TrackML: Tracking Machine Learning challenge

David Rousseau (LAL-Orsay, U Paris-Saclay) (rousseau@lal.in2p3.fr), with Paolo Calafiura, Steven Farrell, Heather Gray (LBNL-Berkeley), Jean-Roch Vlimant (CalTech), Yetkin Yilnaz (LAL), Cécile Germain (LAL/LRI), Isabelle Guyon (ChaLearn, U Paris Saclay), Vincenzo Innocente, Andreas Salzburger (CERN), Tobias Golling, Moritz Kiehn, Sabrina Amrouche (U Geneva), Vava Gligorov (LPNHE-Paris), Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex)

Special thanks for the preparation of the slides: Andreas Salzburger, Jean-Roch Vlimant

LRI-Orsay seminar, 13th Mar 2018
Outline

- Particle Physics context
- Why a Tracking challenge now?
- HiggsML challenge recap
- Simulation
- Metric
- Conclusion

David Rousseau, LRI-Orsay Seminar, 13th March 2018
Who are we?

Paolo Calafiura, Steven Farrell, Heather Gray (LBNL-Berkeley), Jean-Roch Vlimant (CalTech), Cécile Germain (LAL/LRI U Paris Saclay), Isabelle Guyon (ChaLearn, U Paris Saclay), David Rousseau, Yetkin Yilnaz (LAL Orsay U Paris Saclay), Vincenzo Innocente, Andreas Salzburger (CERN), Tobias Golling, Moritz Kiehn, Sabrina Amrouche (U Geneva), Vava Gligorov (LPNHE-Paris), Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex)

- Particle physics tracking experts from three large CERN experiments on the LHC ATLAS, CMS and LHCb
- Machine Learning scientists
- Some of us have organised challenges on Kaggle
  - The Higgs Machine Learning challenge 2014 (proceedings of NIPS 2014 workshop)
  - Flavour of Physics challenge 2015
- We have been preparing this new challenge for 3 years...
Partners

kaggle

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Center for Data Science

CERN openlab
LHC purpose in a nutshell
Proton collisions

E=mc²

Create heavy short-lived Particles (e.g. Higgs boson)

Most decay immediately

A few types (pions, electron, muons…) live long enough to be detected

See Kyle Cranmer keynote NIPS 2016
A proton collision in ATLAS detector

David Rousseau, TrackML challenge, CiML NIPS 2017
Current situation: 20 parasitic collisions
High Lumi-LHC : 200 parasitic collisions
Future of LHC beyond Higgs boson discovery
Vaccuum stability depends of the exact top quark and Higgs boson mass.

We’re leaving at the edge!

This is suspicious!
→ hint for « new physics »
The two infinite (Pascal, Newton,...)

Luminous matter
Scale $\sim 10^{22}$ m

Dark matter

Luminosity by 10 in 2025

Gravitational lensing

How?

$\Rightarrow$ HL-LHC, increase LHC

David Rousseau, TrackML challenge, CiML NIPS 2017
Particle Tracking at LHC
LHC tracking...
...fascinates ML scientists
Current situation

- High luminosity means high pileup
- Combinatorics of charged particle tracking become extremely challenging for GPDs
- Generally sub-linear scaling for track reconstruction time with $m$

Impressive improvements for Run 2, but we need to go much further

Point precision $\sim 5 \, \mu m$ to 3mm

100k points 10k tracks / event

10-100 billion events/year
• Tracking (in particular pattern recognition) dominates reconstruction CPU time at LHC
• High Luminosity-LHC perspective: increased rate of parasitic collisions from 40 (2017) to 200
• CPU time of current software quadratic/exponential extrapolation (difficult to quote any number)
• (current software give sufficiently good results in terms of accuracy, but x10 too slow)
• Distant future FCC-hh would reach 1000
Motivation

- LHC experiments future computing budget flat (at best) (LHC experiments use 300,000 CPU cores on the LHC world wide computing grid)
- Installed CPU power per $==€==CHF expected increase factor <10 in 2025
- Experiments plan on increase of amount of data recorded (by a factor ~10)
  - HighLumi reconstruction to be as fast as current reconstruction despite factor 10 in complexity
  - Requires very significant software CPU improvement, factor ~10
- Large effort to optimise current software and tackle micro and macro parallelism
  - Also development of dedicated hardware for fast tracking
- >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
- Need to engage a wide community to tackle this problem
Particle Tracking algorithms
Current Algorithms

