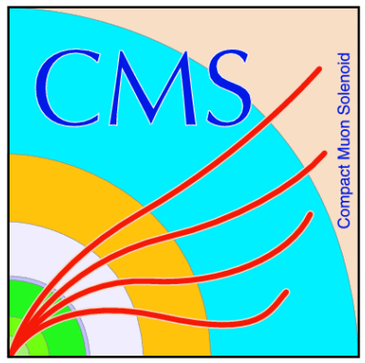
CMS Open Data

Status and plans

Clemens Lange (CERN),
on behalf of the CMS Collaboration

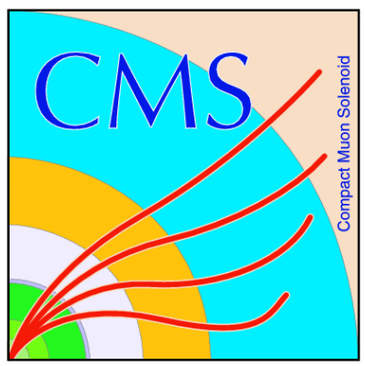
(Re)interpreting the results of new physics searches at the LHC
Imperial College, London
2nd April 2019

What is CMS Open Data about



- **Provide a setup** to do whatever a CMS member did, could have done or could still do with the CMS data, without any formal constraint for non-CMS members
- (Approximately) **reproduce** the **results**, or **produce** new ones
- Modify whatever you want to modify
- Compare to your favourite hypothesis
- Drawbacks:
  - can only be done on already released data sets
  - will probably need a similar **effort** as if a CMS person or group would have done it

