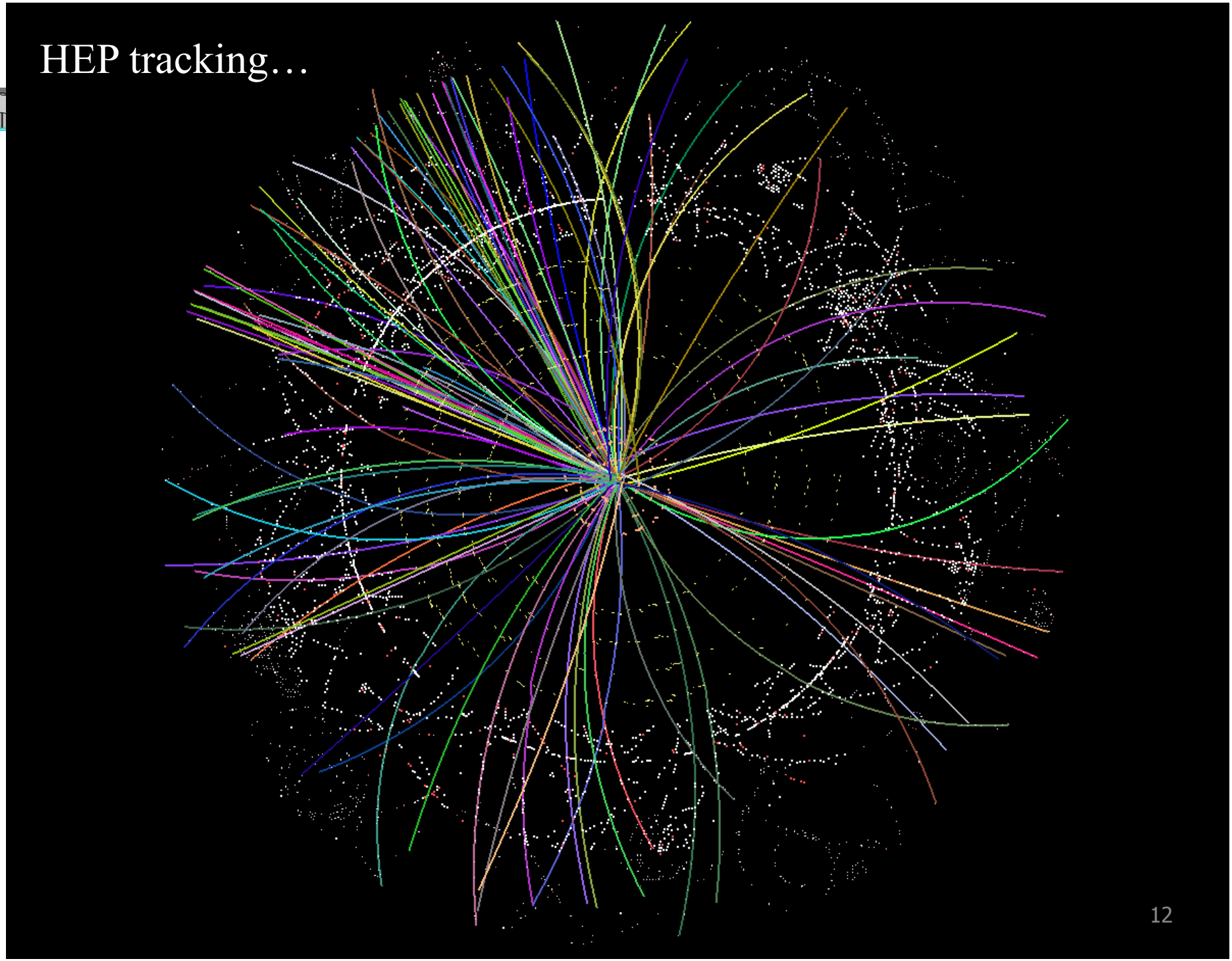




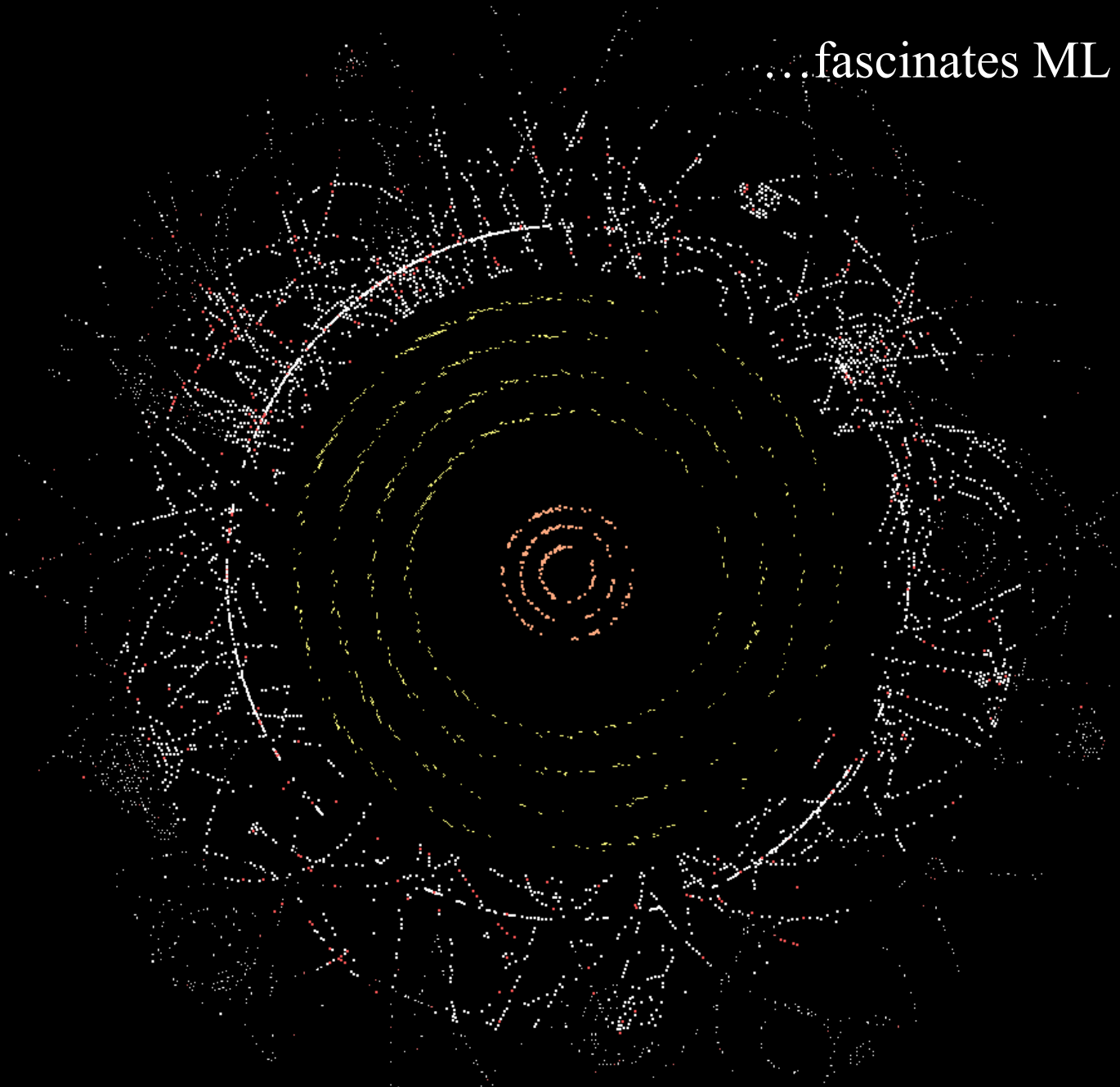
Tracking challenge



HEP tracking…



…fascinates ML experts



TrackML : current thinking

- Use ACTS (check Andreas Salzburger talk on Thursday) to generate fast simulation of a generic Silicon detector at HL-LHC (cylinder and disks)
  - With non uniform magnetic field, multiple scattering, energy loss, cluster smearing, nuclear interaction
  - ➔simplified simulation but not too simple (otherwise a simple Hough transform would probably work)
  - ➔ATLAS and CMS collaboration
  - "cheap" but realistic events which do not "belong" to any collaboration
- Dataset:
  - 3D points and truth track parameters, also tracking surfaces and dead material. ASCII
  - (also thinking of giving cluster splitting as a sub-task)
  - Type of events to be decided, typically pythia tt (includes e/μ/τ/b/jets), +~200 pileup (~10.000 tracks/event)
  - Number of events to be decided but we should not be shy on the training sample (people have been complaining about the lack of data in previous challenges, preventing use of e.g. deep learning)
  - But ~0.5MB per event, so 1 million event ➔500GB (could have in addition a ~GB one for easy access)
  - Instead or in addition to dataset, can distribute simulation engine (ACTS)
  - ~seconds to produce one event➔~day for the whole sample
  - A separate small test sample without the truth : higgs ? Black hole ?
- Participants are given the test sample. They should upload the tracks they have found
  - A track is a list of points belonging to it
  - We don't ask for track parameters, nothing will beat Kalman filter (right ?)
  - Figure of merit built from efficiency, fake rate, CPU time (see later)
- Leaderboard automatically updated

