Introduction
RAPID 2018 workshop

Workshop at TU Dortmund, 19th - 21st of November 2018
• Challenge: Fast primary vertex reconstruction
• Bringing together experts from HEP and industry
Introduction
What is a RAMP challenge?

Building Egypt's Great Pyramid
This 4,500-year-old system used to pull alabaster stones up a steep slope was discovered at Hatnub, an ancient quarry in the Eastern Desert of Egypt.

[source]

Yet another data challenge (framework)? Why? Who ordered that?
Two different approaches in data science to tackle problems:

**Hackathons**
- Encourage cooperation
- Not well-defined
- Open-ended

**Challenges**
- Well-defined problem and goal
- Competitive
- Might discourage code sharing

→ RAMP aims to bring these concepts together
Why RAMP?
Combining competition...

Start with 'closed' competitive phase ...

drug_spectra_mines_201617

Leaderboard

Combined score: 0.023

<table>
<thead>
<tr>
<th>team</th>
<th>submission</th>
<th>contributivity</th>
<th>historical contributivity</th>
<th>combined</th>
<th>err</th>
</tr>
</thead>
<tbody>
<tr>
<td>clement.acher</td>
<td>reg_nn2</td>
<td>0</td>
<td>0</td>
<td>35.214</td>
<td>0.023</td>
</tr>
<tr>
<td>clement.acher</td>
<td>reg_nn</td>
<td>0</td>
<td>0</td>
<td>12.949</td>
<td>0.027</td>
</tr>
<tr>
<td>mohamed.trabelsi</td>
<td>tack15</td>
<td>0</td>
<td>0</td>
<td>1.179</td>
<td>0.021</td>
</tr>
<tr>
<td>mohamed.trabelsi</td>
<td>tack16</td>
<td>0</td>
<td>0</td>
<td>1.042</td>
<td>0.020</td>
</tr>
<tr>
<td>jeang</td>
<td>submission24-2</td>
<td>0</td>
<td>0</td>
<td>0.899</td>
<td>0.023</td>
</tr>
<tr>
<td>jeang</td>
<td>submission24</td>
<td>0</td>
<td>0</td>
<td>0.899</td>
<td>0.023</td>
</tr>
<tr>
<td>samuel.kron</td>
<td>ETLACAMARGHE3?</td>
<td>0</td>
<td>0</td>
<td>0.558</td>
<td>0.612</td>
</tr>
</tbody>
</table>

More information at RAMP studio
Why RAMP?
... with collaboration... followed by 'open' collaborative phase

Good ideas propagate, Participants learn from each other and co-create.

Links correspond to solutions that were re-used.

More information at RAMP studio
What is the problem we want to solve using RAMP?
The problem to solve

Primary vertex (PV) reconstruction (at LHCb):

- Why reconstructing primary vertices is important:
  - Discriminating short- and long-lived particles
  - Trigger on displaced signatures
The problem to solve

Primary vertex reconstruction (at LHCb):

 Upgrade I conditions:

• On average about 6 'visible' primary vertices per event
• PVs are distributed in z with a spread of about 6cm
• Spread in x,y about 20µm
• $\mathcal{O}(200)$ tracks in Vertex Locator
The problem to solve

Primary vertex reconstruction (at LHCb):

• How to reconstruct close PVs?
• How to assign tracks to close PVs?
Setting the stage
Input data:

- LHCb Upgrade I simulation
- Reconstructed track states
  - extrapolated to point of closest approach to beamline
  - covariance matrix from simplified Kalman filter
- Hit positions in Vertex Locator
- Underlying truth: MC primary vertex positions
Setting the stage

Baseline solution

LHCb PV reconstruction performance in Run2:

The algorithm consists of two stages:

- Seeding in $z$
- Fitting to get PV positions and covariance matrices

The baseline solution is basically a worse, untuned version of this:

- Efficiency $\sim 90\%$
- Fake rate $\sim 10\%$
Setting the stage

Scoring

Important physics performance parameters:

- PV reconstruction efficiency $\epsilon$
- Fake (or false positive) rate $f$
- PV position resolutions $\sigma_{x,y,z}$

→ not obvious what total score should be

orange and red better than blue
But is orange $>$ red?
Setting the stage

Scoring

Rather arbitrarily define:

- total score $\sim \epsilon \cdot (1 - f)^2 \cdot \frac{1}{\sigma_x \sigma_y \sigma_z}$

