Karlsruhe Institute of Technology

CERN

Usage of SWAN with the CERN Open Data portal for education and outreach

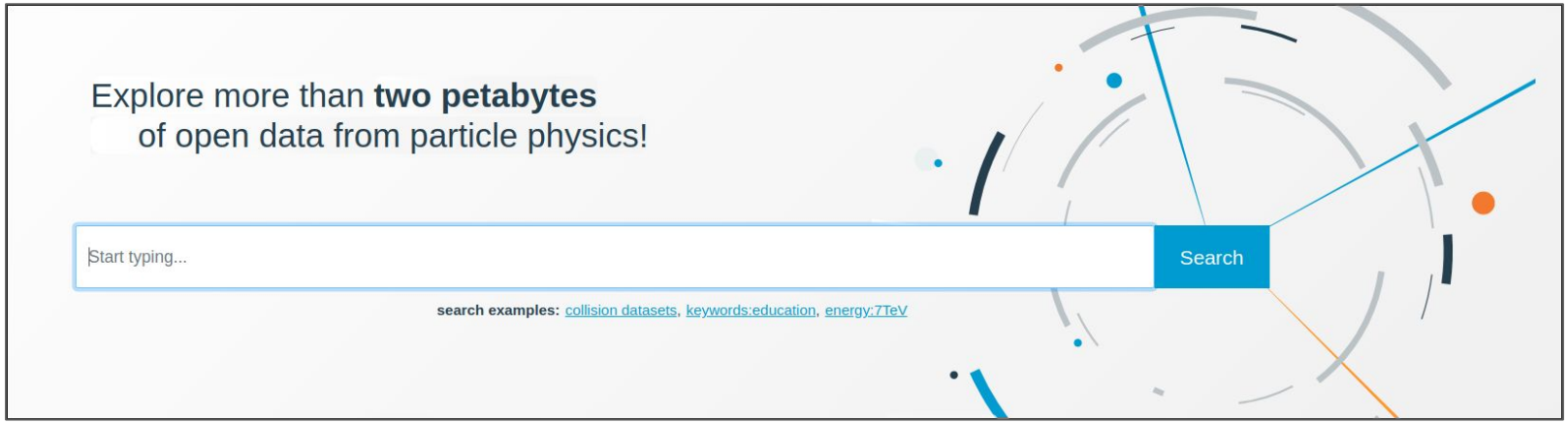
Stefan Wunsch
stefan.wunsch@cern.ch

CERN EP-SFT / KIT ETP

What is the CERN Open Data portal?

- Access point to a rapidly growing collection of data and other material originating from the research at CERN

- Well suited to preserve, document and publish information for newcomers in the HEP community and the general public



What the portal provides

- **Original datasets**
  From LHC experiments in AOD format with extensive documentation
- **Derived datasets**
  For specific analyses, examples, simplified datasets, ...
- **Software**
  Examples, analysis code, validation code, tools, ...
- General information and documentation

What people do with the resources

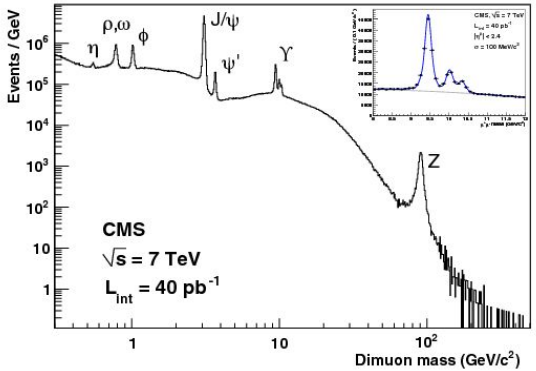
- **Research** resulting in publication in peer reviewed journals
  - Jet Substructure Studies with CMS Open Data
  - Fast and accurate simulation of particle detectors using generative adversarial networks
  - Searching in CMS Open Data for Dimuon Resonances with Substantial Transverse Momentum
  - ...