- Pattern: connect 3D points into tracks
- Essentially combinatorial approach
- Tracks are (not perfect) helices pointing (approximately) to the origin
- Challenge: explore completely new approaches
- (not part of the challenge: given the points, estimate the track parameters)
Hough transform: principle

- Toy: 2D, track coming from origin → 2 parameters phi, rho0 (radius of curvature)
- Find an excess in image plane
- → go back to real plane
Hough Transform: toy 1

- 6 particles, no hit smearing
Hough Transform: toy 2

- 6 particles, with hit smearing
Hough Transform: final comments

- Mapping x,y,z to 5 helix parameters
- Generalised Hough Transform
- Excess to be found in 5D image space
- Difficult to take into account point measurement anisotropy
- Multiple scattering broadens the possible trajectory
- Excess in image space is blurred
- High multiplicity → confusion
- However: linear time at first order
- Approach still promising
Kalman filter

- initially developed by I. Kalman to track missiles (for HEP pioneered by Billoir and R. Fruehwirth)
- performs a progressive way of least square estimation equivalent to a $\chi^2$ fit (if run with a smoother)
- start with transport of track parameters (and covariances) to measurement surface, create predicted parameters ("predicted state")
- combine/update predicted parameters with measurement to updated parameters ("filtered state")
- Also used for local pattern recognition (outlier)
- Computation intensive
Pattern recognition in ML

- Pattern recognition, tracking, is a very old, very hot topic in Artificial Intelligence: examples

- Note that these are real-time applications, with CPU constraints
- Worry about efficiency, "track swap"...
- But no on-the-shelf algorithm will solve our problem
- (in fact a few lines calling DBScan in sklearn does find some tracks)

David Rousseau, LRI-Orsay
An early attempt

known

- Losely inspired from Traveling Salesman Problem with NN by Hopfield & Tank Biological Cybernetics 52 (1985) 141. or with Minimal Tree Span Cassel & Kowalski Nucl Inst; and Meth 185 (1981) 235

- (large litterature since, e.g. Neural Combinatorial Optimization with reinforcement learning, Bello et al Google Brain 1611.0994)


- However never deployed
A recent attempt: NOVA

Neutrino interaction classification using Convolutional Neural Network
No attempt to separate individual tracks.

Used in published results
No attempt to identify separate tracks.
A simplified tracking challenge setup on RAMP (Center for Data Science Paris-Saclay platform, Balazs Kégl)

A (non completely trivial) 2D simulation with ~10 tracks instead of 3D/10,000 tracks

Run as a 40 hours hackathon during CTDWIT 6-9th March 2017 LAL-Orsay

Allowed to validate robustness a scoring variable and show richness of possible algorithms: combinatorial (HEP baseline), conformal mapping, MCTS, LSTM (See also S. Farrell et al paper accepted by NIPS 2017 “Deep Learning for Physical Science”)

Published in proceedings EPJ Web Conf., 150 (2017) 00015
Attention mechanisms [16–18], giving the models the capability to focus on particular parts of the input or intermediate feature representations to produce a desired output.

Such rich learned representations have also proven highly beneficial for tracking-based problems in non-HEP applications, such as sports analytics and computational neuro-science. For instance, [19] uses attention-based LSTMs to learn hierarchical models of basketball player behavior from tracking data, while [20] applies recurrent neural networks to generate realistic fruit-fly behavior and handwriting.

4 Datasets

Simple toy datasets were used to study and demonstrate the ideas discussed in this paper. The “detectors” are made of perfect pixel planes in 2D or 3D. Tracks are sampled from straight lines contained within the detector volume, and binary hits are recorded in each intercepting discrete pixel on each layer. No trajectory curvature, material effects, or detector inefficiencies are modeled. These toy datasets are highly simplistic compared to real tracking detector data, which means that quantitative results are likely not indicative of algorithm performance in realistic scenarios. Nonetheless, this simple toy data provides a useful environment to test out various models. Figure 2 shows example 2D data generated with tracks as well as uniform noise. Figure 3 shows an example 3D event.

