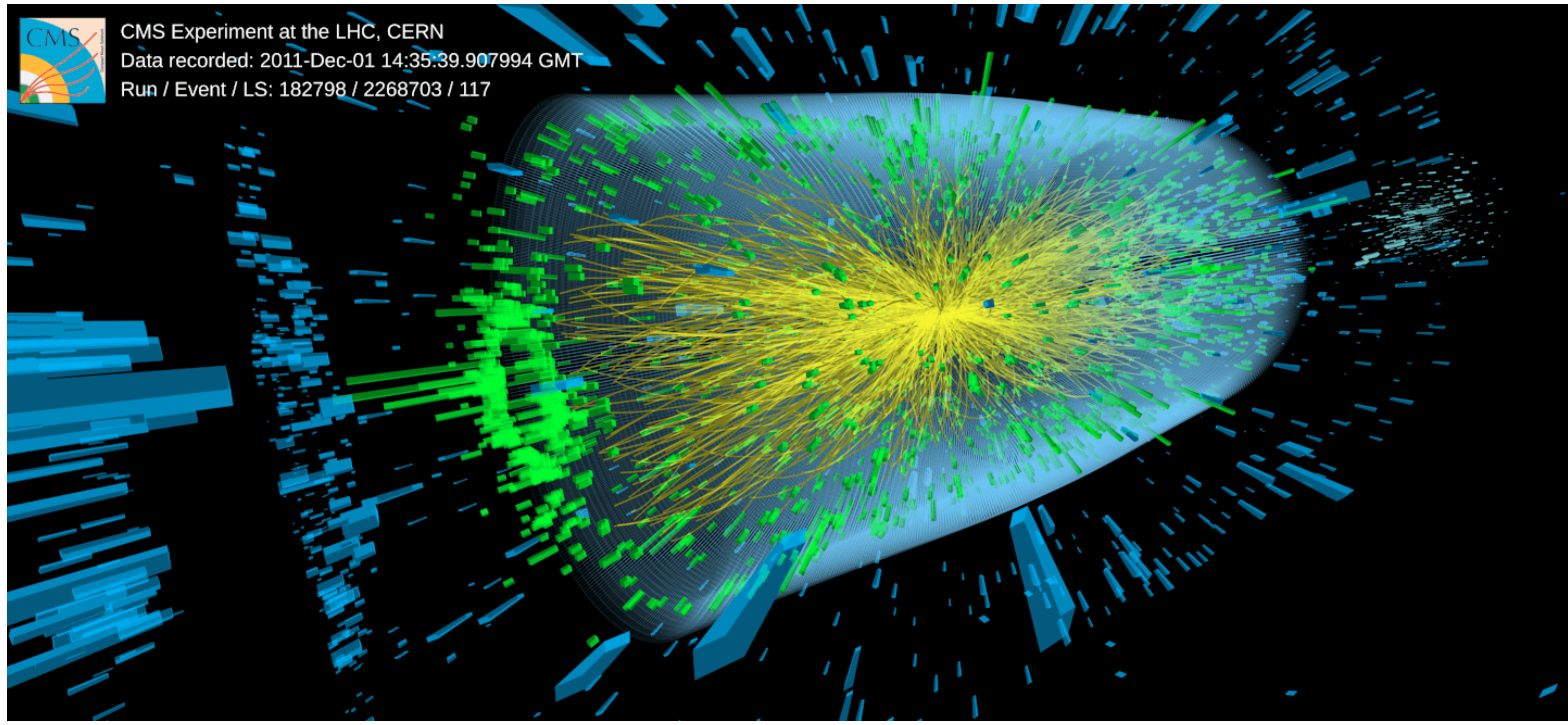
Unveiling Hidden Physics at the LHC using Open Data

Making particle physics Open Data usable


CMS Experiment at the LHC, CERN
Data recorded: 2011-Dec-01 14:35:39.907994 GMT
Run / Event / LS: 182798 / 2268703 / 117

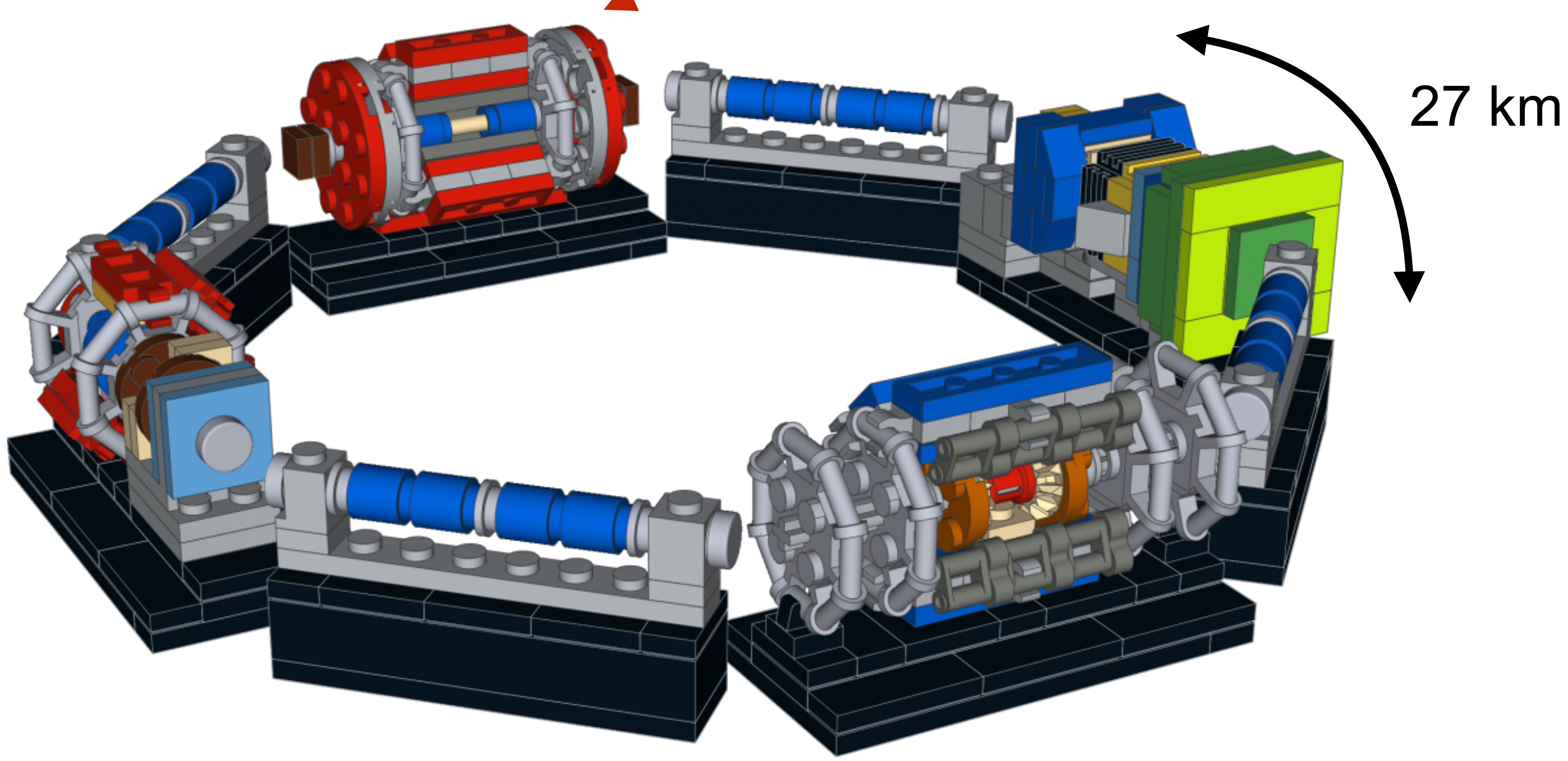
Clemens Lange (@clelange)

Open Research Tools and Technologies devroom
FOSDEM'22

5th February 2022

Hello :-)

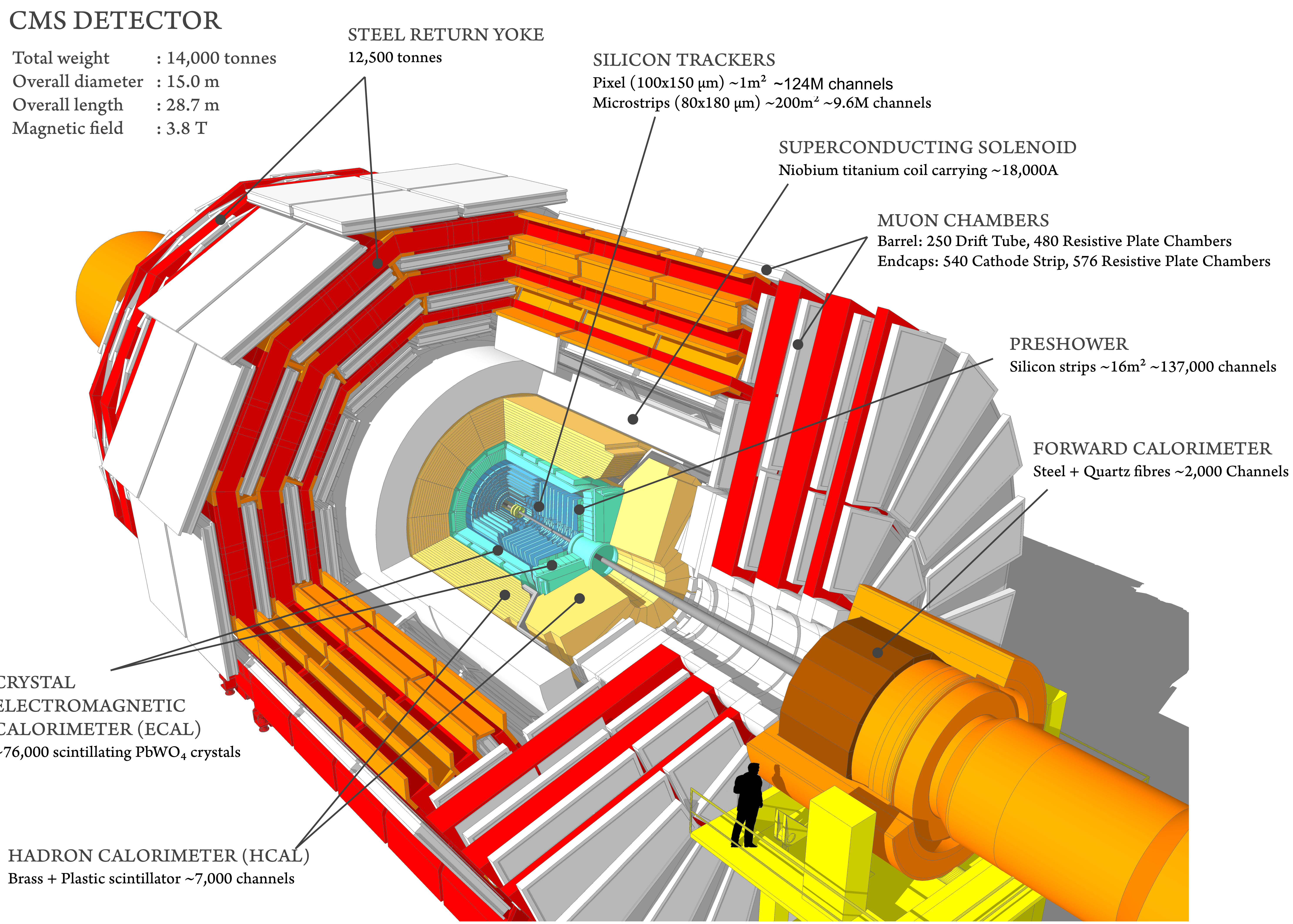
- I'm a particle physicist working on the CMS experiment at the Large Hadron Collider (LHC) at CERN, Switzerland
- Analysing the particle collisions provided by the LHC, recorded by the CMS detector



