Unveiling Hidden Physics at the LHC using Open Data

Making particle physics Open Data usable

Clemens Lange (@clelange)
Open Research Tools and Technologies devroom
FOSDEM’22

5th February 2022
Hello :-)  

I’m a particle physicist working on the CMS experiment at the Large Hadron Collider (LHC) at CERN, Switzerland.

Analysing the particle collisions provided by the LHC, recorded by the CMS detector.

27 km

LHC — the coolest place in the universe

Credits: https://build-your-own-particle-detector.org/
Particle detectors such as the CMS detector take up to 40,000,000 “3D photos” of the LHC collisions per second, 24/7 (almost) all year long.

Goal: understand the smallest building blocks of matter.

〜134 million readout channels — extraordinary levels of technical sophistication.

CMS is a collaboration of over 4000 particle physicists, engineers, computer scientists, technicians, and students from around 200 institutes and universities from more than 40 countries.
Need to translate detector data into “reconstructed” events ➔ software

CMSSW framework under Apache-2.0 license: https://github.com/cms-sw/cmssw/
> Since 2008, >1000 peer-reviewed papers published
  - Among them the discovery of the Higgs boson (No. 183)

> All published under open access (since 2014 under SCOAP³)
  - Preprints available on arXiv
  - Tabulated results largely available on HEPData portal

> Since 2014, have released 2.8 petabytes of open data available on the CERN Open Data Portal
We have measured processes that only happen a few dozen times per year.

However, there is still incredibly much that we do not (yet) understand!
We have measured processes that only happen a few dozen times per year.

However, there is still incredibly much that we do not (yet) understand!

What are dark matter particles? Are there other forces? Are there any other particles? Where do 18+7 standard model parameters come from? Why 3 sets of matter particles?

Credits: https://www.maxpixel.net/Haystack-Search-Find-Needle-In-A-Haystack-Needle-1706106
We have measured processes that only happen a few dozen times per year.

However, there is still incredibly much that we do not (yet) understand!

What are dark matter particles?
We have spin 0, 1/2, 1 elementary particles, expect spin 2 - where is spin 3/2?
Are there other forces?
Are there any other particles?
Where do 18+7 standard model parameters come from?
Why 3 sets of matter particles?

Why is universe matter? Where is anti-matter?
Why is the weak force left-handed?
Why is CP violated only a little in the standard model?
Why does the strong force not violate CP?
Why do quarks and leptons mix so differently?
Are neutrinos their own anti-particle?
Why no symmetry between fermions and bosons?

Credits: https://www.maxpixel.net/Haystack-Search-Find-Needle-In-A-Haystack-Needle-1706106
We have measured processes that only happen a few dozen times per year.

However, there is still incredibly much that we do not (yet) understand!

What are dark matter particles?
We have spin 0, 1/2, 1 elementary particles, expect spin 2 - where is spin 3/2?
Are there other forces?
Are there any other particles?
Where do 18+7 standard model parameters come from?
Why 3 sets of matter particles?

Why is universe matter? Where is anti-matter?
Why is the weak force left-handed?
Why is CP violated only a little in the standard model?
Why does the strong force not violate CP?
Why do quarks and leptons mix so differently?
Are neutrinos their own anti-particle?
Why no symmetry between fermions and bosons?

Why is dark matter ≈ matter?
Why does the proton not decay?
How can the weak scale be so much larger than gravity?
Can the different force strengths unify?
Why is the universe expanding?
What is dark energy?
How many dimensions are there?

Credits: https://www.maxpixel.net/Haystack-Search-Find-Needle-In-A-Haystack-Needle-1706106
We have measured processes that only happen a few dozen times per year. However, there is still incredibly much that we do not (yet) understand!

What are dark matter particles?
- We have spin 0, 1/2, 1 elementary particles, expect spin 2 - where is spin 3/2?
- Are there other forces?
- Are there any other particles?
- Where do 18+7 standard model parameters come from?
- Why 3 sets of matter particles?

Why do particles have such different masses?
- How do neutrinos get their mass?
- Why is the top quark so heavy?
- Why is the up quark so light?
- Why is the Higgs mass so small?

Why is universe matter? Where is anti-matter?
- Why is the weak force left-handed?
- Why is CP violated only a little in the standard model?
- Why does the strong force not violate CP?
- Why do quarks and leptons mix so differently?
- Are neutrinos their own anti-particle?
- Why no symmetry between fermions and bosons?