For the experiments described in section 5, the following data configurations were used. 2D toy experiments used one million 2D events with 50 detector layers of 50 pixels each, one signal track, and five background tracks for training. The 3D toy experiments used a detector with 10 layers and $50 \times 50$ pixels in each layer. Events were generated with a random number of background tracks sampled from a Poisson distribution with mean values varied from 1 to 100. At each point, five million events were generated for training and one hundred thousand events for testing.

5 Track finding with LSTMs and CNNs

The goal of this line of study is to identify models which can do the assignment of pixel hits to a track candidate by extrapolating from a partial track (a seed) through detector layers. When considering a single track at a time, the problem can be formulated as one of multi-class classification. The pixels in one detector layer make up the possible “classes”, and the model must identify which one is traversed by the target track candidate. Modeling of track dynamics can be handled by LSTMs or CNNs.

A basic LSTM model for 2D track finding is shown in figure 4. This model consists of an LSTM layer which reads the input pixel arrays and a single fully-connected layer which is applied separately to each LSTM output to produce the pixel predictions for the same detector layer. The seed is specified.

See:
See:
2014 HiggsML challenge recap
HiggsML in a nutshell

- (see JMLR proceedings http://proceedings.mlr.press/v42/cowa14.html)
- ATLAS Htautau MC analysis ntuple released
- Competition on kaggle to optimise Higgs selection: https://higgsml.lal.in2p3.fr
- 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 (until spring 2015)
  - 35772 solutions uploaded
- 136 forum topics with 1100 posts
What data did we release?

- From ATLAS full sim Geant4 MC12 production
- 30 variables
- Signal is $H\rightarrow\tau\tau$, Background a mixture of: Z, top, W
- Based on November 2013 ATLAS $H\tau\tau$ conf note ATLAS-CONF-2013-108
- Preselection for lep-had topology: single lepton trigger, one lepton identified, one hadronic tau identified
- $\approx 800,000$ events (all that was available):
  - 250,000 training data set
  - 550,000 test data set without label and weight
- Reproduces reasonably well (~20%) content of 3 highest sensitivity bins (x 2 categories) in conf note
- (some background and many correction factors deliberately omitted so that the sample cannot be used for physics, only for machine learning studies)
Permanently available and usable by anyone (also non ATLAS) on CERN Open Data:
ASCII csv file, with mixture of Higgs to tautau (lephad) signal and corresponding backgrounds, from official GEANT4 ATLAS simulation
Weight and signal/background (for training dataset only)
weight (fully normalised)
label : « s » or « b »
Conf note variables used for categorization or BDT:
   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
   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

Primitive 3-vectors allowing to compute the conf note variables (mass neglected),
16 independent variables:
Real life vs challenge

1. Systematics (and data vs MC)
2. 2 categories x n BDT score bins
3. Background estimated from data (embedded, anti tau, control region) and some MC
4. Weights include all corrections. Some negative weights (tt)
5. Potentially use any information from all 2012 data and MC events
6. Few variables fed in two BDT
7. Significance from complete fit with NP etc...
8. 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!