Credits: https://build-your-own-particle-detector.org/

LHC — the **coolest** place in the universe

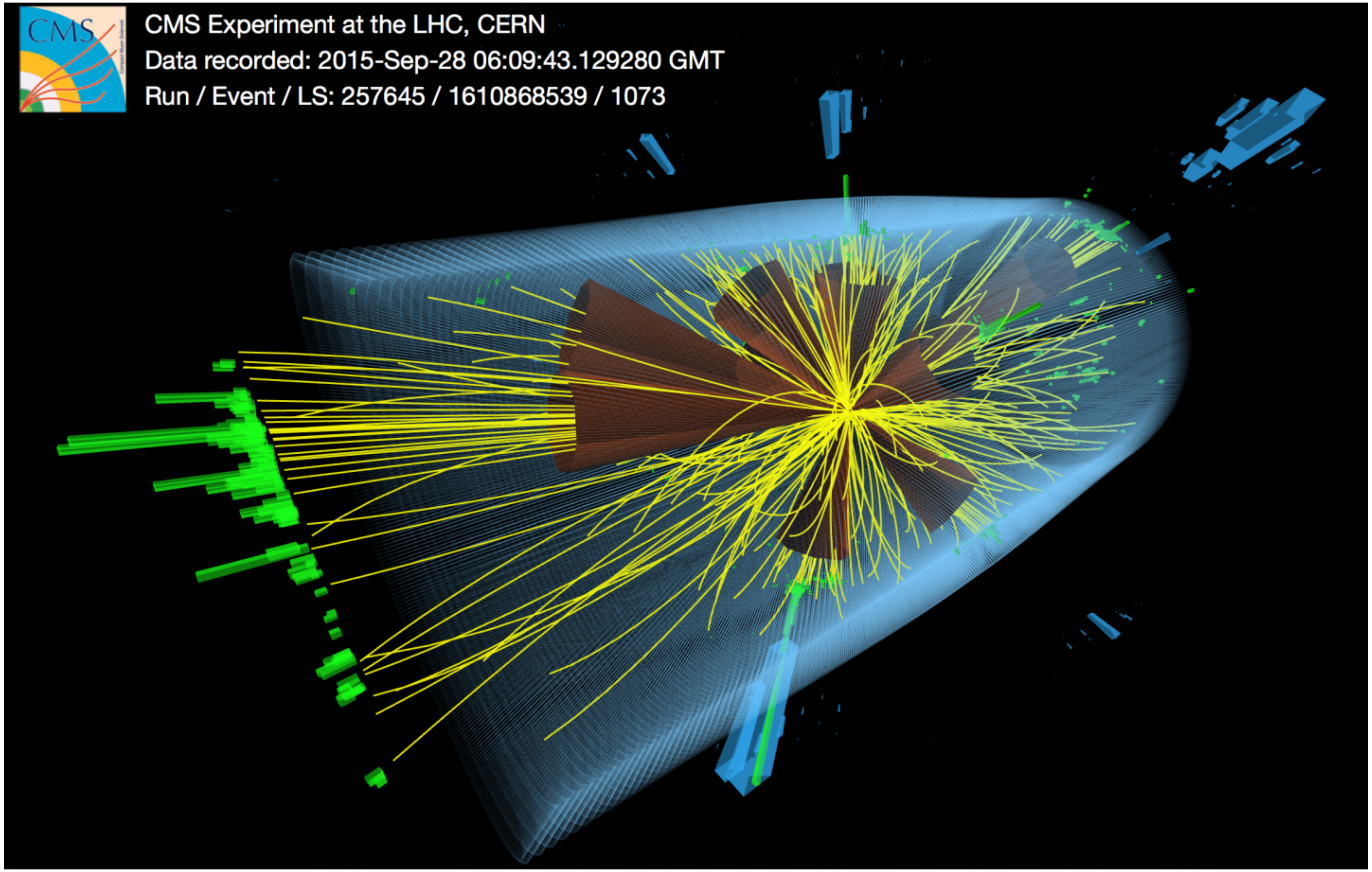
The CMS experiment: a huge and fast camera


CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~1m² ~124M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL
ELECTROMAGNETIC
CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

- 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

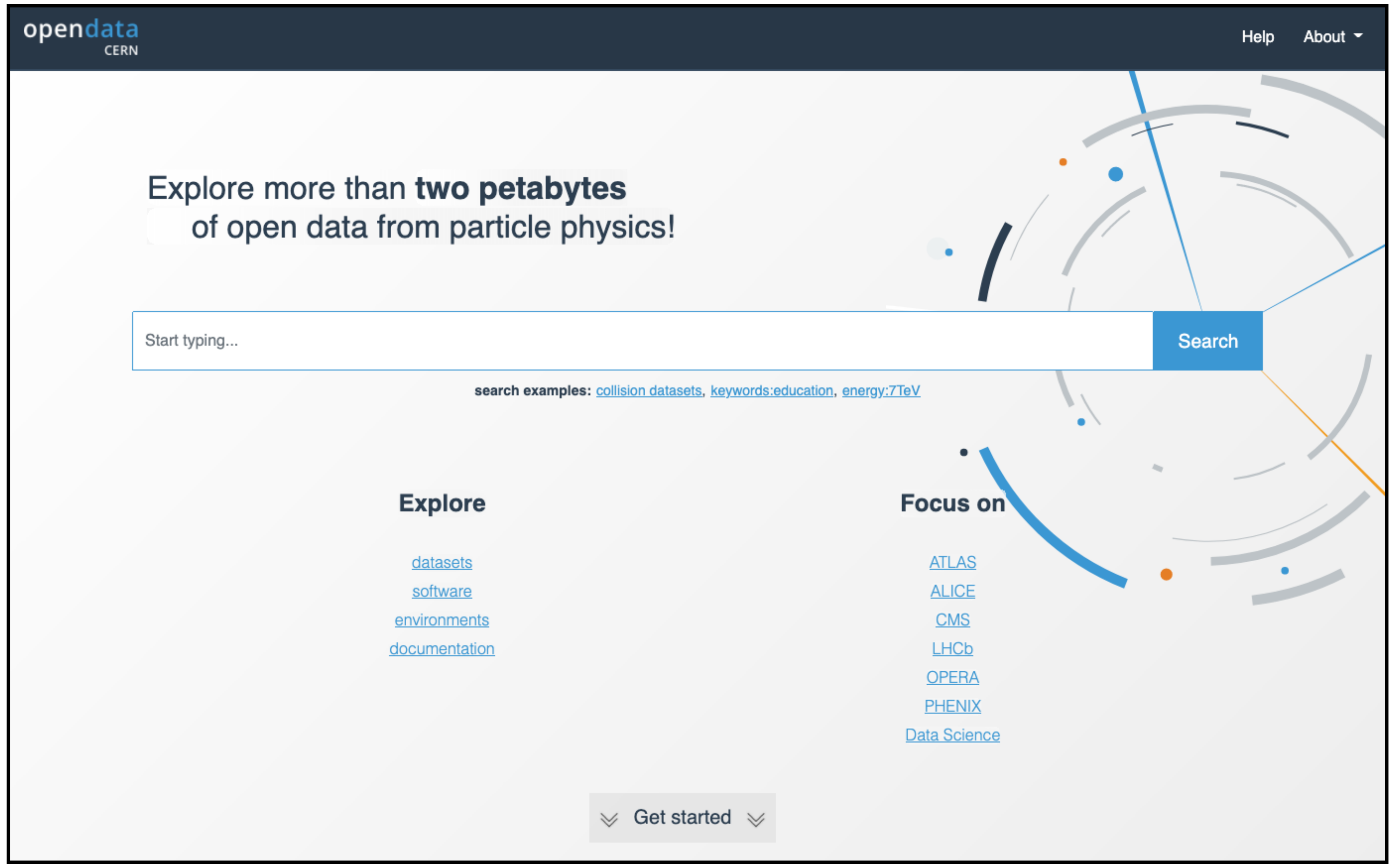
From detector signals to “events”


CMS Experiment at the LHC, CERN
Data recorded: 2015-Sep-28 06:09:43.129280 GMT
Run / Event / LS: 257645 / 1610868539 / 1073

- Need to translate detector data into “reconstructed” events ➜ software
- CMSSW framework under Apache-2.0 license: https://github.com/cms-sw/cmssw/

Physics results

- 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



Explore more than **two petabytes** of open data from particle physics!

search examples: collision datasets, keywords:education, energy:7TeV

Explore

datasets
software
environments
documentation

Focus on

ATLAS
ALICE
CMS
LHCb
OPERA
PHENIX
Data Science

Finding the needle in the haystack

- 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!