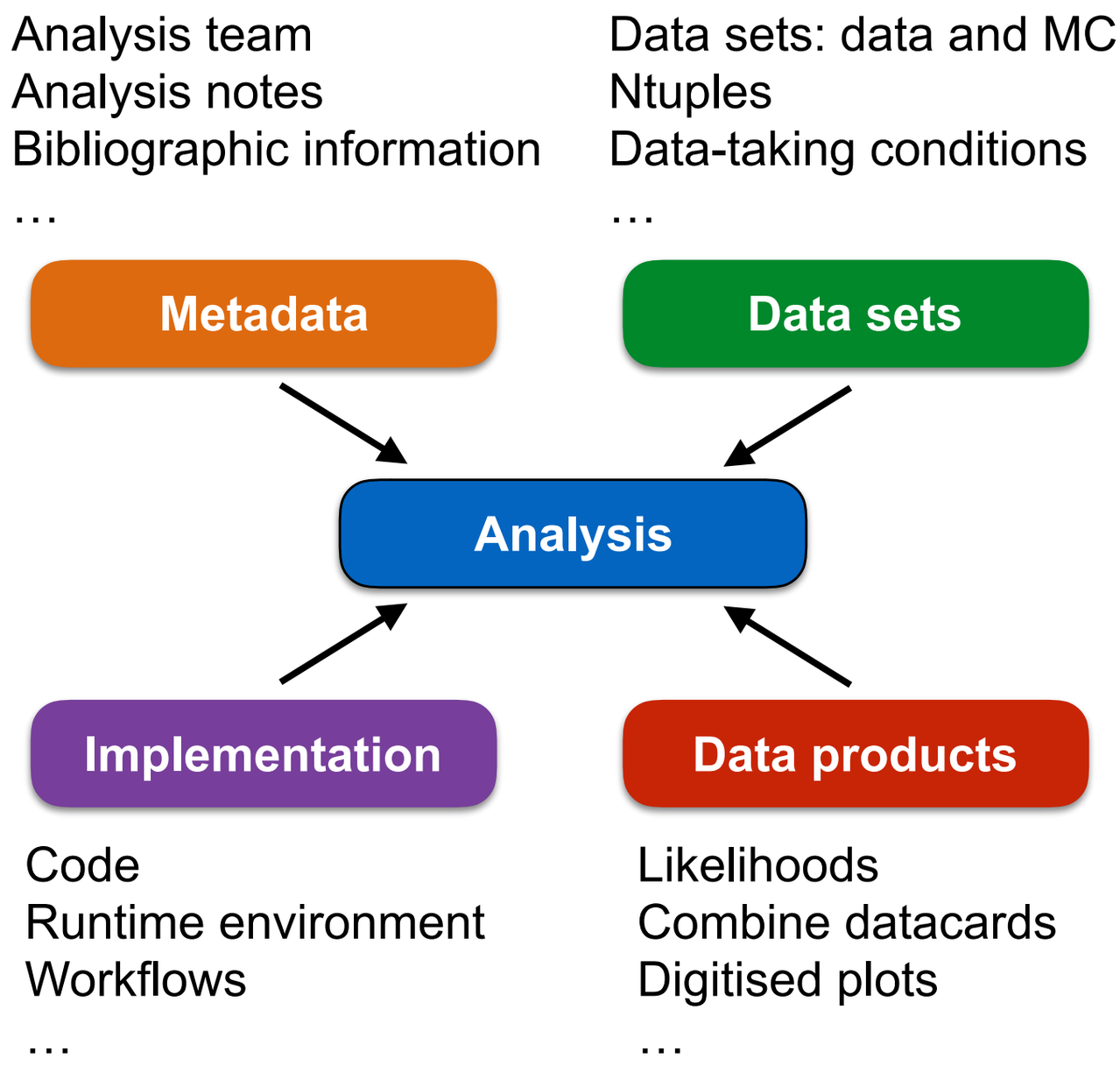
What is CMS Open Data NOT about



- Not a tool to browse existing published CMS results
  - use e.g. INSPIRE-HEP, arXiv, …
- Not a tool to (re)interpret published results by comparing with theory
  - use e.g. HEPData, Rivet, …
- Not a toolbox to recast published results into a different form
  - use recasting tools (see preceding and later contributions)
  - maybe in the future…?