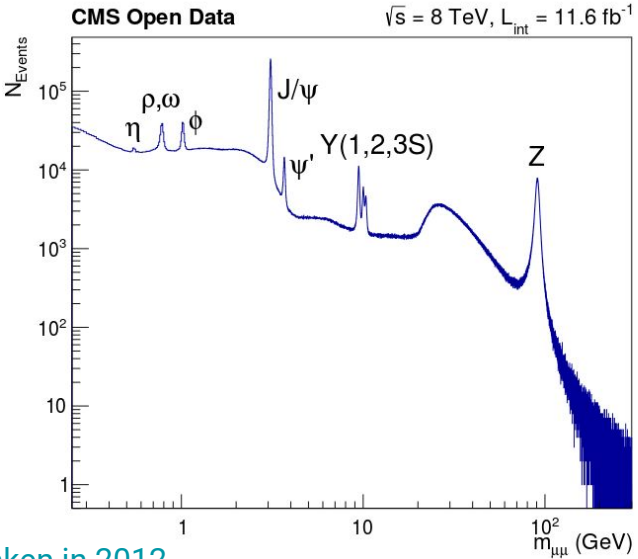
- **Education and outreach**
  - CMS education and outreach material
  - Machine learning tutorials
  - Training courses organized for teachers at Helsinki Institute of Physics
  - Software tutorials and workshops
  - Teaching & Data workshop at University of Florida
  - Projects for bachelor theses and summer students
  - ...

Example: Di-muon spectrum with CMS Open Data

- Rediscover particle resonances in a wide energy range up to the Z boson
- About 62 mio. events with four vectors of all muon candidates of data taken at CMS served as a 2.1 GB ROOT file via EOS and XRootD
- Analysis code in Python or C++ with less than 50 lines (less than 100 lines with plotting), only ROOT as dependency

Official CMS plot





Remake using CMS Open Data

Analysis of the di-muon spectrum using data from the CMS detector taken in 2012

How to get people to try the examples and learn about what we do at CERN?

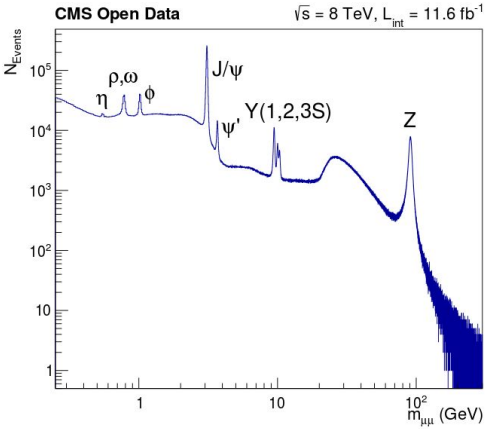
- **Key:** Keep the initial effort as low as possible

- **Installing software?** → To be avoided!
  - Different operating systems, technical know-how of the user, …
  - Complicated to support

- **Solution:** Jupyter notebooks
  - Minimal software requirements from the user (a web browser)
  - Proper user interface
  - Nice documentation using markdown
  - Interactive graphics, widgets, ...

Why is Binder not always sufficient?

- Binder works perfectly fine out-of-the-box for small examples
  - CMS education notebook reading reduced dataset in CSV format
- **New:** ROOT release on conda-forge with XRootD and opening XRootD port in Binder makes service now widely usable in HEP
- **Problem:** Di-muon example already on the edge for Binder's computing capacity
  - Fully IO bound
- **Software requirements**
  - C++ and Python notebooks
  - File transfer via XRootD
  - ROOT and (experiment) specific software
- **Hardware requirements**
  - Fast access to data on the EOS space of the CERN Open Data portal
  - Fast access to software on CVMFS