David Rousseau, LRI-Orsay Seminar, 13th March 2018
# Final leaderboard

<table>
<thead>
<tr>
<th>#</th>
<th>Δrank</th>
<th>Team Name</th>
<th>Score</th>
<th>Entries</th>
<th>Last Submission UTC (Best – Last Submission)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Gábor Melis ‡ * 7000$ « deep » learning</td>
<td>3.80581</td>
<td>110</td>
<td>Sun, 14 Sep 2014 09:10:04 (-0h)</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Tim Salimans ‡ * 4000$ BDT ensemble</td>
<td>3.78913</td>
<td>57</td>
<td>Mon, 15 Sep 2014 23:49:02 (-40.6d)</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>nhlxShaze ‡ * 2000$</td>
<td>3.78682</td>
<td>254</td>
<td>Mon, 15 Sep 2014 16:50:01 (-76.3d)</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>ChoKo Team</td>
<td>3.77526</td>
<td>216</td>
<td>Mon, 15 Sep 2014 15:21:36 (-42.1h)</td>
</tr>
<tr>
<td>5</td>
<td>35</td>
<td>cheng chen</td>
<td>3.77384</td>
<td>21</td>
<td>Mon, 15 Sep 2014 23:29:29 (-0h)</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>quantify</td>
<td>3.77086</td>
<td>8</td>
<td>Mon, 15 Sep 2014 16:12:48 (-7.3h)</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>Stanislav Semenov &amp; Co (HSE Yandex)</td>
<td>3.76211</td>
<td>68</td>
<td>Mon, 15 Sep 2014 20:19:03</td>
</tr>
<tr>
<td>8</td>
<td>17</td>
<td>Luboš Motl's team ‡ Best physicist</td>
<td>3.76050</td>
<td>589</td>
<td>Mon, 15 Sep 2014 08:38:49 (-1.6h)</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>Roberto-UCIIM</td>
<td>3.75864</td>
<td>292</td>
<td>Mon, 15 Sep 2014 23:44:42 (-44d)</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>Davut &amp; Josef</td>
<td>3.75838</td>
<td>161</td>
<td>Mon, 15 Sep 2014 23:24:32 (-4.5d)</td>
</tr>
<tr>
<td>45</td>
<td>15</td>
<td>crowwork ‡ HEP meets ML award XGBoost authors</td>
<td>3.71885</td>
<td>94</td>
<td>Mon, 15 Sep 2014 23:45:00 (-5.1d)</td>
</tr>
<tr>
<td>782</td>
<td>149</td>
<td>Eckhard TMVA expert, with TMVA improvements</td>
<td>3.49945</td>
<td>29</td>
<td>Mon, 15 Sep 2014 07:26:13 (-46.1h)</td>
</tr>
<tr>
<td>991</td>
<td>14</td>
<td>Rem.</td>
<td>3.20423</td>
<td>2</td>
<td>Mon, 16 Jun 2014 21:53:43 (-30.4h)</td>
</tr>
</tbody>
</table>

- **Simple TMVA boosted trees**: 3.19956
From domain to challenge and back

Domain e.g. HEP

Problem

Domain experts solve the domain problem

Solution

Challenge organisation

simplify

reimport

Challenge

Problem

The crowd solves the challenge problem

Solution

David Rousseau, LRI-Orsay Seminar, 13th March 2018
The tracking challenge
In a nutshell

- Accurate simulation engine (ACTS https://gitlab.cern.ch/acts/acts-core) to produce realistic events
  - 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 100k events, 10 billion tracks ~100GByte

- 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
  - (do not consider estimation of particle parameter)
  - Score: fraction of points correctly grouped together
  - Evaluation on test sample with per-mille precision on 100 event
(note the measurement anisotropy)
Detector: layout

Long strips

Short strips

Pixels

~12 points per tracks

David Rousseau, LRI-Ursay Seminar, 13th March 2018
Clustering: analog in Pixel, digital in Strips

- very different residuals (see examples)
- we’ll let participants figure out given \((x,y,z)_{\text{measured}} \Leftrightarrow (x,y,z)_{\text{true}}\)

Non trivial simplification: one true track \(\Leftrightarrow\) one reco hit (except for 1% inefficiency)

=> no hit merging/splitting
Some details on simulation

- Particles bent by quasi-solenoidal magnetic field ➔ quasi-helicoidal trajectories
- Deterministic trajectory except for multiple scattering

Exaggerated Multiple Scattering

Same particle, simulated 100 times
Event simulation

- Typical LHC event simulated
  - Pythia tt-bar event
  - Overlaid with Poisson(200) Pythia minimum bias
  - ~10'000 tracks

- Most tracks are coming from a central region: gaussian $\sigma_z=5.5$ cm, transverse $\sigma=15\mu$m, some from a larger cylinder

- 15% of random hits

- Trajectories are deterministic, except for Multiple Scattering, Energy Loss and hadronic interaction
# Datasets

## Hit file

(measured position mm)                  (pixel location and charge)

<table>
<thead>
<tr>
<th>hit_id</th>
<th>volume_id</th>
<th>layer_id</th>
<th>module_id</th>
<th>x</th>
<th>y</th>
<th>z</th>
<th>ncells</th>
<th>pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>-63.9659</td>
<td>-3.70513</td>
<td>-1502.5</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>-40.2738</td>
<td>2.82386</td>
<td>-1502.5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>-88.1049</td>
<td>-11.72380</td>
<td>-1502.5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>-39.7041</td>
<td>-8.71702</td>
<td>-1502.5</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>-30.4918</td>
<td>-8.19262</td>
<td>-1502.5</td>
<td>1</td>
</tr>
</tbody>
</table>