Figure of merit

- We're more interested by CPU gain, than efficiency or fake rate reduction, provided they are "good enough"
  - Our algorithms find correctly tracks in HL-LHC, they are just not fast enough by factors
- The f.o.m should favour the algorithm which is the most likely to become part of the HL-LHC ATLAS/CMS reconstruction
  - (even if we keep only the algorithm ideas and rewrite the software)
- We're focussing on pattern recognition, presuming it is followed by a kalman filter
- Actually efficiency more important than fake rate (fakes and duplicates can be removed later on by the final fit)
- Penalise holes on tracks, especially in the inner layers
- Bulk of the tracks are pileup tracks, but they are the uninteresting ones! ➔Need to give more weight in the evaluation to:
  - Higher momentum tracks
  - Tracks in dense jets
  - Tracks from displaced vertices, b, tau (not K0s, conversion ?)
  - Electrons (?)
  - Most tracks from pileup vertices should still be found
- ➔ current baseline : define a few bins based on the above. Define $\varepsilon = \min \varepsilon_i$ ➔algorithm should have homogeneous efficiency across the board
- Deal with fake rate with something like: f.o.m= ε-a*fake rate (with a~1/10)
- (another option is to set a hard limit on fake rate but challenge professional do not like that, because it introduces a risk factor)

CPU measurement

- Contrary to HiggsML or flavour of physic challenge need to evaluate CPU time
  - CPU time to find the tracks
  - Cap on memory used (e.g. one x86-64 core with 2GB)
  - Training time unlimited
- Some platforms (see AutoML, Codalab, Topcoder) now allow to automatically upload, compile and run software
  - ➔well defined hardware (CPU and memory available)
  - ➔uniform comparison
  - Could also use an Amazon instance
    - (amazon could also host the large dataset and allow resources for the training)
  - Use e.g. Docker for software packaging
- Positive side-effect : limit diversity of software languages and libraries
- We're more interested in the detailed algorithm (as it would be explained in a technical paper) rather than the software itself (we do want to see the software)
- We're more interested in new approaches than in super-optimised version of old approaches

Challenge sequencing

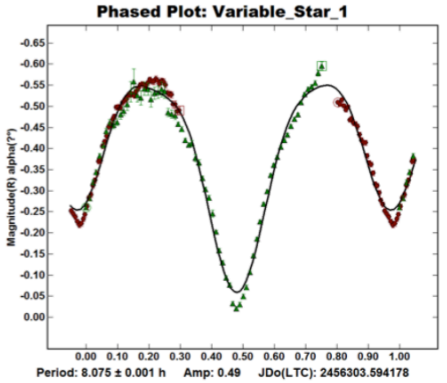
- Building collaborations more important than the competition
- Strive to promote "coopetition" so that participant collaborates (tricky): example in HiggsML challenge XGBoost released very early and use by many participants
- ➔foresee long term interaction between participants and HEP
- ➔dataset and evaluation should remain available after the challenge is done
- Baseline is to have a two step challenges (best algorithm designers might not be expert in code optimisation)
  - First step with little CPU component➔goal is to see new ideas
  - Second step with strong CPU component
- Which platform ?
  - Kaggle ? Informal talk : >100.000$
  - Codalab (Isabelle)
  - Yandex/Everware (Andrei/Mikhail)
  - RAMP (Balazs)
- Can foresee several prices for different category
- Foresee of releasing the sample publicly (e.g. CERN Open Data Portal) just after the challenge
- Foresee a publication outlet (e.g. a satellite NIPS workshop proceedings or see later)
- Anticipate from the very beginning the final re-import stage