Credits: https://www.maxpixel.net/Haystack-Search-Find-Needle-In-A-Haystack-Needle-1706106

Finding the needle in the haystack

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?

- that only dozen times per year

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

Credits: https://www.maxpixel.net/Haystack-Search-Find-Needle-In-A-Haystack-Needle-1706106

Finding the needle in the haystack

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?

- that only dozen times per year
- However, there is still incredibly much that we do not (yet) understand!

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?

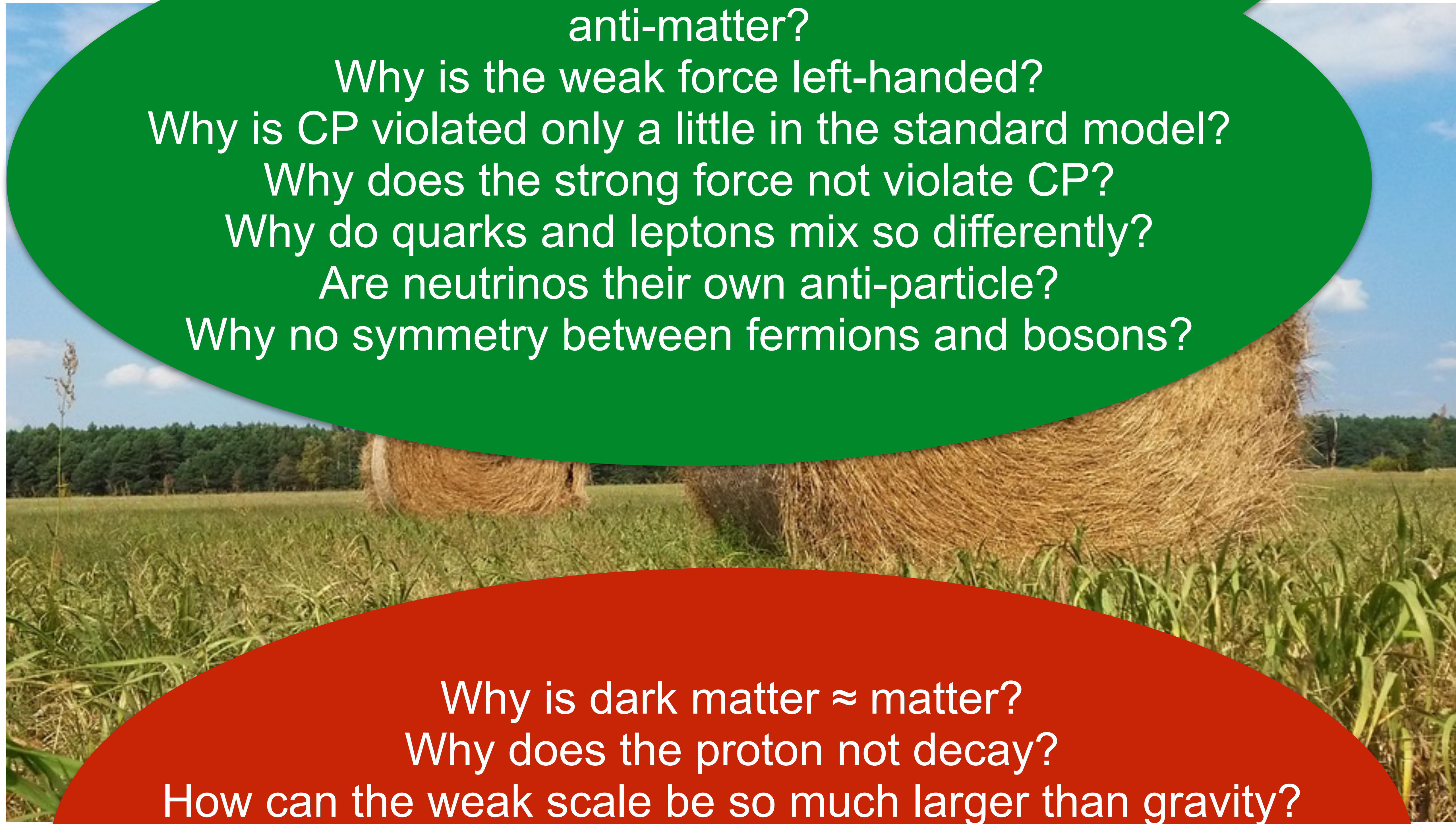
Credits: https://www.maxpixel.net/Haystack-Search-Find-Needle-In-A-Haystack-Needle-1706106

Finding the needle in the haystack

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?

- s
that only happ ew dozen times per year

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



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

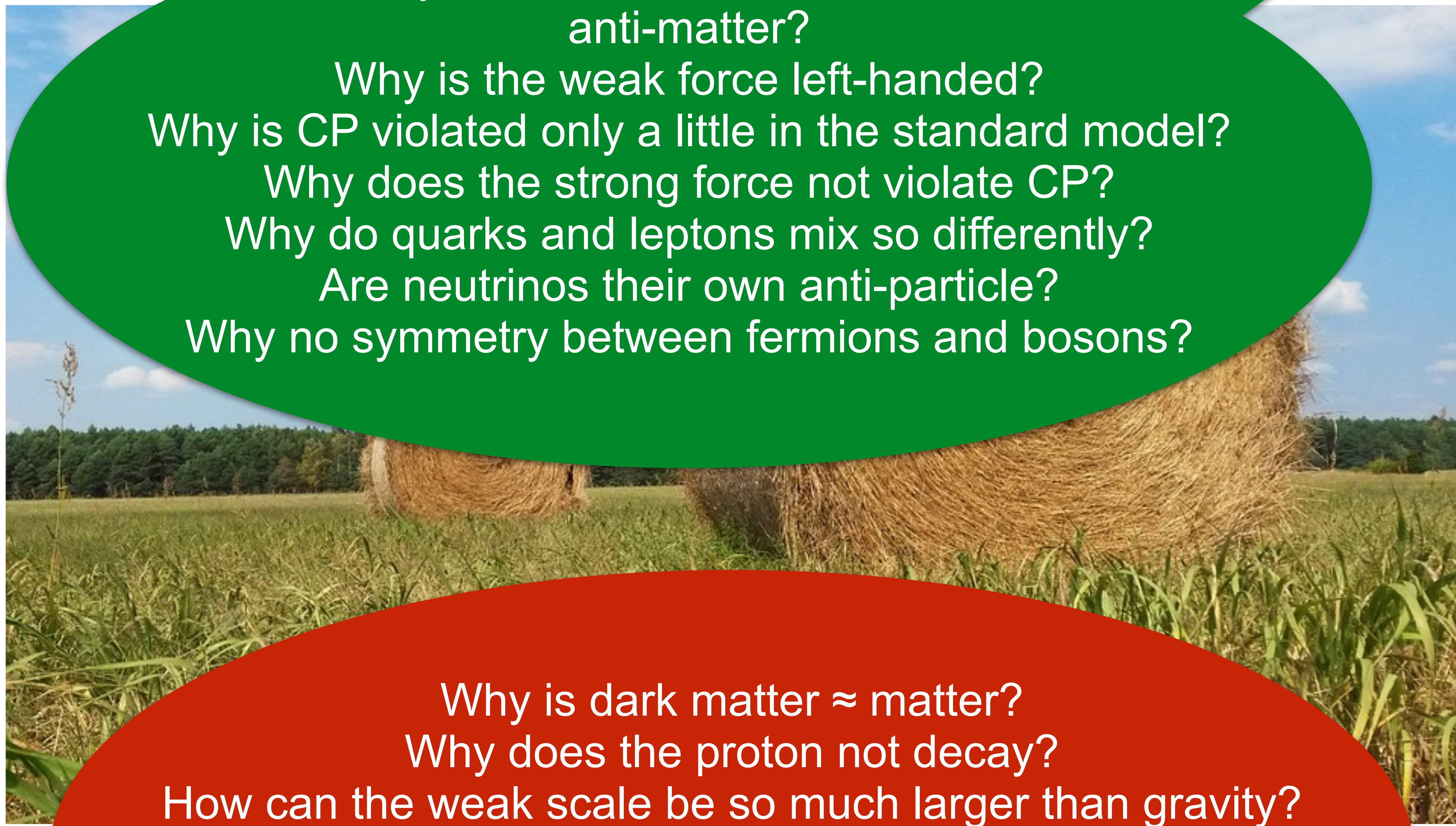
Finding the needle in the haystack

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?

that only dozen
times per year

>How till incredibly

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

Finding the needle in the haystack

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 matte

that only happ
times per year

>How

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?
trong force not violate CP?
l leptons mix so differently?
heir own anti-particle?
tween fermions and bosons?

Maybe the answer to one of these questions lies hidden in the LHC data — with the data openly available, anyone could unveil it!

Why do particles have such diffe
How do neutrinos get thei
Why is the top quark so heavy?
Why is the up quark so light?
Why is the Higgs mass so small?