## Truth file

(true position mm      particle momentum GeV)

<table>
<thead>
<tr>
<th>hit_id</th>
<th>particle_id</th>
<th>tx</th>
<th>ty</th>
<th>tz</th>
<th>tpx</th>
<th>tpy</th>
<th>tpz</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>-63.972698</td>
<td>-3.72889</td>
<td>-1502.5</td>
<td>-0.342366</td>
<td>-0.001899</td>
<td>-7.83544</td>
<td>0.018565</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>-40.287201</td>
<td>2.84328</td>
<td>-1502.5</td>
<td>-0.366049</td>
<td>0.013878</td>
<td>-13.55470</td>
<td>0.035088</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>-88.089600</td>
<td>-11.72360</td>
<td>-1502.5</td>
<td>-0.550128</td>
<td>-0.041929</td>
<td>-9.22279</td>
<td>0.018542</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>-39.712601</td>
<td>-8.71581</td>
<td>-1502.5</td>
<td>-0.363936</td>
<td>-0.094646</td>
<td>-14.01150</td>
<td>0.035088</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>-30.470400</td>
<td>-8.18647</td>
<td>-1502.5</td>
<td>-0.413489</td>
<td>-0.123403</td>
<td>-20.65790</td>
<td>0.000000</td>
</tr>
</tbody>
</table>
Datasets

- **Particle file**
  - origin vertex (mm)
  - momentum (GeV)
  - charge

<table>
<thead>
<tr>
<th>particle_id</th>
<th>vx</th>
<th>vy</th>
<th>vz</th>
<th>px</th>
<th>py</th>
<th>pz</th>
<th>q</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.021389</td>
<td>-0.012618</td>
<td>-0.624757</td>
<td>38.907001</td>
<td>-16.146099</td>
<td>-84.311096</td>
<td>-1</td>
</tr>
<tr>
<td>1</td>
<td>-0.021389</td>
<td>-0.012618</td>
<td>-0.624757</td>
<td>-0.661993</td>
<td>0.118267</td>
<td>249.181000</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>-0.021389</td>
<td>-0.012618</td>
<td>-0.624757</td>
<td>0.821614</td>
<td>0.954217</td>
<td>0.948994</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>-0.021389</td>
<td>-0.012618</td>
<td>-0.624757</td>
<td>0.300791</td>
<td>0.080450</td>
<td>2.656530</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>-0.021389</td>
<td>-0.012618</td>
<td>-0.624757</td>
<td>-0.552250</td>
<td>-0.481988</td>
<td>-0.888733</td>
<td>1</td>
</tr>
</tbody>
</table>

(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

<table>
<thead>
<tr>
<th>volume_id</th>
<th>layer_id</th>
<th>module_id</th>
<th>cx</th>
<th>cy</th>
<th>cz</th>
<th>rot_xu</th>
<th>rot_xv</th>
<th>rot_xw</th>
<th>rc</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>-65.7965</td>
<td>-5.17830</td>
<td>-1502.5</td>
<td>0.078459</td>
<td>-0.996917</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>-139.8510</td>
<td>-6.46568</td>
<td>-1502.0</td>
<td>0.046183</td>
<td>-0.998933</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>-138.6570</td>
<td>-19.34190</td>
<td>-1498.0</td>
<td>0.138156</td>
<td>-0.990410</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>2</td>
<td>4</td>
<td>-64.1764</td>
<td>-15.40740</td>
<td>-1498.0</td>
<td>0.233445</td>
<td>-0.972370</td>
<td>0.0</td>
</tr>
</tbody>
</table>
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
**good track**

many compatible hits

completeness

uniqueness

low \( \chi^2/\text{ndf} \)

small impact parameter (for primaries)

clusters are compatible

**not so good track**

short tracks

holes

shared hits

bad fit quality, outliers
Define: \( \text{weight} = \text{weight}_{\text{order}} \times \text{weight}_{\text{pt}} \)

Weighted track score

- \( \text{Weight}_{\text{order}} \): more emphasis on first and last hits
- \( \text{Weight}_{\text{pt}} \): more emphasis on high pT tracks
- \( \text{Weight} = 0 \) for noise hits or hits from particle with \( \leq 3 \) hits