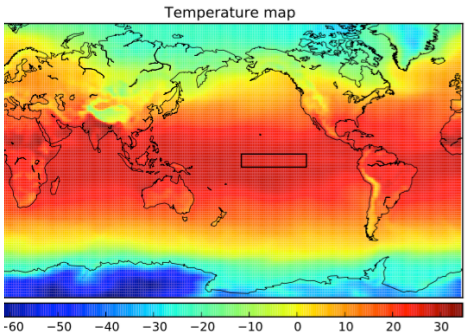
TrackMLRamp introduction

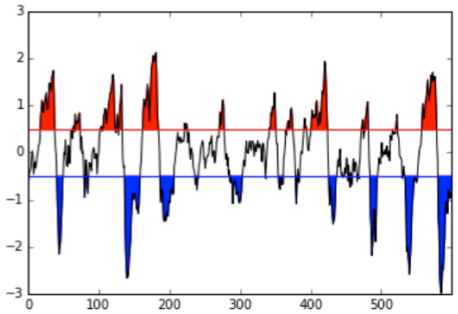
After coffee break in room 101

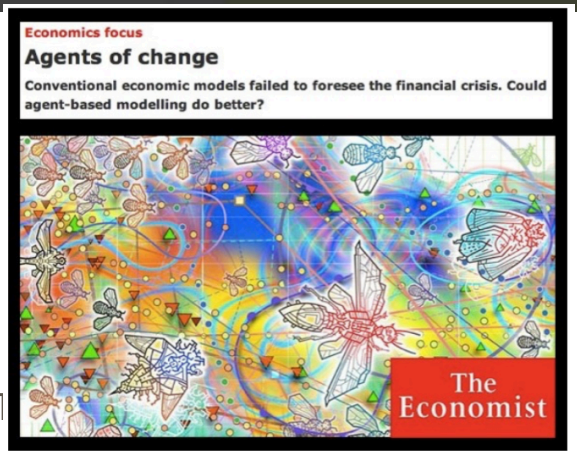
Beyond challenges : RAMP

- http://www.ramp.studio : Rapid Analysis and Model Prototyping
- Run by CDS Paris Saclay
- Main difference wrt Kaggle-like challenge
  - participants post their software (python), which is run by the RAMP platform (training+testing)
  - one day hackathon
  - participants are encouraged to re-use other people's software
- Can adapt to all domains:







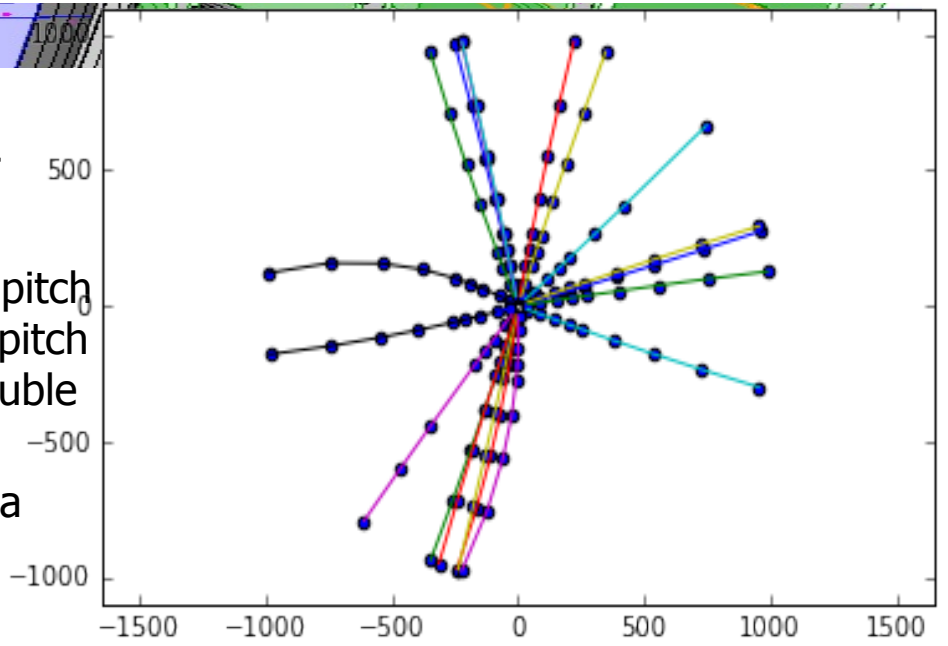


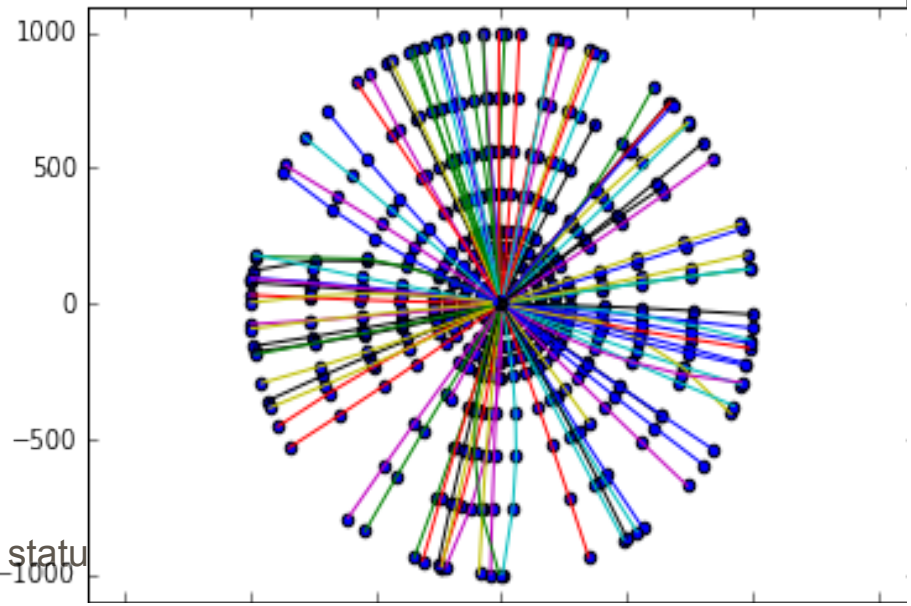
TrackMLRamp mini challenge for CTD/WIT

- Allow us a "dry-run" for the real challenge, without giving away data similar to the challenge
- Very simplified problem, so that significant results can be obtained in 2-3h (?) (but not too much so that trivial method do not work)
- Audience are HEP tracking experts, no need to prepare tracking introduction for ML/ non HEP
- 42 people registered
- In practice, only 2-3 hours : do not expect brand new methods to be coded from scratch
- Will remain open : winner will be whoever (not organisers) has the best score at 9AM Thursday
- Will remain open even later : interesting playground for ideas