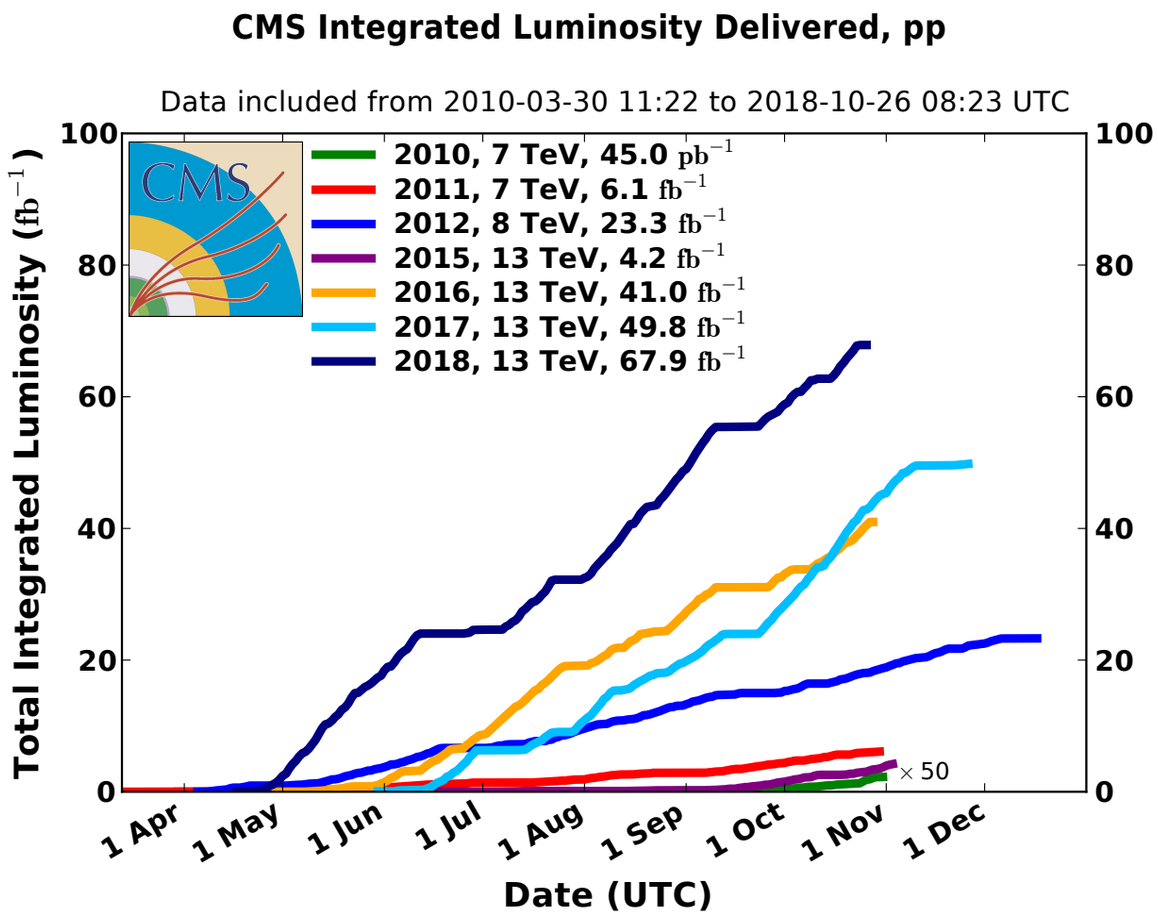
The vision

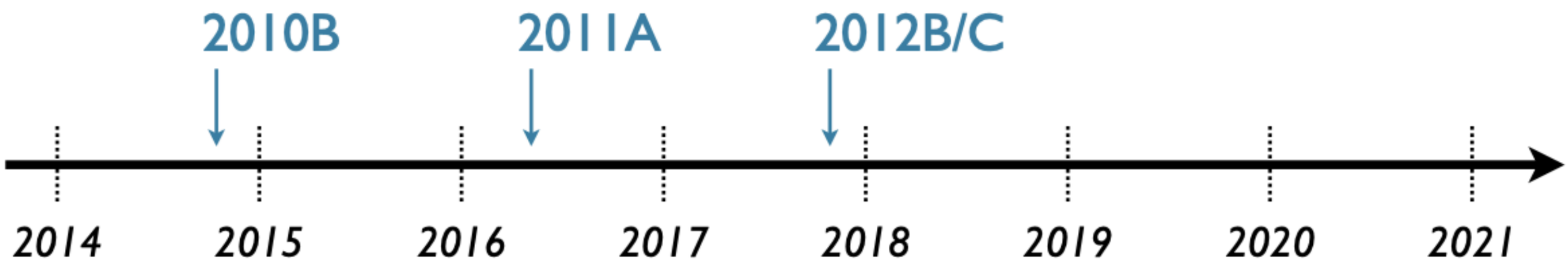
- **Preserve data and knowledge**
- **Open sharing** — data and knowledge more likely to survive if constantly used
- Make data available to school **pupils** and **researchers alike** — allow them e.g. to reconstruct the Higgs discovery
- Mine data to test new theories and provide crucial references



CMS Open Data

- **CMS publishes 50% of its collision data** three years after data taking
  - practically, ~five years after data taking
  - policy updated in 2018 (compare to other LHC experiments)
  - open data are released under the Creative Commons CC0 waiver
- **Up to 100% within ten years**
- Currently available:
  - 2010: 32 pb^{-1}
  - 2011: 2.3 fb^{-1},
  - 2012: 11.6 fb^{-1}





Information on Open Data

CERN Open Data Portal: http://opendata.cern.ch/about/CMS

CMS (DPHEP) **Open Data levels**:

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

higher computational effort

Challenges

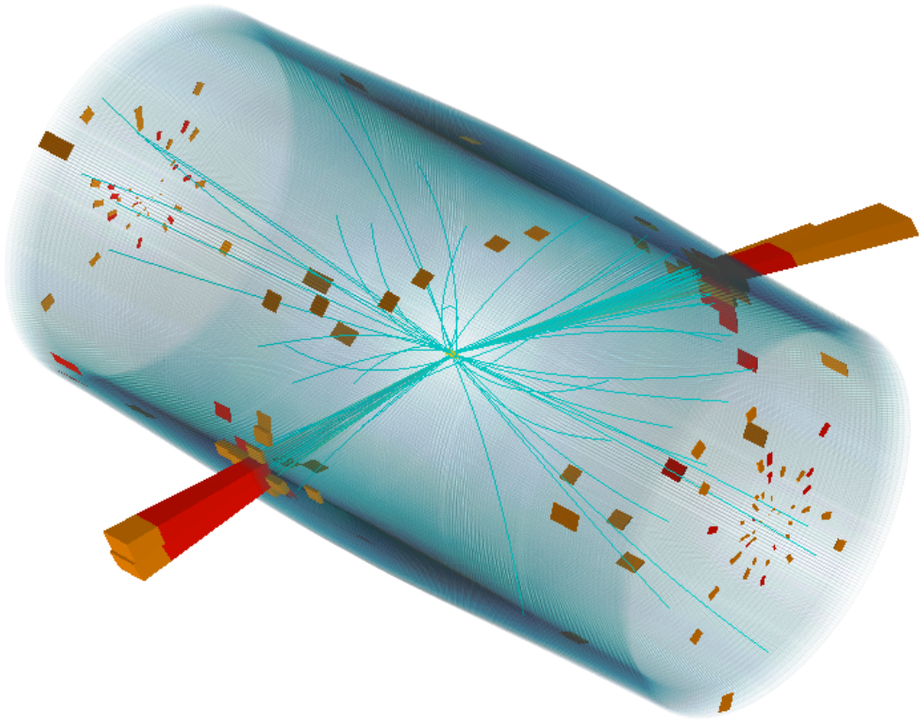
https://www.lojic.org/explore-local-data-through-lojics-new-online-metadata-search

- Need to preserve knowledge
- **immediate metadata**:
  - beam conditions, event and run numbers, provenance information (software versions, reconstruction chains)
- **context metadata**:
  - select correct objects, document them
  - apply further corrections, document them
  - information available at time of analysis, but often not preserved
- Need to collect all information and release it together with the data

Example why preserving a full CMS analysis is difficult

- Take e.g. a “simple” dijet analysis
- Select two jets and calculate dijet invariant mass
- **Jets are complicated**
  - need to be calibrated, need to preserve exact version of jet energy corrections
  - correct energy resolution according to measured differences in data and simulation
  - noise rejection cuts
  - pileup rejection strategies
  - trigger efficiencies
  - …
- **Inputs are from different groups** of people within the collaboration
- Preserve data sets and **exact versions** of reconstruction algorithms used
  - we re-reconstruct our data several times
  - older versions cannot be preserved (but the corresponding software), only latest-greatest versions are kept