$\rightarrow$ same score can describe vastly different physics performances
(further complicated by different $\sigma_{x,y,z}$)
Analyzing the submissions
Submissions

≈ 40 submissions by 21 people over the course of the workshop

(No strict separation between competitive and collaborative phases)
Submissions

Comparison - Total score

Compare a few submissions:

What causes small scores?
Submissions approaching the baseline efficiency
• Fake rates usually problematic
Submissions

Comparison - Effective efficiency

Effective efficiency $\hat{\epsilon} = \epsilon \cdot (1 - f)^2$

- Effective efficiency dominated by high fake rates

Report from RAMP challenge on fast vertexing

Florian Reiss
CTD 2019 03.04.19
• PV x-position resolution usually close to baseline
PV z-position resolution often degraded by $\sim 60\%$
Submissions

Approaches

Different approaches explored by participants, for example:

- Tuning baseline + improvements using machine learning
- Neural networks
- Clustering: from "simple" histogramming to DBScan

→ PV finding is in essence searching of accumulations of 3D lines

![Graph showing track z-position and MC PV z-position]
Submissions
Closer look - total score

- **M. Stahl**: tuning of baseline parameters
- **V. Macko**: baseline + decision based on ML (XGBClassifier)
- **J. Shlomi**: SciPy hierarchical clustering
- **L. Funke**: SciKit DBSCAN clustering
Submissions

Closer look - efficiency

- **M. Stahl**: tuning of baseline parameters
- **V. Macko**: baseline + decision based on ML (XGBC_classifier)
- **J. Shlomi**: SciPy hierarchical clustering
- **L. Funke**: SciKit DBSCAN clustering
Submissions
Closer look - fake rate

- **M. Stahl**: tuning of baseline parameters
- **V. Macko**: baseline + decision based on ML (XGBClассifier)
- **J. Shlomi**: SciPy hierarchical clustering
- **L. Funke**: SciKit DBSCAN clustering
Fake killing using XGBClassifier:
- highest score
- eliminates fakes and improves resolution
- sacrifices efficiency
Hierarchical clustering:

- one of the highest effective efficiencies
- decreased efficiency, fake rate similar to baseline
- suffers from z-resolution
DBSCAN clustering

- decent score
- low efficiency, low fake rate
- improves position resolution
Very enthusiastic participation, despite some challenges:
- Submitting non-python code not straight-forward
- Limited amount of training/testing data
- Rather short time scale (~ two days to work on problem)
- Scoreboard not working (mea culpa)

→ might have hindered effective collaboration
→ still many submissions exploring various approaches with reasonable results
Conclusions

Building LHCb's Great Reconstruction

This 4,500-year-old system used to pull alabaster stones up a steep slope was discovered at Hatnub, an ancient quarry in the Eastern Desert of Egypt.

Gradient was up to 20%

Posts, 50cm in diameter were buried deep into the slope so blocks could be hauled up
Conclusions

I would say it was a successful challenge:

• Various novel approaches explored
• Reasonable physics performances despite short time frame
• Participation by people not familiar with LHCb data/problem

What can we learn for our use case at LHCb?

• Fake killing using ML very interesting
• Clustering approaches could be useful for seeding in current algorithm

Many thanks to the RAMP team, with special thanks to:

• Akin Kazakci for RAMP introduction slides
• Balázs Kégl for his support when setting up the challenge
Thanks for your attention!
Back-up
From current to upgrade

- **VELO**: Strips to pixels, VeloPix chip, New RF boxes
- **TT to UT**: Smaller strips, SALT chip
- **IT/OT**: To SciFi+SiPM+PACIFIC
- **CAKO**: Remove PRS/SPD, New electronics
- **RICH**: HPDs to PMTs, Claro chip, New optics
- **MUON**: Remove M1, Add shielding, New electronics

All: 40MHz FE readout, software trigger only

Figure from here
Submissions
Closer look - x-position resolution

![Graph showing x-position resolution for various submissions]

- baseline
- Stumpf1stTrumpf3.0
- CheckThatOut
- cluster_v4
- rampensaeue_1

Florian Reiss
CTD 2019 03.04.19
Report from RAMP challenge on fast vertexing
Submissions
Closer look - z-position resolution

![Graph showing z-position resolution for different submissions]
How to set-up a challenge?

Ingredients:

- Data
- RAMP kit
  - introduction to problem
  - description of data set
  - example solution (baseline)
- Scoring
- Eager participants