TrackMLRamp 2D simulation

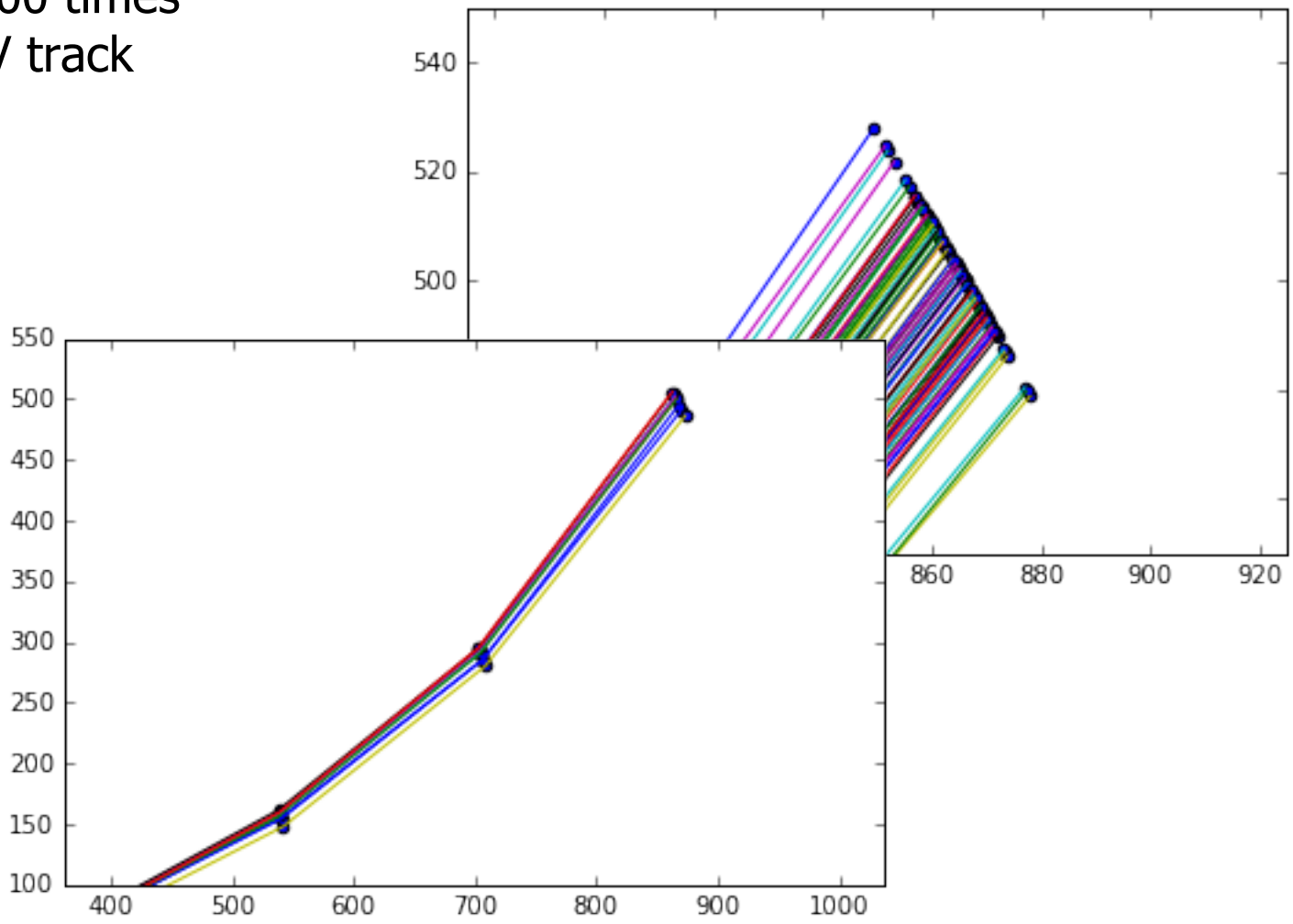
- Written in python (numpy, pandas)
- 2D simulation, detectors are perfectly circular
- Unit mm and MeV
- Use typical HL-LHC detector layout : 5 layers pitch 25um, radii {39,85,155,213,271} , +4 layers pitch 50um radiii {405,562,762,1000} (simulate double layer strip 75um)
- Digital read out : a hit is a "pixel" crossed by a track
- Constant magnetic field 2T
- Multiple scattering 2% radiation length each layer: σ_ϕ=13.6 MeV $\sqrt{(0.02)}$/P (MeV)
- Hit inefficiency 3%
- Particle stopping probability 1% per layer
- Particle gun :
  - uniform phi distribution baseline
  - Poisson ~10 tracks per event
  - Momentum : flat 300 MeV to 10 GeV
  - Origin vertex spread : σ_x=σ_y =2/3. mm
    - Each track has a different vertex