Why is dark matter ≈ matter?
- Why does the proton not decay?
- How can the weak scale be so much larger than gravity?
- Can the different force strengths unify?
- Why is the universe expanding?
- What is dark energy?
- How many dimensions are there?

Credits: https://www.maxpixel.net/Haystack-Search-Find-Needle-In-A-Haystack-Needle-1706106
We have measured processes that only happen a few dozen times per year. However, there is still incredibly much that we do not (yet) understand!

Maybe the answer to one of these questions lies hidden in the LHC data — with the data openly available, anyone could unveil it!
At the end of 2020, all large LHC experimental collaborations have endorsed a new open data policy  
- Following existing CMS policy

Commit to publicly **releasing data required to make scientific studies**

Data and simulation will start to be released approximately five years after collection (50%)
- Released under the Creative Commons CC0 waiver
- Full dataset by the close of the experiment

CMS Open Data

- **Level 1**: Open access publication and additional numerical data
- **Level 2**: Simplified data for Outreach and Education
- **Level 3**: Reconstructed data and the software to analyse them
- **Level 4**: Raw data, and the software to reconstruct and analyse them
Open Data needs to be FAIR:

- **Findable** ➞ CERN Open Data Portal records
- **Accessible** ➞ reliable storage and access technology
- **Interoperable** ➞ provide good documentation, avoid jargon
- **Reusable** ➞ preserve software (and hardware to run it if needed), data provenance, workflows
Collider data is complex

Theory
(perturbation theory)
/ LHC pp collisions

Parton Shower
+ Hadronisation
(non-perturbative)

Experiment
Collider data is complex

Theory (perturbation theory) / LHC pp collisions

On average, 32 simultaneous proton-proton collisions
Collider data is complex

Theory
(perturbation theory)
/ LHC pp collisions

Parton Shower
+ Hadronisation
(non-perturbative)

Experiment
Collider data is complex

**Theory** (perturbation theory) / LHC pp collisions

**Parton Shower** + Hadronisation (non-perturbative)

**Experiment**

**Pileup+Underlying Event**

- $q, g$
- $p$
- $\pi, K, ...$
Collider data is complex

Theory (perturbation theory) / LHC pp collisions

Parton Shower + Hadronisation (non-perturbative)

Pileup+Underlying Event

Detector noise

q, g

p

π, K, ...
Analysing collider data is very challenging

- We can only store 0.025% of the collisions (1 in 40,000 events or 1,000 events per second)
  - A multi-stage trigger system selects events of interest — this bias needs to be taken into account when performing an analysis

- A raw event has the size of about 2 megabytes
  - We have recorded tens of billions of events, and simulated even more
  - Size can be reduced at the cost of information loss — expertise required
  - We currently release largely “Analysis object data” (500 kB/event)

- Billions of events need significant computing power for processing

- A complete physics analysis needs to take dozens of systematic uncertainties into account
  - Understanding the relevance of individual uncertainties needs expertise

- Statistical interpretation needs particular care
Collaborations make huge internal review effort (months to years) to **ensure accurate interpretation of the data**

- False claims (also from OD users) could risk erosion of public trust

**Small deviations can make a big difference**

- A few events could mean a discovery

**Physics objects definitions are analysis-dependent**

- An electron in one analysis might not be one in another due to different reconstruction algorithms used
Beyond the data sets available on the CERN Open Data Portal, we provide:

- Analysis examples with different levels of complexity (scientific and education)
- A separate CMS Open Data Guide
  - In particular, trying to explain how to use the data and what to do with them in addition to what is in the data
- Workshops with Software Carpentry style tutorials:
  - 2020 CMS Open Data Workshop for Theorists
  - 2021 CMS Open Data Workshop
The analysis part usually takes a lot of iterations.
We provide simplified analysis examples to lower the threshold to get started

- **Pro:** users can obtain a result/plot rather quickly
- **Contra:** these are usually far from realistic

At least the first step of the analysis chain requires substantial computing resources, ideally high-throughput batch processing systems

- Data sets can be processed in an “embarrassingly parallel” way
- We provide examples/tutorials on using public cloud resources

Simulation of new processes needs CMSSW

- Parts of the software are more than a decade old → interfacing can be difficult
When developing examples, we now aim to use open tools combined with container technologies for automatic and regular validation:

- Continuous integration using CERN’s GitLab installation
- Simpler examples also run as GitHub actions

For easier usability, we provide examples on how to get out of the HEP-specific software tool chain to industry standard tools.

By now, CMS Open Data have been used for both actual physics results and also several computing-related projects.

Eventually, the data might be used to unveil hidden physics!