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

CMS Open Data

- 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

higher computational effort

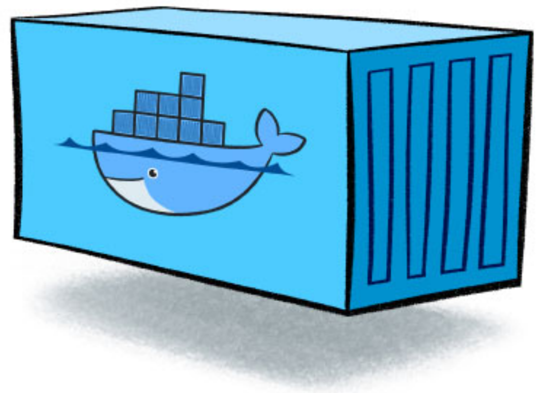
- Level 1: Open access publication and additional numerical data
- Level 2: Simplified data for Outreach and Education
- **Level 3**: Reconstructed data and the software to analyse them
- Level 4: Raw data, and the software to reconstruct and analyse them

Challenges with open data

Data: available ≠ usable

Open Data needs to be FAIR:

- **F**indable ➜ CERN Open Data Portal records
- **A**ccessible ➜ reliable storage and access technology
- **I**nteroperable ➜ provide good documentation, avoid jargon
- **R**eusable ➜ preserve software (and hardware to run it if needed), data provenance, workflows



Collider data is complex

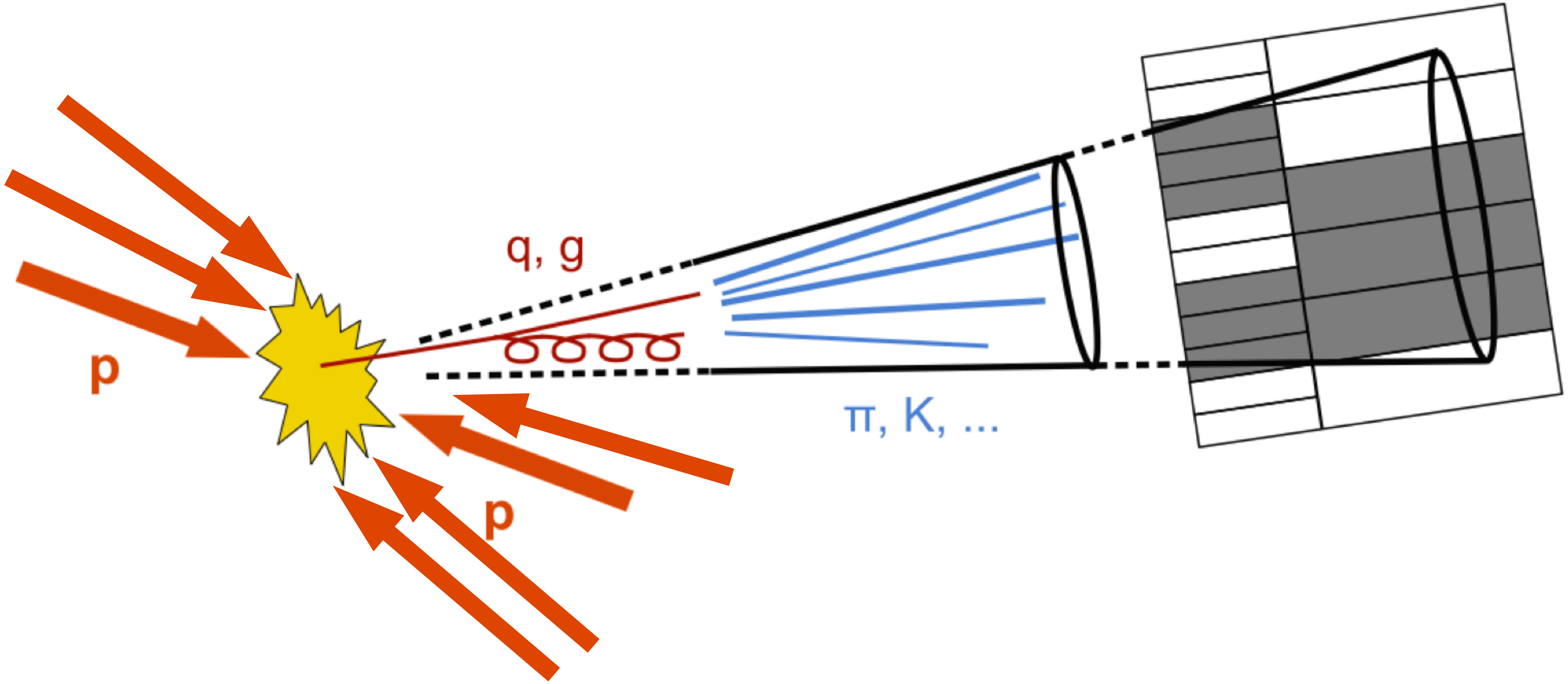
Theory
(perturbation theory)
/ LHC pp collisions

↔

Parton Shower
+ Hadronisation
(non-perturbative)

↔

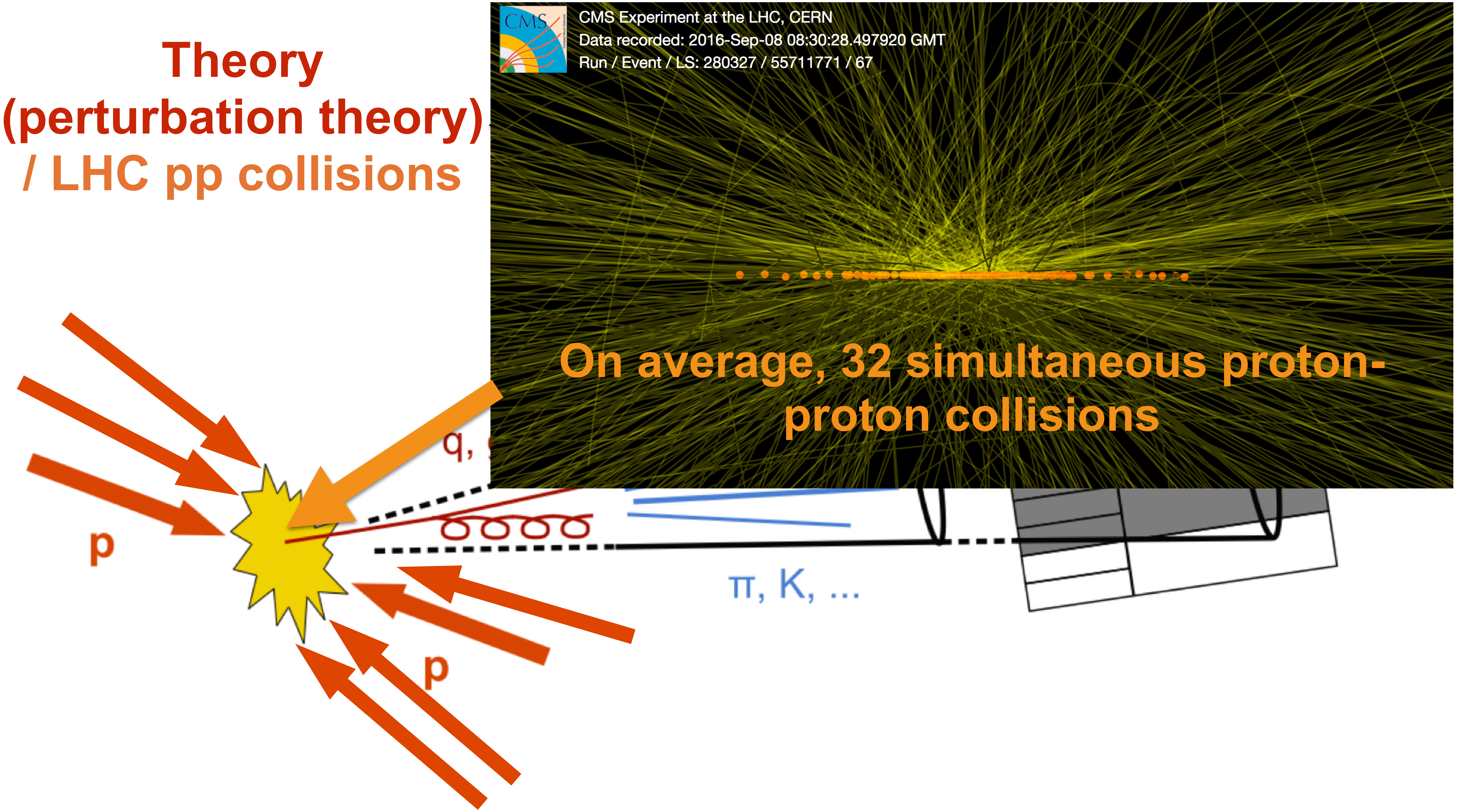
Experiment



Collider data is complex

Theory (perturbation theory)