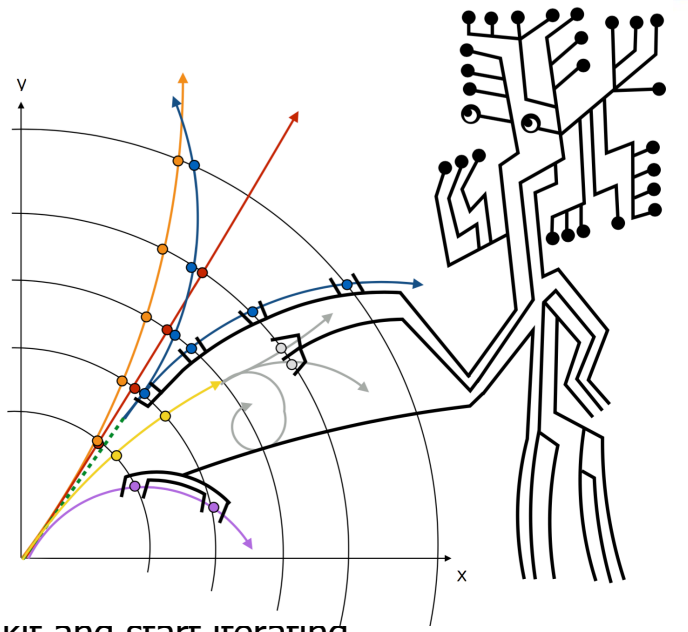


2D simulation:MS

- Evidence of multiple scattering : 100 times same 300MeV track



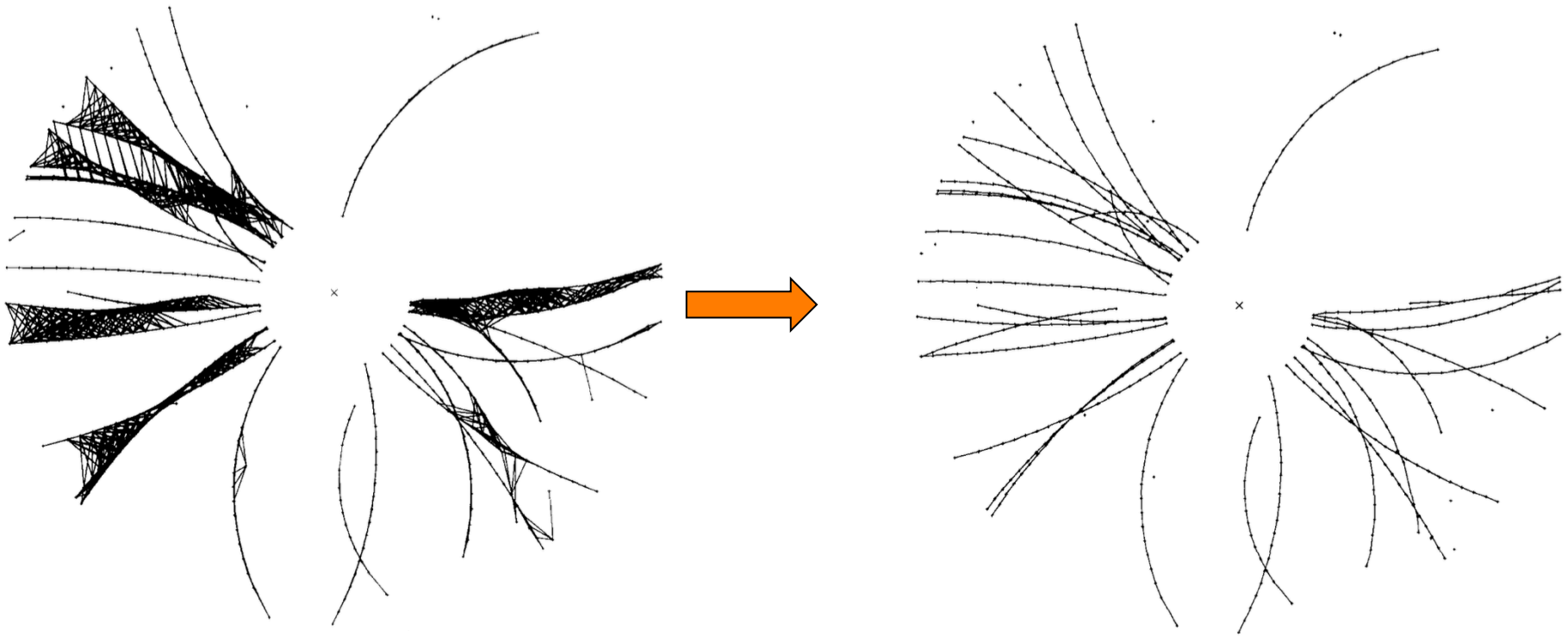
Conclusion



- Track Machine Learning challenge taking shape
- Broad lines are defined but still a lot of work:
  - finalise the figure of merit (including CPU)
  - write documentation
  - starting kit software (many)
  - development/test platforms
  - budget (>>10k€), sponsoring
  - legal matter
  - outreach, challenge publicity, social media
  - post challenge
- Most important is what remains after the challenge
- Next step is to define a v0 of dataset, testing framework, starting kit and start iterating
- Define work packages and implement
- If you want to help organise this challenge:
  - Register to trackml-challenge@googlegroups.com (note that you'll not be able to claim a prize)
  - Regular bi-weekly next Tuesday 5PM
- Call for NIPS competition just out https://nips.cc/Conferences/2017/CallForCompetitions
  - Submit by 15 March 2017
  - Competition to be finished by end October 2017
  - Competition' competition in NIPS, then publication spring 2018
- Becoming to be tight but worthwhile