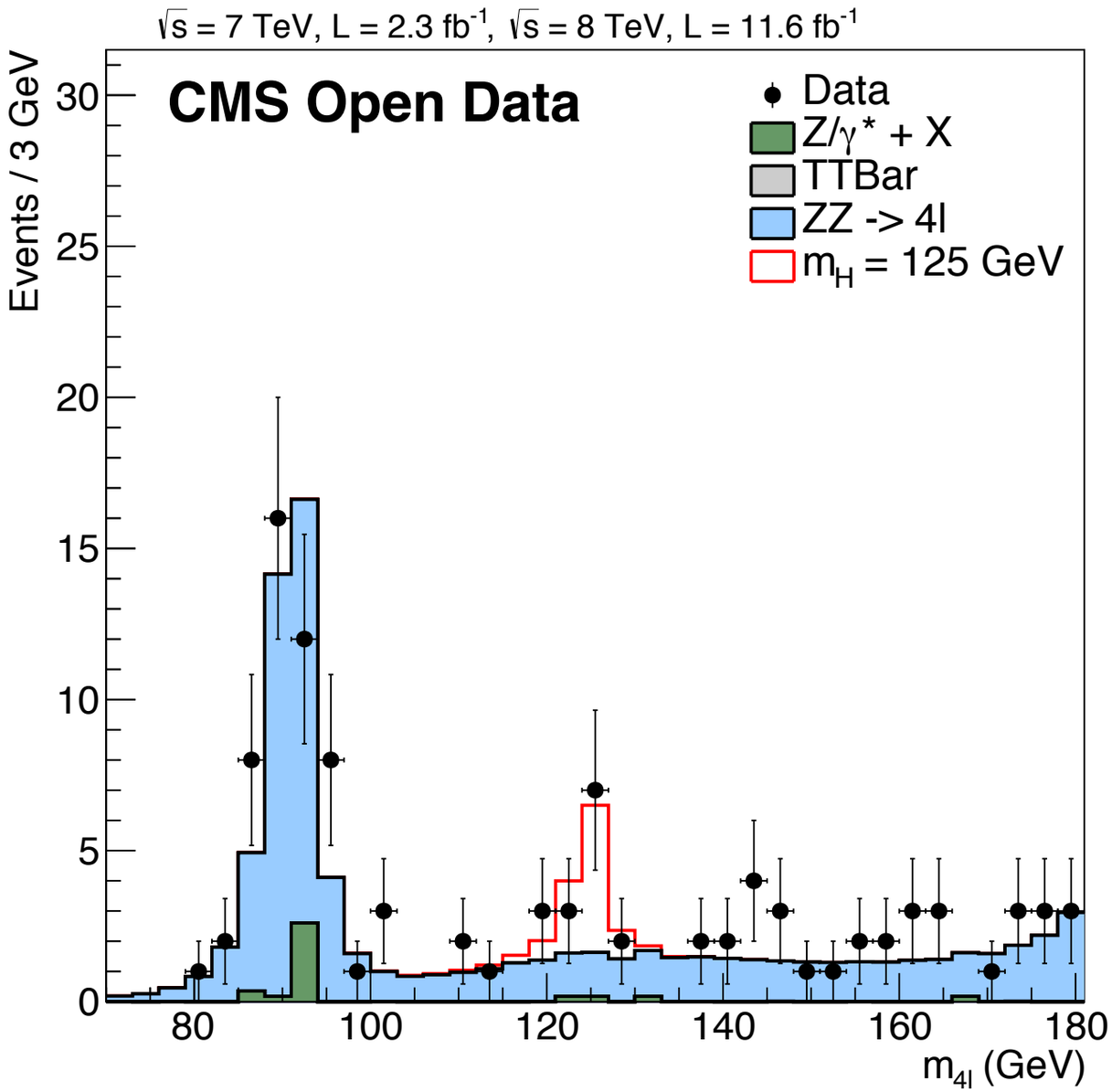
+ conditions of each subdetector
+ selection of “good runs”
+ luminosity
+ …