David Rousseau, LRI-Orsay Seminar, 13th March 2018
Track scoring

- Overall scoring defined at hit level
- Loop on reco tracks
  - Require >50% of hits from same true particle
  - Require >50% of hits from this true particle in this reco track
  - At this point 1 ⇔ 1 relationship between true and reco tracks
  - Sum the weights of the intersection (hits belonging both to true and reco track)
- Event score normalised to the sum of weights of all the hits
  - ⇒ ideal algorithm has score==1.
- Final score averaged of 100 events⇒ statistical precision
  ~0.1%

David Rousseau, LRI-Orsay Seminar, 13th March 2018
Attempt with 2 simple algs

DBScan (sk-learn clustering)

Multiplicity

Hough Transform
## Real life vs challenge

<table>
<thead>
<tr>
<th>Real life</th>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Wide type of physics events</td>
<td>1. One event type (ttbar)</td>
</tr>
<tr>
<td>2. Full detailed Geant 4 / data</td>
<td>2. ACTS (MS, energy loss, hadronic interaction, solenoidal magnetic field, inefficiency)</td>
</tr>
<tr>
<td>3. Detailed dead matter description</td>
<td>3. Cylinders and slabs</td>
</tr>
<tr>
<td>4. Complex geometry (tilted modules, double layers, misalignments...)</td>
<td>4. Simple, ideal, geometry (cylinders and disks)</td>
</tr>
<tr>
<td>5. Hit merging</td>
<td>5. No hit merging</td>
</tr>
<tr>
<td>6. Allow shared hits</td>
<td>6. Disallow shared hits</td>
</tr>
<tr>
<td>7. Output is hit clustering, track parameter and covariance matrix</td>
<td>7. Output is hit clustering</td>
</tr>
<tr>
<td>8. Multiple metrics (see TDR’s)</td>
<td>8. Single number metrics</td>
</tr>
</tbody>
</table>

Simpler, but not too simple!

David Rousseau, LRI-Orsay Seminar, 13th March 2018
We have decided to run in two phases

- **Accuracy Phase**: focus only on accuracy, no CPU incentive
  - Goal is to expose innovative algorithms
  - Training time unlimited
  - Evaluation time unlimited
  - To run on Kaggle May-August 2018

- **Throughput Phase**: focus on CPU, preserving accuracy
  - Goal is to expose the fastest algorithms
  - Training time (still) unlimited
  - Require the challenge platform to run the algorithm evaluation within fully reproducible controlled environment (VM with x86 processor with 2GB memory, but do not exclude a GPU track in addition)
  - To run in July-October 2018

**Prizes**:
- From leaderboards of both phases
- From jury examining the algorithms: what are the more likely to be beneficial to HEP?
  - Invitation to NIPS workshop (if confirmed) and to CERN workshop
Events

- **Challenge Schedules**
  - May to August: Run challenge Accuracy phase
  - July to October: Run challenge Throughput phase

- **Conference/workshops**
  - Connecting The Dots 20-22nd March 2018 Seattle hackathon
  - July 2018: Accuracy Phase accepted as an official competition for the IEEE World Congress on Computational Intelligence at Rio de Janeiro
  - July 2018: (submitted) as a talk at CHEP Sofia and ICHEP Seoul
  - December 2018: Throughput Phase as a NIPS 2018 competition and possibly workshop
  - Spring 2019: grand finale workshop at CERN with prize delivery
Conclusion

- Setting up TrackML: a particle tracking challenge
- Goal is to involve ML community in overhauling core algorithms of CERN LHC experiments.
  - Looking for new approaches rather than hyper-optimised (HEP) approaches
- Very large training dataset ~100GB
  - Will be released (CERN Open Data portal most likely) after the challenge
- Wealth of possible ML techniques (NN, CNN, RNN, Reinforcement learning, clustering techniques, MCTS...) ... which makes it all the more interesting
- Separate Accuracy phase (most accurate algorithm) and Throughput phase (fastest algorithm to reach similar accuracy)
- Sponsorship more or less OK for Accuracy Phase, still looking for ~40k€ for Throughput phase
- Contact: trackml.contact@gmail.com
- More details, news, etc...: https://sites.google.com/site/trackmlparticle/, twitter @trackmllhc
- We’ve beeing accepted as a NIPS 2018 competition (Throughput phase)