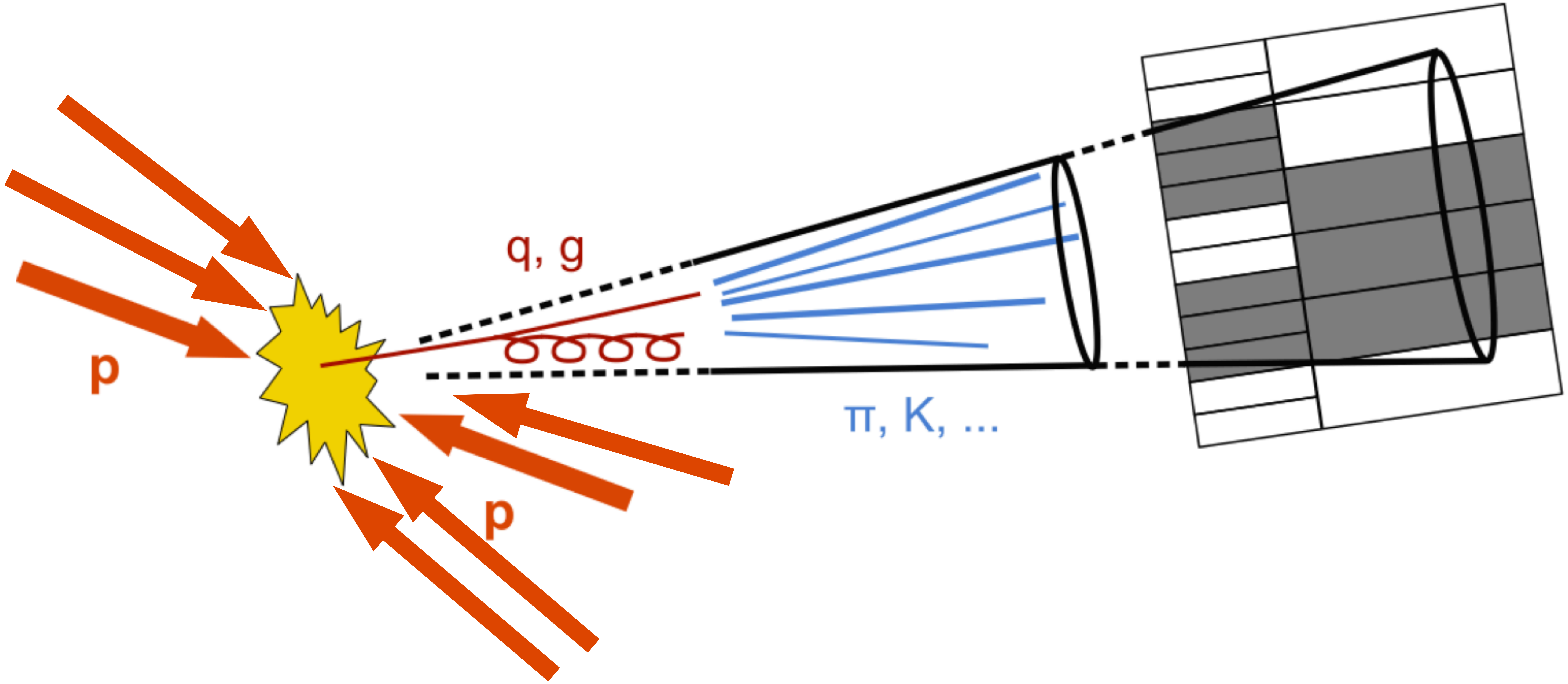
/ LHC pp collisions


CMS Experiment at the LHC, CERN
Data recorded: 2016-Sep-08 08:30:28.497920 GMT
Run / Event / LS: 280327 / 55711771 / 67
q,
p
p
π, K, ...

On average, 32 simultaneous proton-proton collisions

Collider data is complex

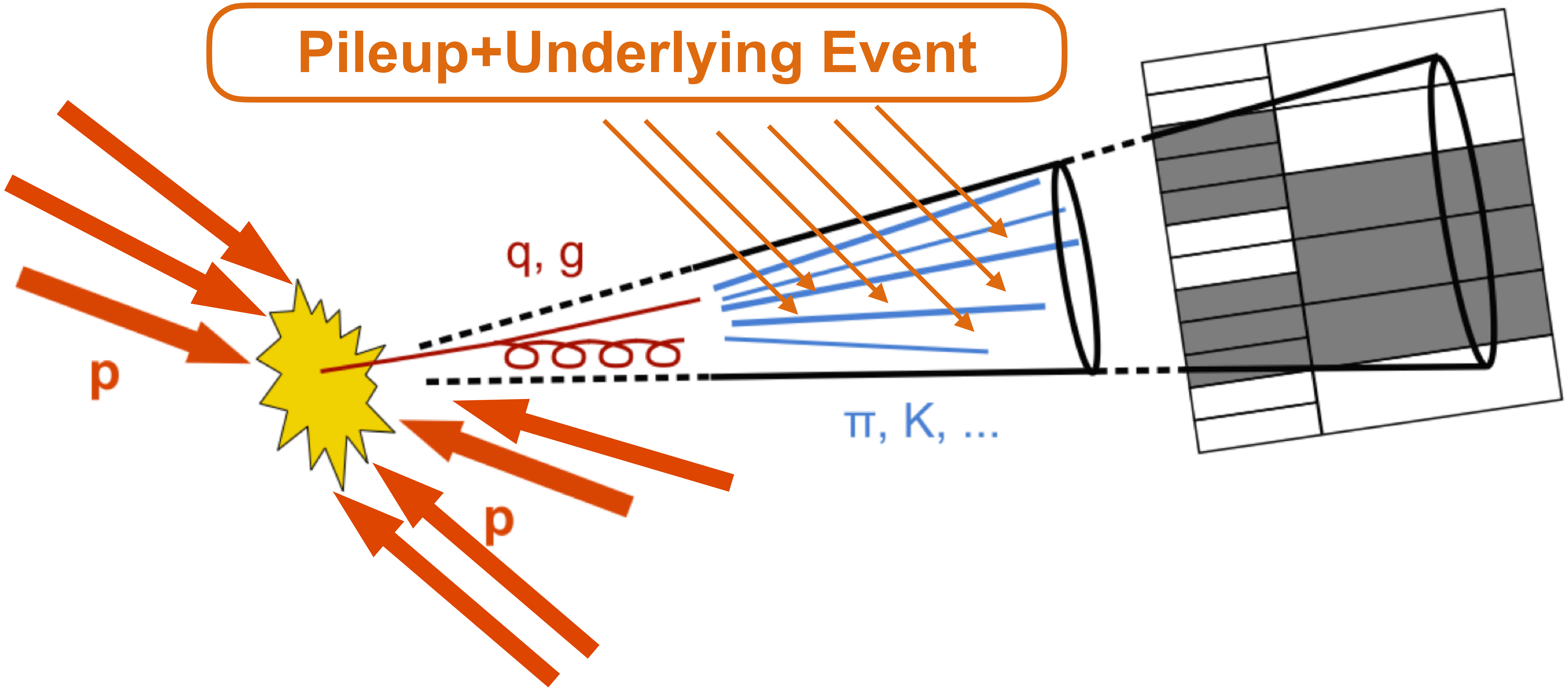
Theory (perturbation theory) / LHC pp collisions ↔ **Parton Shower + Hadronisation (non-perturbative)** ↔ **Experiment**



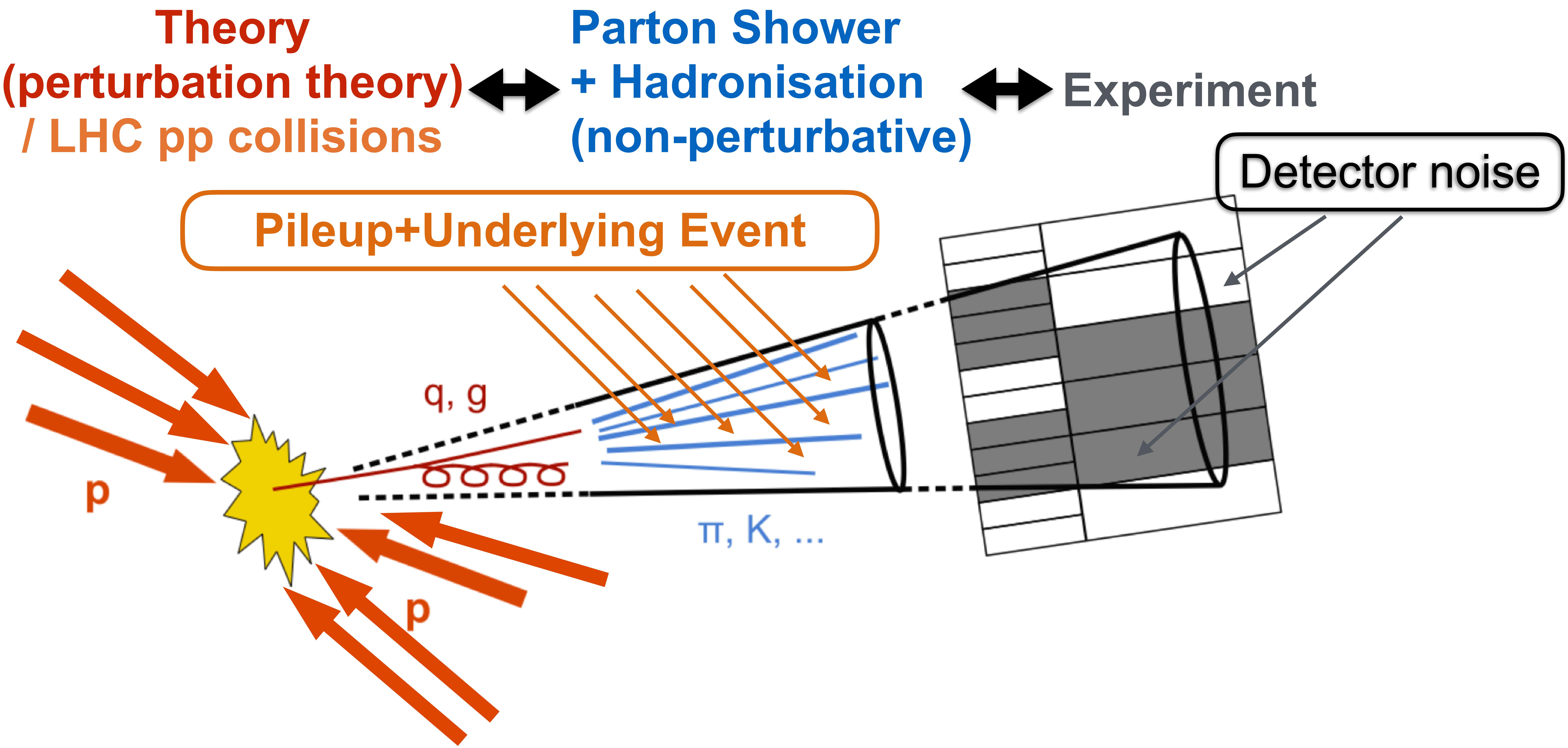
Collider data is complex

Theory (perturbation theory) / LHC pp collisions ⟷ **Parton Shower + Hadronisation (non-perturbative)** ⟷ **Experiment**