Back-up

An early attempt

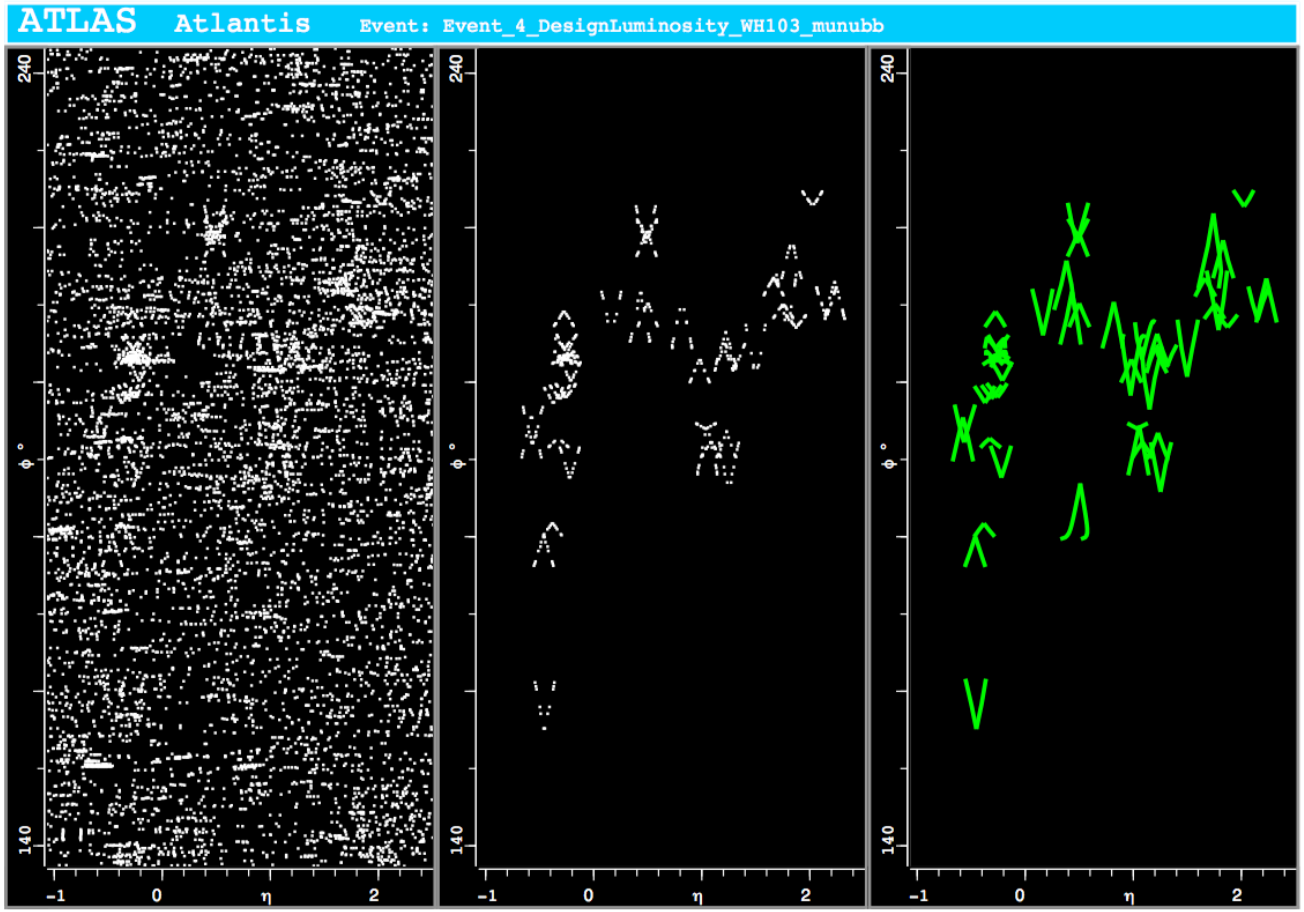


- Stimpfl-Abele and Garrido (1990) (ALEPH)
- All posssible neighbor connections are built, the correct ones selected by the NN
- Also PhD Vicens Gaitan 1993, winner of Flavour of Physics challenge

V plots: connection to computer vision ?

- Computer vision : try to do as well as human
- Tracking : tracks are not visible by eye!

CHEP04

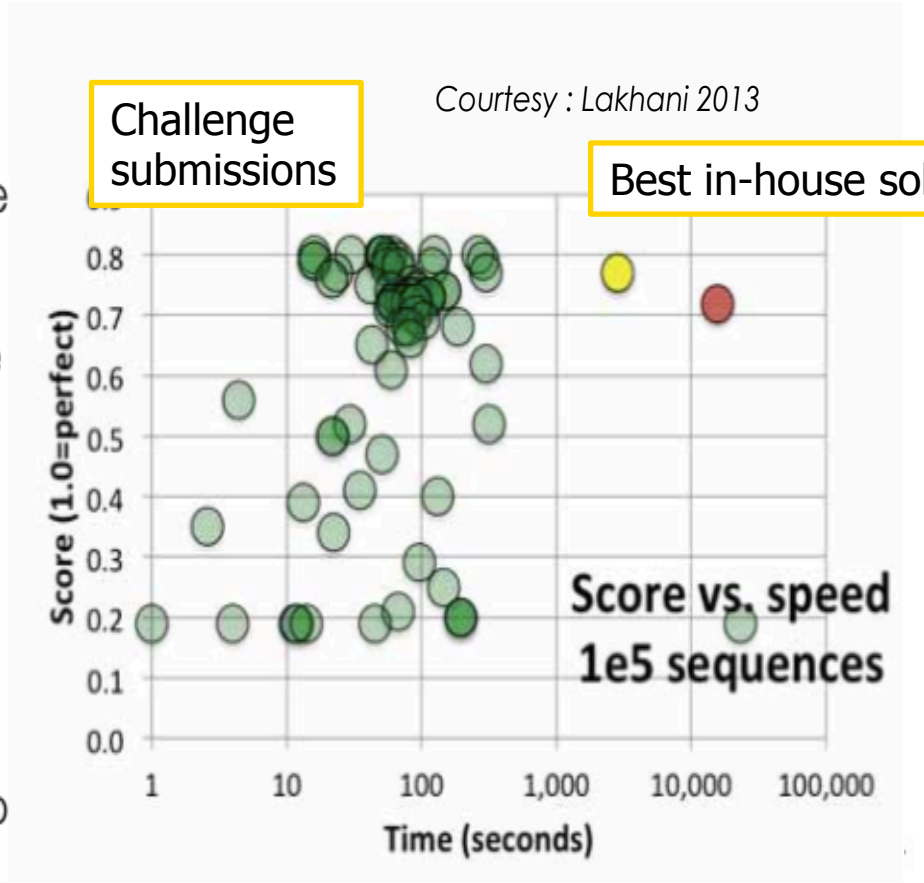


- Hans Drevermann, ALEPH/DALI then ATLAS/ATLANTIS event display
- Eta phi projection with $\delta\eta = +/- \varepsilon(r_{max} - r)$

CPU challenges have been done

Harvard Medical School Contest for Biology Big Data Problem in Genomics
Two week long competition - $2000 prize pot x 3 on TopCoder.com

- 122 coders submitted 654 submissions
- 34 coders exceeded state of the art by 10^2 - 10^5
- 89 different approaches to solve problem identified
- Winners from Russia, France, Egypt, Belgium & US
- Annotate 10 million sequences in < 3 mins; Quarter billion sequences in ~ 1 hour on laptop



Olga Kokshagina 2015

Starting kit

- Starting kit ==all that we provide on top of the data set
- Difficult to get right : between PR and technical documentation
- Web pages (see current Kaggle challenges), videos…
- Document with HEP tracking for the dummies … guiding people to more complex algorithms, without scaring them
- Software which allows to get a very first solution in <1 hour, addressing different communities:
  - Jupyter notebook
  - Simple python
  - SciKit-learn
  - Theano
  - ACTS nicely packaged
  - Etc….