Can only approximate with reasonable computational effort

What's possible with CMS Open Data

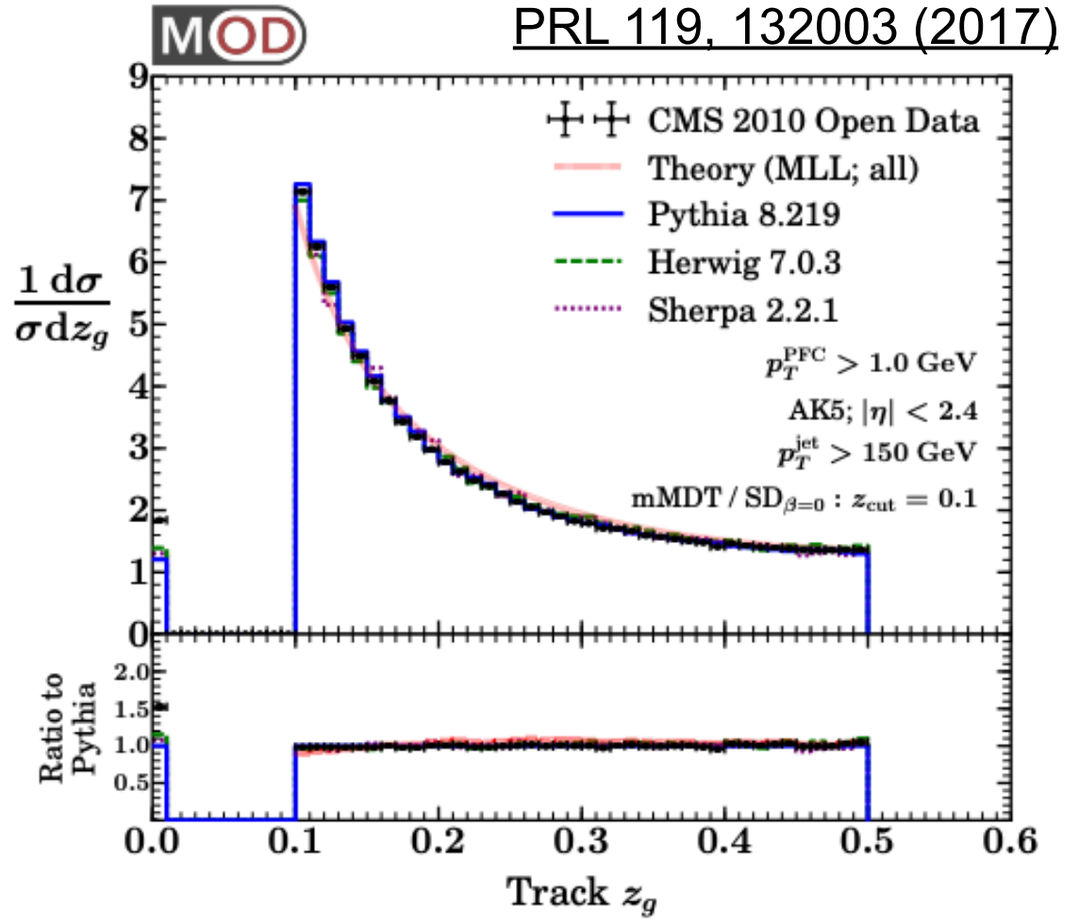
- You can **rediscover the Higgs** boson
  - see e.g. http://opendata.cern.ch/record/5500
- Different levels of computational complexity available
  - from reproducing the plot from pre-processed files
  - to processing ~80 TB of CMS AOD files in CMSSW
- Perform a full-fledged physics analysis!