Pileup+Underlying Event



Collider data is complex



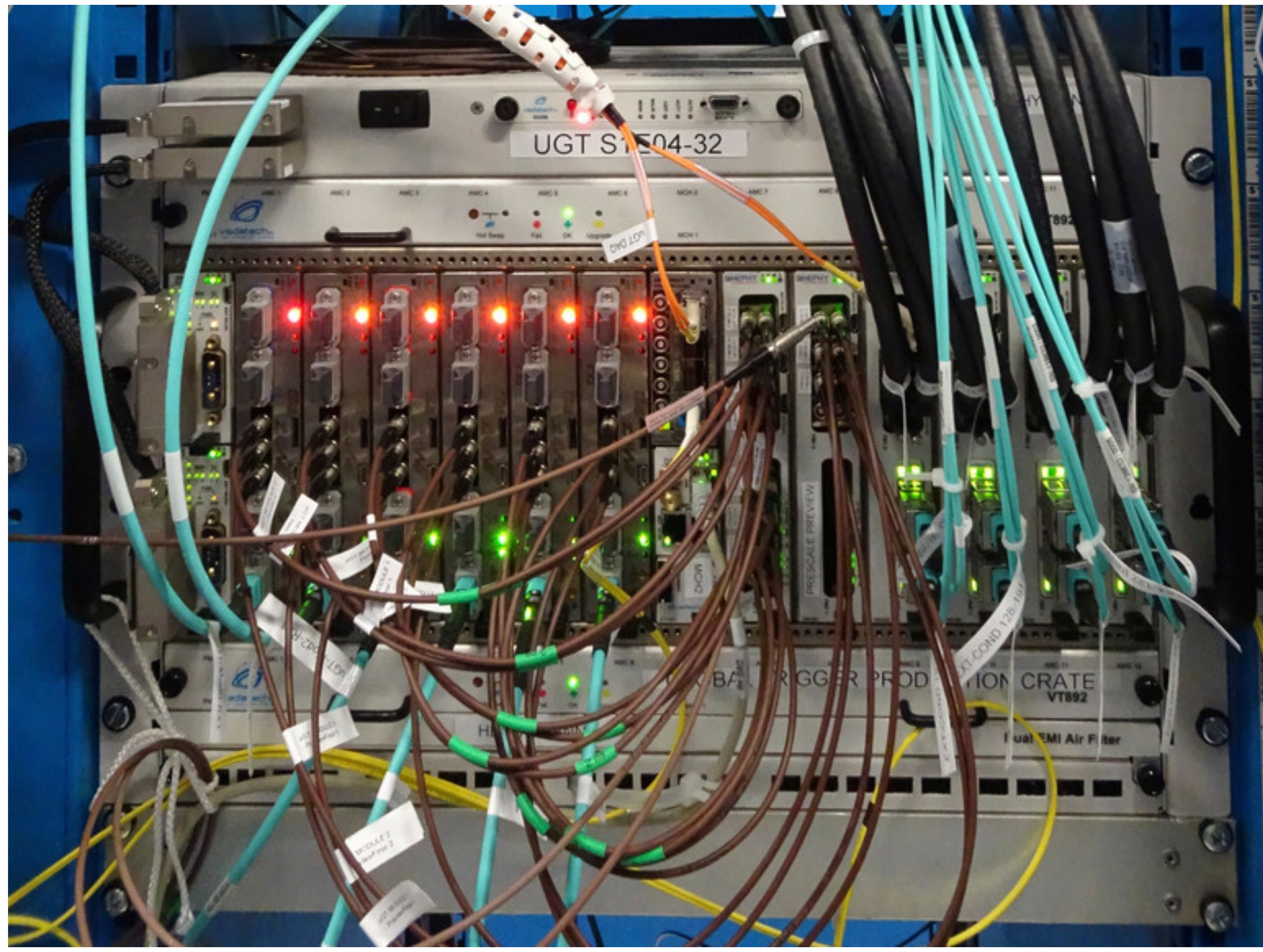
Theory (perturbation theory) / LHC pp collisions ↔ Parton Shower + Hadronisation (non-perturbative) ↔ Experiment

Pileup+Underlying Event

Detector noise

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



Further challenges in high-energy physics data analysis

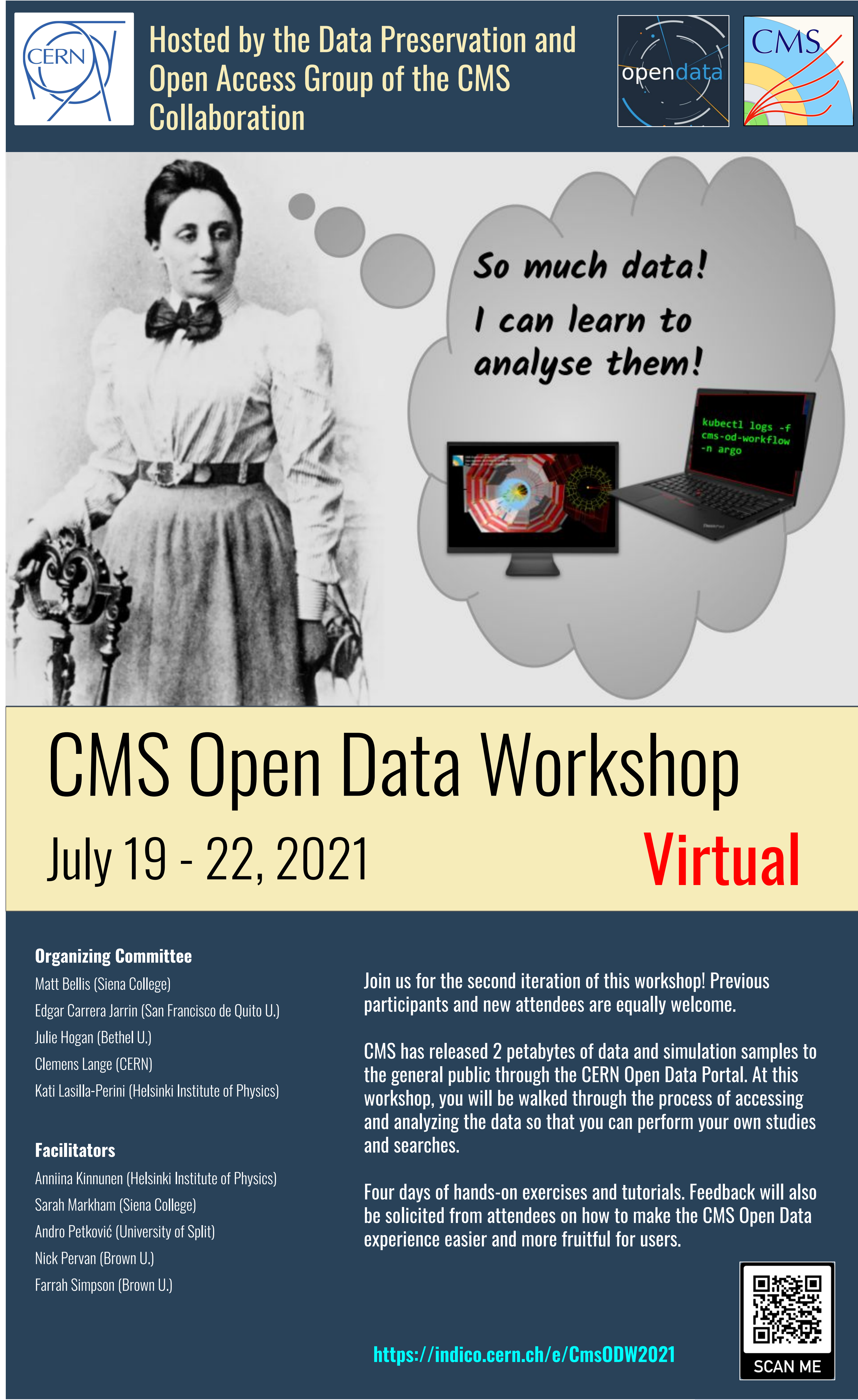
- Collaborations make huge internal review effort (months to years) to **ensure accurate interpretation of the data**
  - False claims (also from OD users) could risk erosion of public trust
- **Small deviations can make a big difference**
  - A few events could mean a discovery
- Physics objects definitions are analysis-dependent
  - An electron in one analysis might not be one in another due to different reconstruction algorithms used

“Review” by Nick Youngson CC-BY-SA 3.0 Alpha Stock Images

Addressing the challenges

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



Hosted by the Data Preservation and Open Access Group of the CMS Collaboration

```
kubectl logs -f
cms-od-workflow
-n argo
```

CMS Open Data Workshop

July 19 - 22, 2021 Virtual

Organizing Committee

Matt Bellis (Siena College)
Edgar Carrera Jarrin (San Francisco de Quito U.)
Julie Hogan (Bethel U.)
Clemens Lange (CERN)
Kati Lasilla-Perini (Helsinki Institute of Physics)

Facilitators

Anniina Kinnunen (Helsinki Institute of Physics)
Sarah Markham (Siena College)
Andro Petković (University of Split)
Nick Pervan (Brown U.)
Farrah Simpson (Brown U.)

Join us for the second iteration of this workshop! Previous participants and new attendees are equally welcome.