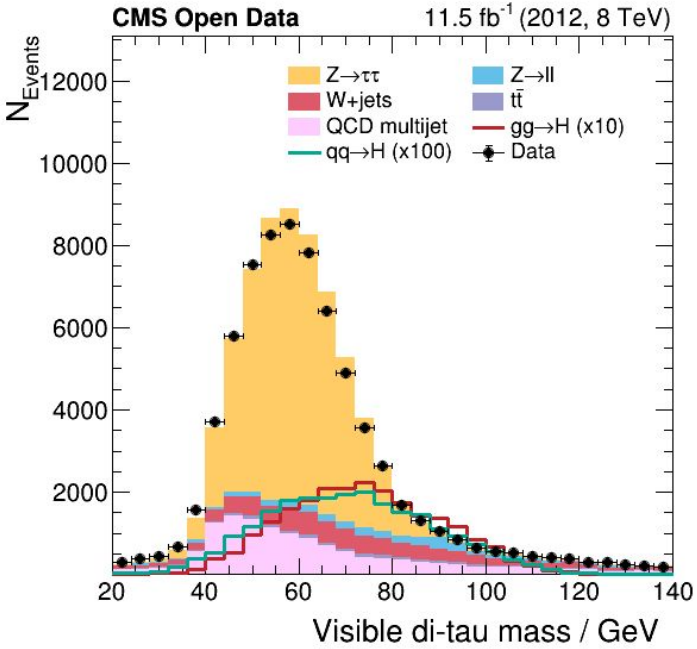
launch binder

Example: H→ττ analysis with CMS Open Data

- Providing ready-to-run examples for more realistic analyses is currently very challenging
  - HEP is inherently large scale
  - Binder-like services and internet connection of individuals provide only limited bandwidth

- Scale of an example minimal H→ττ analysis
  - “University-level” example
  - 9 derived datasets, in total about 60 GB
  - Stored in ROOT files / NanoAOD-like format
  - Computation mainly IO bound (skimming and making histograms)

- Runs smoothly on SWAN due to the existing software stacks (CVMFS and LCG stacks) and excellent connectivity to the CERN eco-system (EOS)



Proposal

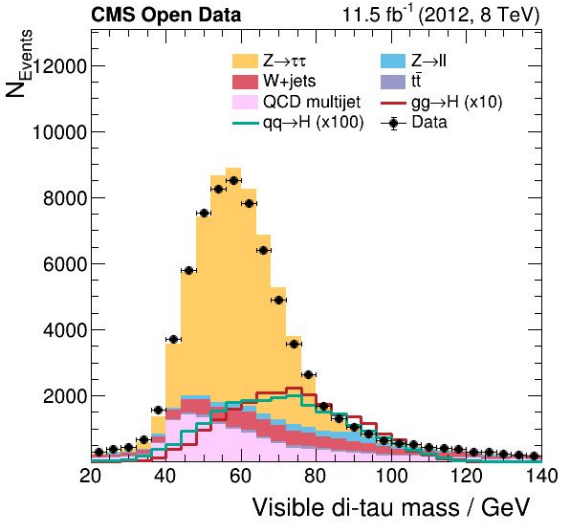
- Backing up the CERN Open Data portal with a service to run small to medium sized examples out-of-the-box
- Low latency for the user heavily influenced by the bandwidth to the data storage and software repositories
- Suitable authentication / accessibility model required
  - Allowing individuals to run whitelisted examples from the CERN Open Data portal?
  - Less restrictive or broader scoped authentication model similar to eduroam?

Outlook

- CERN Open Data portal is constantly growing
  - Data access policies of the ALICE, ATLAS, CMS and LHCb ensure vast amount of new data

- Ongoing effort to publish legacy CMS data in NanoAOD format
  - Reduced data format detached from experiment specific software
  - Suits a wide range of analyses
  - Allows for analyses with simple programming model
    → Bringing students and individuals close to real physics data from the LHC with minimal technical know-how
    → Example: ROOT RDataFrame

- **SWAN together with the Open Data portal would be the perfectly suited to bring HEP as close as possible to students and individuals**



Summary

- The CERN Open Data portal is the access point to a rapidly growing collection of data and other material originating from the research at CERN
- Especially for education and outreach Jupyter notebooks integrate very well with the resources provided by the portal
- Enabling the usage of the SWAN service together with the CERN Open Data portal for a broader audience has the potential to bring HEP as close as possible to students and individuals