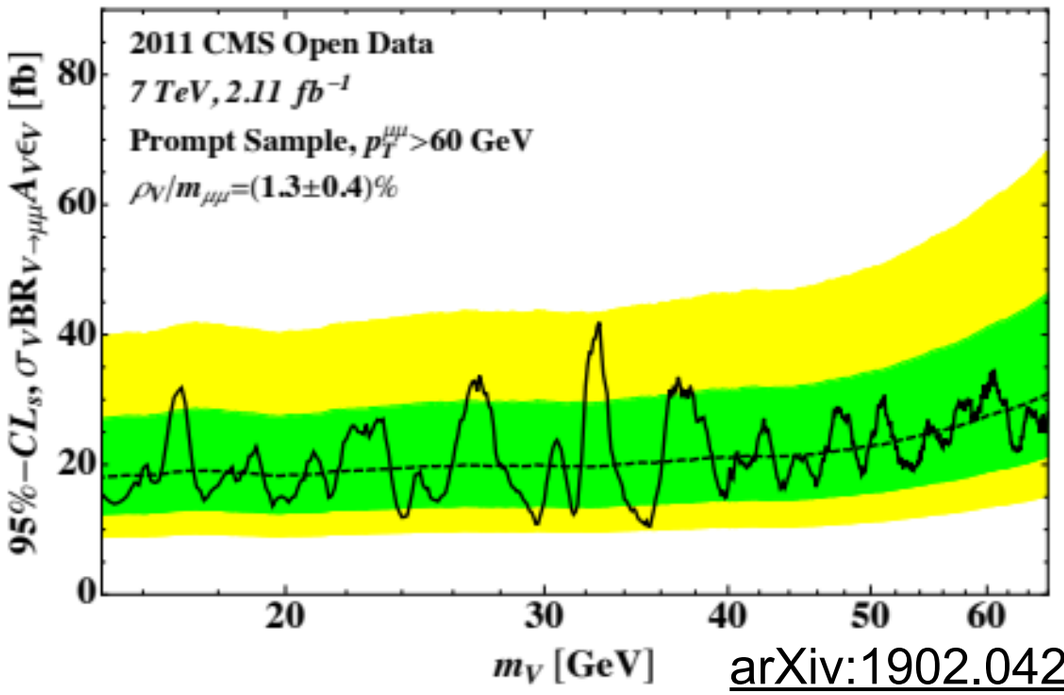


Use of CMS Open Data

- CMS Open Data have been used for physics publications!
  - see e.g. PRL 119, 132003 (2017)
  - and arXiv:1902.04222
- And also for physics education
  - see e.g. Particle Physics Playground, masterclasses, …
- Reaction to use of CMS Open Data within CMS are **uniquely positive**
- We are aware that there are groups that have started to study the CMS Open Data, but have given up because of difficulties
- Rather difficult to analyse the data without **CMS-knowledge**/expertise in **experimental HEP data analysis**
  - planning a **workshop aimed at theorists**

PRL 119, 132003 (2017)





arXiv:1902.04222

Feedback from use of CMS Open Data

- See Phys. Rev. D 96, 074003 (2017) and response at workshop in October 2017 + arXiv:1902.04222
- **Scattered information**: trying to improve Open Data web interface (with CERN scientific information service)
- **Lack of validation examples**: added several more examples, continuously adding more
- **Information overload**: working on simplifying the data formats
- **Presence of superfluous data**: adding documentation on how to filter data sets more efficiently
- **Corrections documented in publications not directly applicable**: see previous slides
- **Provenance information not always complete**: available for most analyses
- …