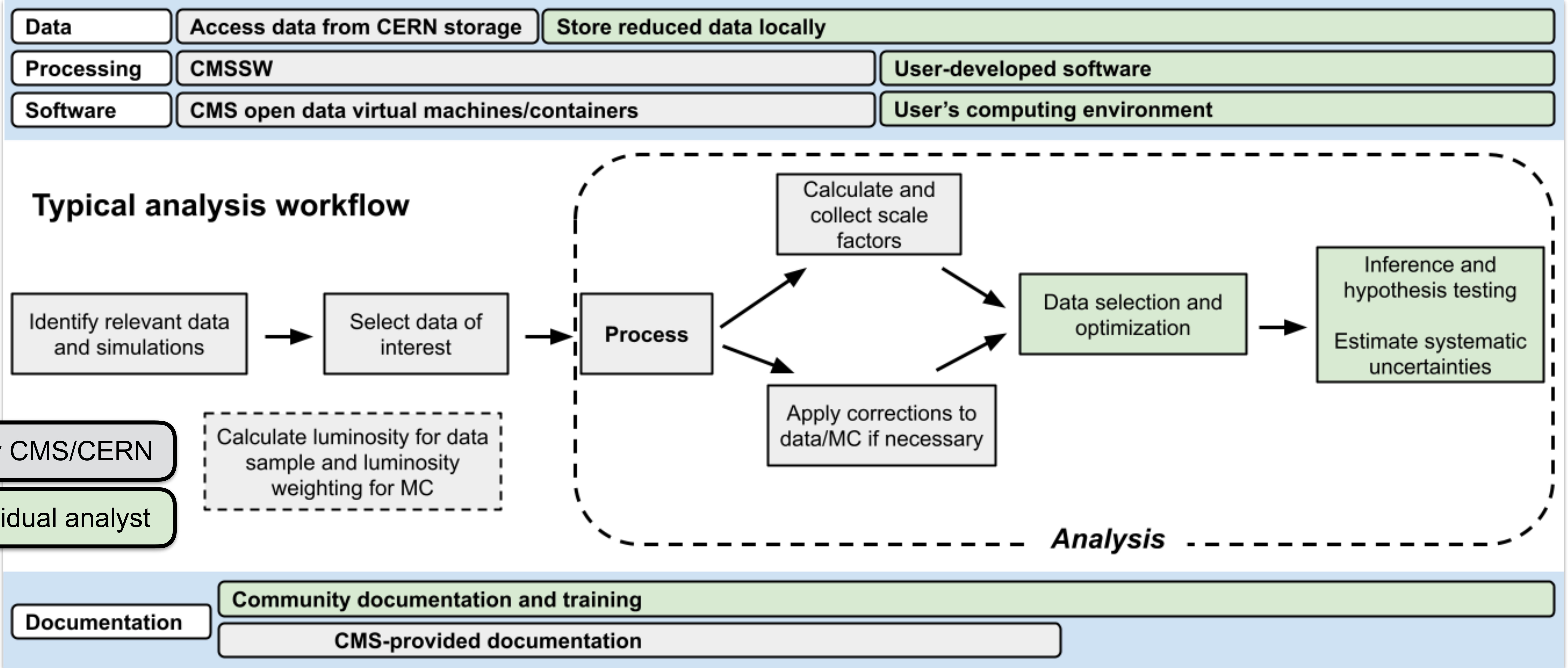
CMS has released 2 petabytes of data and simulation samples to the general public through the CERN Open Data Portal. At this workshop, you will be walked through the process of accessing and analyzing the data so that you can perform your own studies and searches.

Four days of hands-on exercises and tutorials. Feedback will also be solicited from attendees on how to make the CMS Open Data experience easier and more fruitful for users.

https://indico.cern.ch/e/CmsODW2021

SCAN ME

Analysis flowchart


Data
Access data from CERN storage
Store reduced data locally
Processing
CMSSW
User-developed software
Software
CMS open data virtual machines/containers
User's computing environment
Typical analysis workflow
Identify relevant data and simulations
Select data of interest
Calculate luminosity for data sample and luminosity weighting for MC
Process
Calculate and collect scale factors
Apply corrections to data/MC if necessary
Data selection and optimization
Inference and hypothesis testing
Estimate systematic uncertainties
Analysis
Documentation
Community documentation and training
CMS-provided documentation

Provided by CMS/CERN

Left to individual analyst

- The analysis part usually takes a lot of iterations

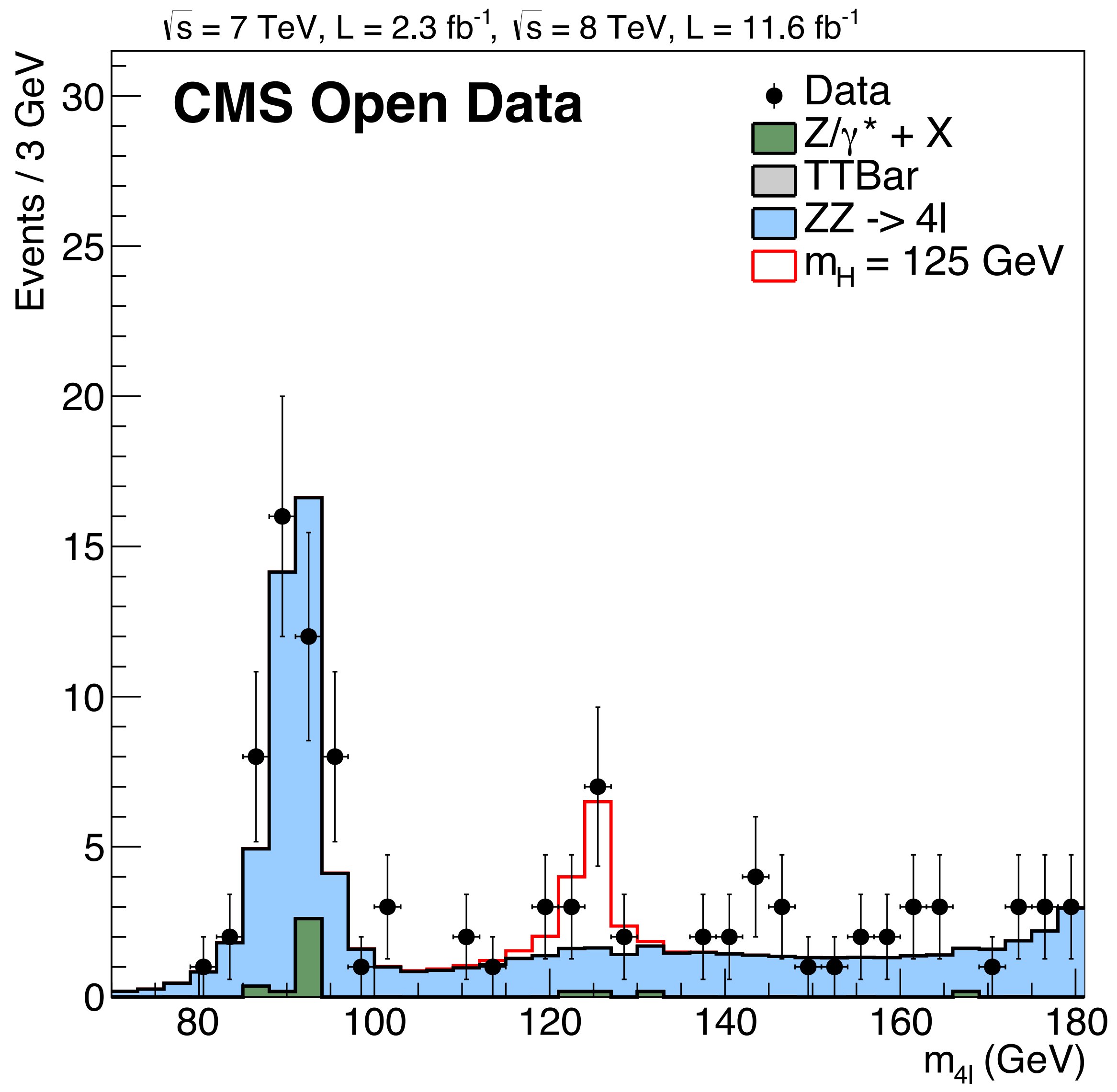
Computational challenges

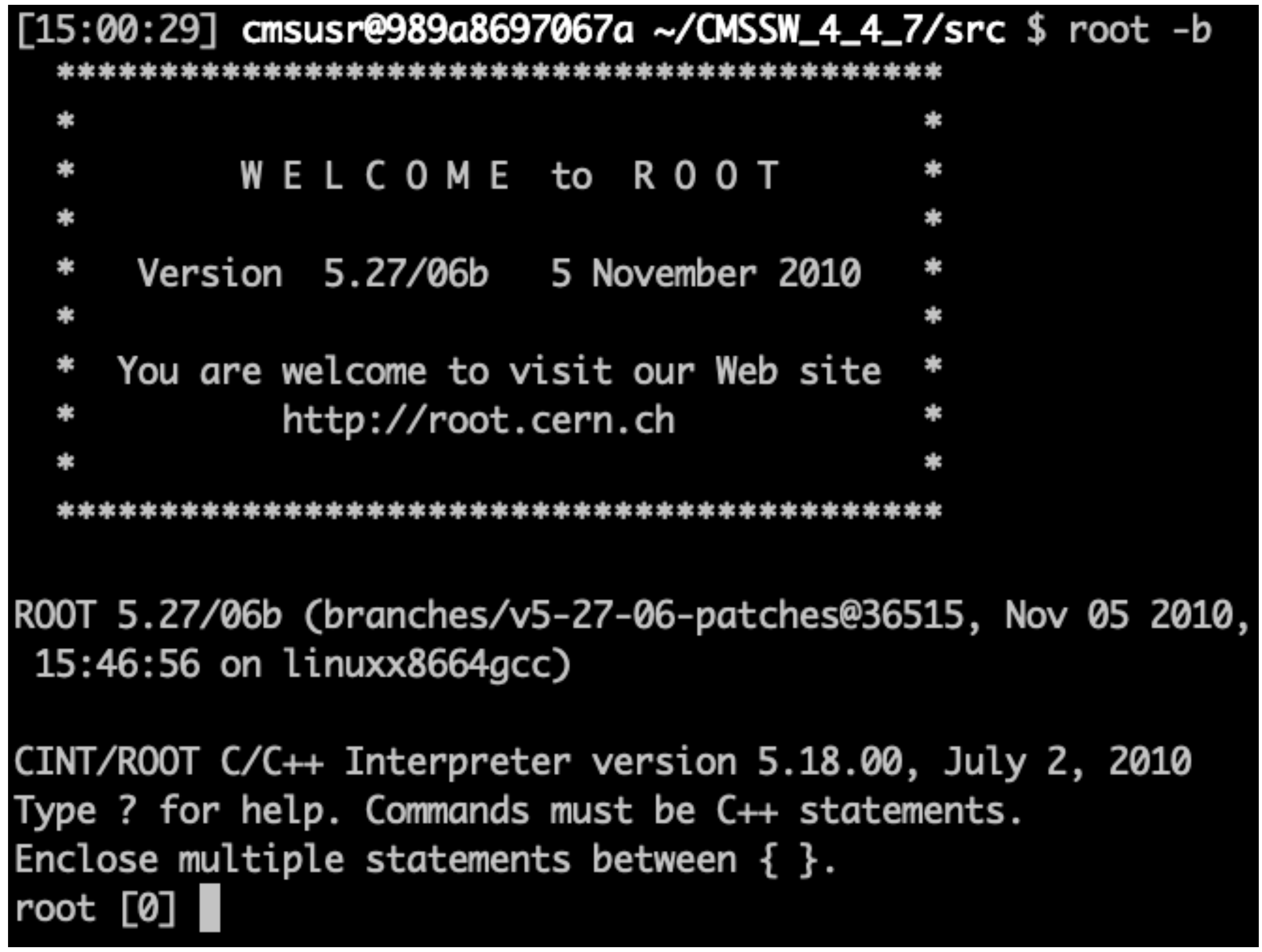
- We provide simplified analysis examples to lower the threshold to get started
  - Pro: users can obtain a result/plot rather quickly
  - Contra: these are usually far from realistic
- At least the first step of the analysis chain requires substantial computing resources, ideally high-throughput batch processing systems
  - Data sets can be processed in an “embarrassingly parallel” way
  - We provide examples/tutorials on using public cloud resources

kubernetes

- Simulation of new processes needs CMSSW
  - Parts of the software are more than a decade old ➜ interfacing can be difficult

