TrackML: the roller coaster of organizing a HEP challenge on Kaggle and Codalab -

David Rousseau
LAL-Orsay
rousseau@lal.in2p3.fr @dhpmrou
EPS-HEP Conference 10-17 Jul 2019, Gand, BElgique
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 Massachusetts), Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex, HSE)

5-6 FTE year

Platforms:

- Kaggle
- CodaLab

Phase 1:
- Accuracy

Phase 2:
- Throughput

TrackML, David Rousseau, EPS-HEP outreach, Jul 2019, Gand
 Tracking crisis

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

\[ \rightarrow \text{challenge} \quad \text{TrackingML} \quad !! \]

Similar plots from CMS
• 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.

23 m

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

100k points 10k tracks / event

10-100 billion events/year

6 m

2 m

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

100k points 10k tracks / event

10-100 billion events/year

6 m

2 m
Tracking outside HEP

- ...is very different
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 problem
  - Strong CPU incentive
From domain to challenge and back

**Domain e.g. HEP**

- **Problem**: Domain experts solve the domain problem
- **Solution**:
  - ~years
  - simplify
  - ~years
  - reimport

**Challenge organisation**

- **Challenge**: The crowd solves the challenge problem
- **Problem**: ~months

TrackML, David Rousseau, EPS-HEP outreach, Jul 2019, Gand
TrackML timeline

- Kick off
- Slow maturation of the problem
- Building of the team
- Decide to Focus on pattern recognition
- 2D Hackathon Orsay HSF workshop
- ACTS simulation
- Dataset definition
- Seattle challenge CTD
- Kaggle implementation, documentation etc...
- Kaggle Accuracy challenge
- Codalab implementation
- Codalab Throughput challenge
- CERN Grande Final
- Spin-offs
- Mar 2015
- Mar 2017
- Mar 2018
- May 2018
- Aug 2018
- Oct 2018
- Mar 2019
- Jul 2019

TrackML, David Rousseau, EPS-HEP outreach, Jul 2019, Gand
3D points
Dataset

3D points ➔ tracks
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

<table>
<thead>
<tr>
<th><strong>Real Life</strong></th>
<th><strong>Challenge</strong></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!**

TrackML, David Rousseau, EPS-HEP outreach, Jul 2019, Gand
Evolution of leaderboard
<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>Score</th>
<th>Rank</th>
<th>Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Top Quarks</td>
<td>0.92182</td>
<td>10</td>
<td>19d</td>
</tr>
<tr>
<td>2</td>
<td>outrunner</td>
<td>0.90302</td>
<td>9</td>
<td>18d</td>
</tr>
<tr>
<td>3 (HEP)</td>
<td>Sergey Gorbunov</td>
<td>0.89353</td>
<td>6</td>
<td>18d</td>
</tr>
<tr>
<td>4 (HEP)</td>
<td>demelian</td>
<td>0.87079</td>
<td>35</td>
<td>1mo</td>
</tr>
<tr>
<td>5</td>
<td>Edwin Steiner</td>
<td>0.86395</td>
<td>5</td>
<td>18d</td>
</tr>
<tr>
<td>6</td>
<td>Komaki</td>
<td>0.83127</td>
<td>22</td>
<td>18d</td>
</tr>
<tr>
<td>7</td>
<td>Yuval &amp; Trian</td>
<td>0.80414</td>
<td>56</td>
<td>18d</td>
</tr>
<tr>
<td>8</td>
<td>bestfitting</td>
<td>0.80341</td>
<td>6</td>
<td>18d</td>
</tr>
<tr>
<td>9</td>
<td>DBSCAN forever</td>
<td>0.80114</td>
<td>23</td>
<td>18d</td>
</tr>
<tr>
<td>10</td>
<td>Zidmie &amp; KhaVo</td>
<td>0.76320</td>
<td>26</td>
<td>18d</td>
</tr>
<tr>
<td>11</td>
<td>Andrea Lonza</td>
<td>0.75845</td>
<td>15</td>
<td>18d</td>
</tr>
<tr>
<td>12</td>
<td>Finnies</td>
<td>0.74827</td>
<td>56</td>
<td>18d</td>
</tr>
<tr>
<td>13</td>
<td>Rei Matsuzaki</td>
<td>0.74035</td>
<td>12</td>
<td>18d</td>
</tr>
<tr>
<td>14</td>
<td>Mickey</td>
<td>0.73217</td>
<td>10</td>
<td>2mo</td>
</tr>
<tr>
<td>15</td>
<td>Vicens Gaitan</td>
<td>0.70429</td>
<td>19</td>
<td>1mo</td>
</tr>
<tr>
<td>16</td>
<td>Robert</td>
<td>0.69955</td>
<td>3</td>
<td>21d</td>
</tr>
<tr>
<td>17</td>
<td>Yuval-CPMP tribute band</td>
<td>0.69364</td>
<td>20</td>
<td>20d</td>
</tr>
<tr>
<td>18</td>
<td>N. Hi. Bouzu</td>
<td>0.67573</td>
<td>9</td>
<td>22d</td>
</tr>
<tr>
<td>19</td>
<td>Steins;Gate</td>
<td>0.66763</td>
<td>12</td>
<td>19d</td>
</tr>
<tr>
<td>20</td>
<td>Victor Nedel'ko</td>
<td>0.66723</td>
<td>4</td>
<td>2mo</td>
</tr>
</tbody>
</table>
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
e.g. Participant Data Analysis