Bottom line: **simplify/facilitate use of Open Data**

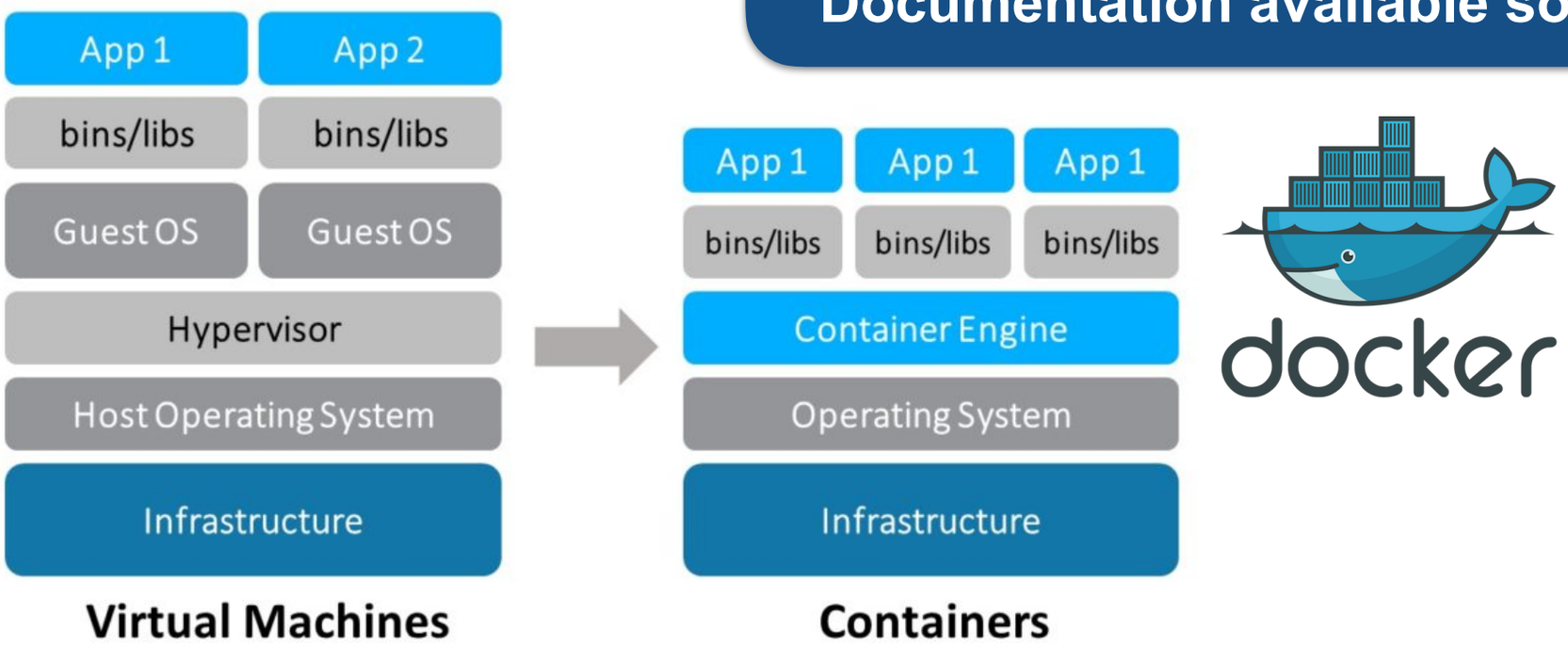
Simplify data formats

- Most of CMS open data users do not necessarily want to **learn CMSSW**
- "Get rid of CMSSW" as soon as possible
- ➜ Developing Physics Object Extractor code
- Improving documentation and tools to get scientific results from experimental data
  - Luminosity calculation
  - Experimental methods (tag & probe, MVA…)
  - MC generation
  - Improving/expanding trigger analysis examples
  - …
- Moving towards object-level formats: AOD ➜ MiniAOD/NanoAOD
  - examples will be provided

Software containers vs. virtual machines

- Previously provided on virtual machines with CMSSW installation
  - …and will continue to do so
- **New**: Docker software **containers**
  - allow to preserve full analysis
- Can run these containers on HPC-platforms
  - e.g. using Kubernetes **orchestration**
- Currently working on defining **workflows** (e.g. within **CERN REANA project**)
  - run physics analysis steps

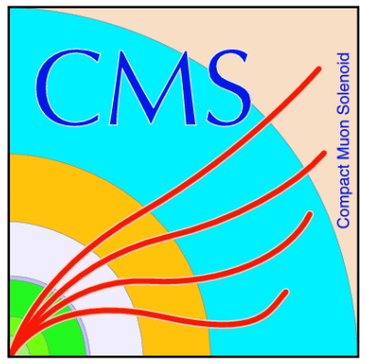
Documentation available soon!



Upcoming data releases

- Provide examples and data sets for use in **machine learning**
  - including their production workflows
- New documentation for **MC production**
  - enable production of custom MC for anyone with the required resources
- Publish **rest of 2010 collision data** and additional 2010 simulation
  - also improved metadata for all 2010 data
- Provide further **2012 simulation samples**
- Improved provenance information and new search functionalities for MC, ”on demand” MC
- First **CASTOR data** (with corresponding metadata and instructions)
- … and more to come later

Summary



- CMS is **leading the LHC Open Data effort**
- We are trying to facilitate the use of CMS Open Data
  - improved documentation and software tools + containers
  - simplified data formats
  - planning on organising a workshop aimed at theorists
- Release of new data imminent
- Please let us know if you have any feedback!


CERN