DOI:10.7483/OPENDATA.CMS.JKB8.RR42





```
[15:00:29] cmsusr@989a8697067a ~/CMSSW_4_4_7/src $ root -b
  *******************************************
  *                                         *
  *        W E L C O M E  to  R O O T       *
  *                                         *
  *   Version   5.27/06b    5 November 2010 *
  *                                         *
  *  You are welcome to visit our Web site  *
  *          http://root.cern.ch            *
  *                                         *
  *******************************************

ROOT 5.27/06b (branches/v5-27-06-patches@36515, Nov 05 2010,
 15:46:56 on linuxx8664gcc)

CINT/ROOT C/C++ Interpreter version 5.18.00, July 2, 2010
Type ? for help. Commands must be C++ statements.
Enclose multiple statements between { }.
root [0] █
```

Keeping up

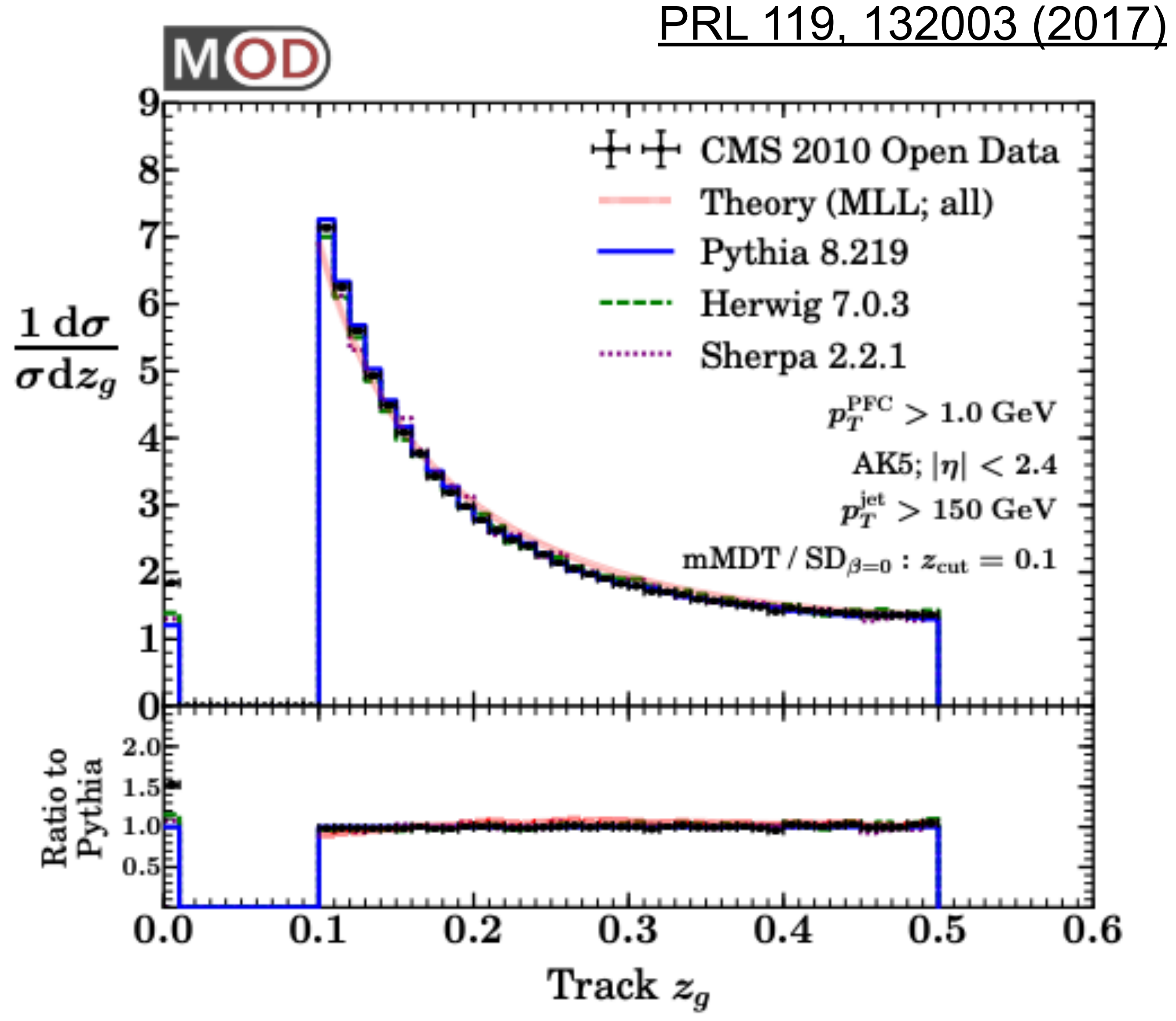
- When developing examples, we now aim to use open tools combined with container technologies for automatic and regular validation
  - Continuous integration using CERN’s GitLab installation
  - Simpler examples also run as GitHub actions
- For easier usability, we provide examples on how get out of the HEP-specific software tool chain to industry standard tools

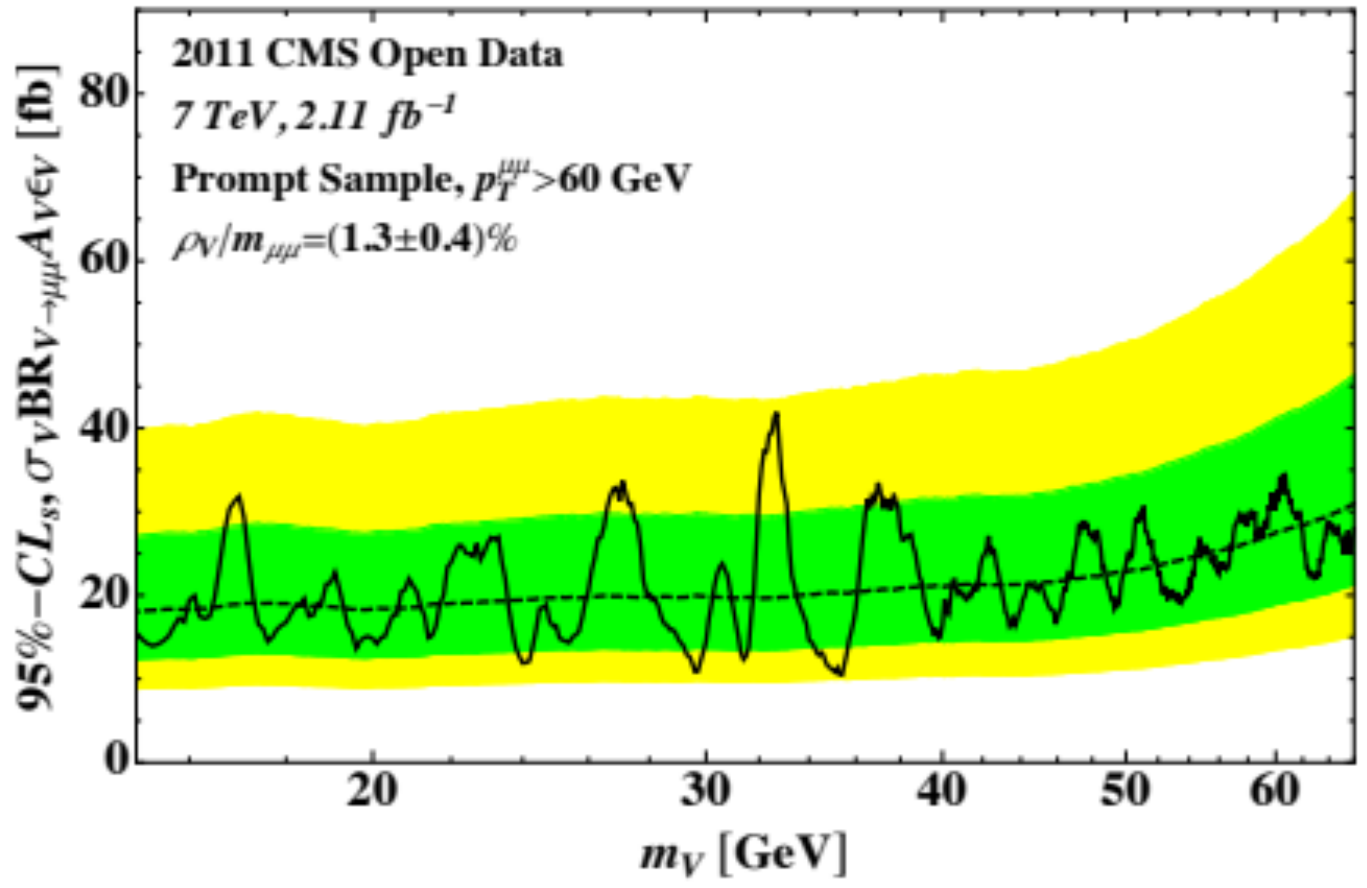
ROOT Data Analysis Framework → NumPy

- 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!

PRL 119, 132003 (2017)





Phys. Rev. D 100, 015021 (2019)