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

See link
Efficiency all

TrackML, David Rousseau, EPS-HEP outreach, Jul 2019, Gand
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 ourselves 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...
- \(\Rightarrow\) decision: remeasure speed at the end of the competition many times on a dedicated machine
  - \(\Rightarrow\) 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 »)
## Throughput phase LB

<table>
<thead>
<tr>
<th>#</th>
<th>User</th>
<th>Entries</th>
<th>Date of Last Entry</th>
<th>score ▲</th>
<th>accuracy_mean ▲</th>
<th>accuracy_std ▲</th>
<th>computation time (sec) ▲</th>
<th>computation speed (Sec/event) ▲</th>
<th>Duration ▲</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sgorbuno</td>
<td>9</td>
<td>03/12/19</td>
<td>1.1727 (1) 1.16</td>
<td>0.944 (2)</td>
<td>0.00 (14)</td>
<td>28.06 (1)</td>
<td>0.56 (1)</td>
<td>0.60</td>
</tr>
<tr>
<td>2</td>
<td>fastrack</td>
<td>53</td>
<td>03/12/19</td>
<td>1.1145 (2) 1.12</td>
<td>0.944 (1)</td>
<td>0.00 (15)</td>
<td>55.51 (16)</td>
<td>1.11 (16)</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>cloudkitchen</td>
<td>73</td>
<td>03/12/19</td>
<td>0.9007 (3) 0.891</td>
<td>0.927 (3)</td>
<td>0.00 (13)</td>
<td>364.00 (18)</td>
<td>7.28 (18)</td>
<td>7.41</td>
</tr>
<tr>
<td>4</td>
<td>cubus</td>
<td>8</td>
<td>09/13/18</td>
<td>0.7719 (4) 0.770</td>
<td>0.895 (4)</td>
<td>0.01 (9)</td>
<td>675.35 (19)</td>
<td>13.51 (19)</td>
<td>13.7</td>
</tr>
<tr>
<td>5</td>
<td>Taka</td>
<td>11</td>
<td>01/13/19</td>
<td>0.5930 (5)</td>
<td>0.875 (5)</td>
<td>0.01 (12)</td>
<td>2668.50 (23)</td>
<td>53.37 (23)</td>
<td>2758.00 (13)</td>
</tr>
<tr>
<td>6</td>
<td>Vicennial</td>
<td>27</td>
<td>02/24/19</td>
<td>0.5634 (6)</td>
<td>0.815 (6)</td>
<td>0.01 (10)</td>
<td>1270.73 (20)</td>
<td>25.41 (20)</td>
<td>1339.00 (10)</td>
</tr>
<tr>
<td>7</td>
<td>Sharad</td>
<td>57</td>
<td>03/10/19</td>
<td>0.2918 (7)</td>
<td>0.674 (7)</td>
<td>0.02 (4)</td>
<td>1902.20 (22)</td>
<td>38.04 (22)</td>
<td>1986.00 (12)</td>
</tr>
<tr>
<td>8</td>
<td>WeizmannAI</td>
<td>5</td>
<td>03/12/19</td>
<td>0.0000 (8)</td>
<td>0.133 (11)</td>
<td>0.01 (11)</td>
<td>88.08 (17)</td>
<td>1.76 (17)</td>
<td>124.00 (7)</td>
</tr>
<tr>
<td>9</td>
<td>harshakoundinya</td>
<td>2</td>
<td>03/12/19</td>
<td>0.0000 (8)</td>
<td>0.085 (13)</td>
<td>0.01 (6)</td>
<td>49.22 (8)</td>
<td>0.98 (8)</td>
<td>86.00 (3)</td>
</tr>
<tr>
<td>10</td>
<td>iWit</td>
<td>6</td>
<td>03/10/19</td>
<td>0.0000 (8)</td>
<td>0.082 (15)</td>
<td>0.01 (8)</td>
<td>48.23 (3)</td>
<td>0.96 (3)</td>
<td>85.00 (2)</td>
</tr>
</tbody>
</table>

TrackML, David Rousseau, EPS-HEP outreach, Jul 2019, Gand
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
Visualisation spin-off

- Visit at CERN Tobias Isenberg visualisation scientist at LRI-Orsay with PhD student Xiyao Wang
- They are using TrackML dataset to experiment with visualisation/interaction with Microsoft’ Hololens (see talk in CERN Grand Finale 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....

TrackML, David Rousseau, EPS-HEP outreach, Jul 2019, Gand
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
- 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/