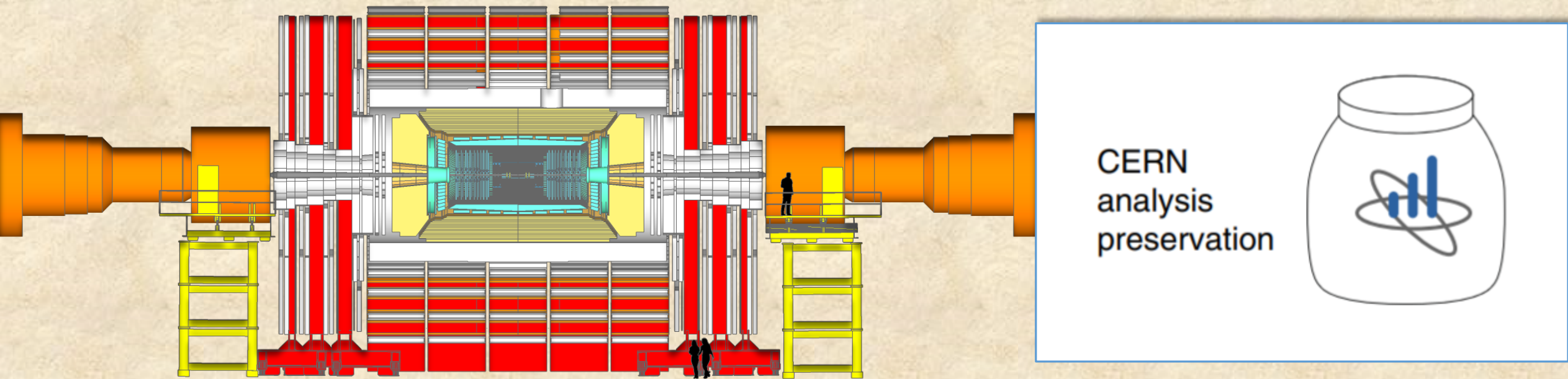


The CMS approach to Analysis Preservation



Lara Lloret Iglesias

Instituto de Física de Cantabria (CSIC-UC)

On behalf of the CMS Analysis Preservation team

Why?

The Analysis Preservation effort aims at responding to two parallel demands:

Inside of the Collaborations:

The high complexity of the analyses create **major challenges** in terms of capturing and preserving the analysis and the knowledge around it.

Externally:

Increasing number of funding agencies have put in place data management policies demanding the **development of comprehensive data management frameworks** for data and knowledge preservation, and for future reuse (or even reinterpretation and reproducibility) of research outcomes.



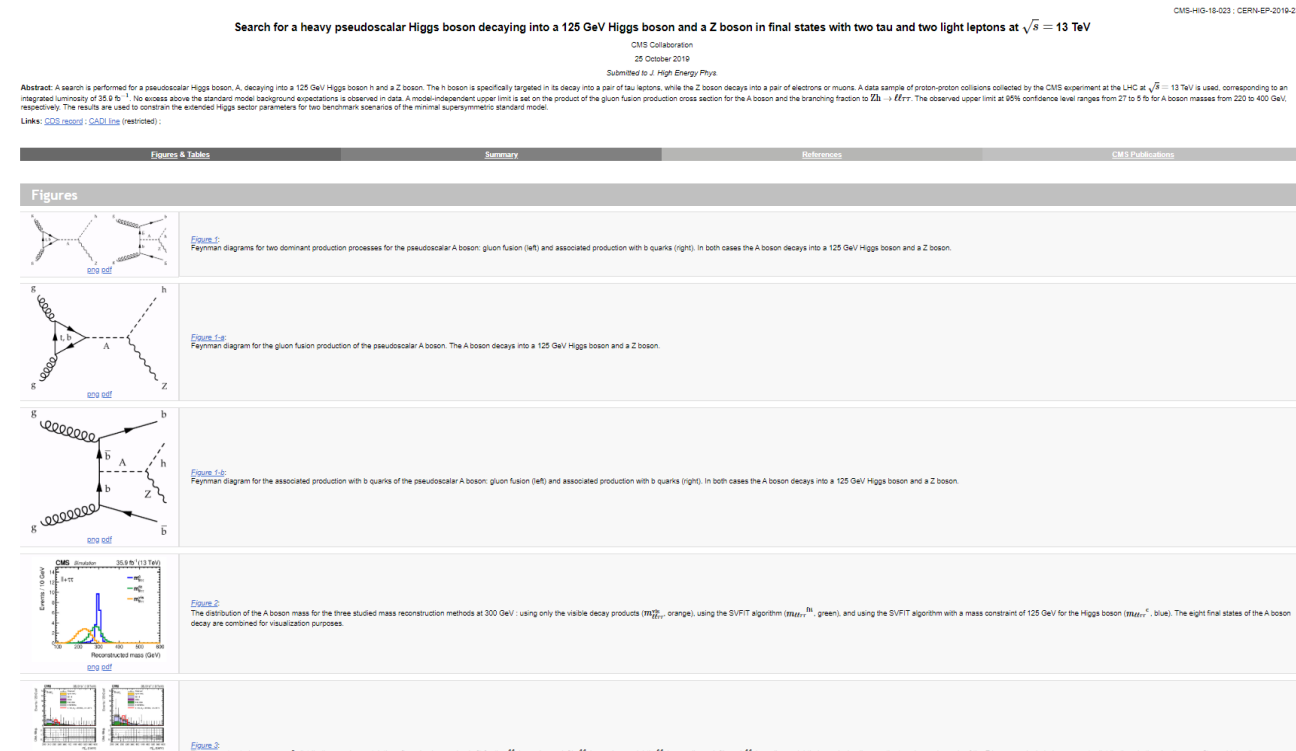
Reinterpretation effort at CMS

The CMS Collaboration is making an effort to make analyses reinterpretable

The central place for accessing the material is the [CMS public results webpage](#)

Each analysis has its own webpage on which one can find:

- All figures and tables that are part of the publication
- Additional figures and tables
- Links to the HEPData entry

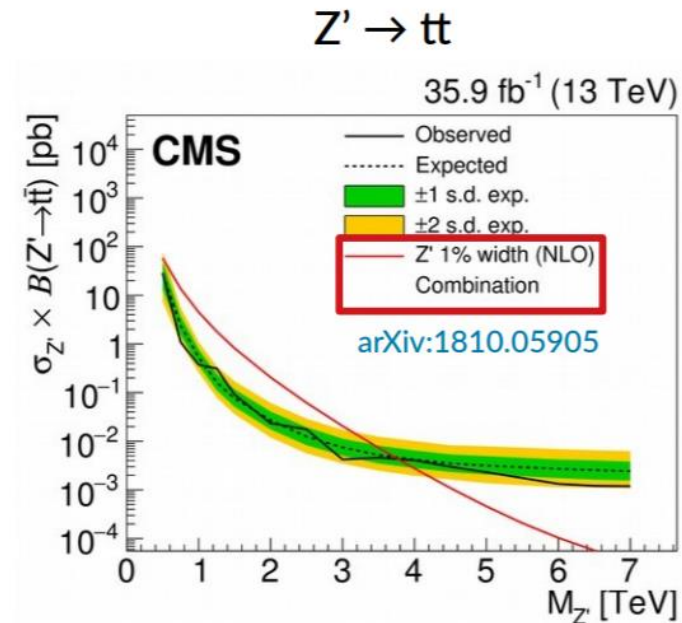
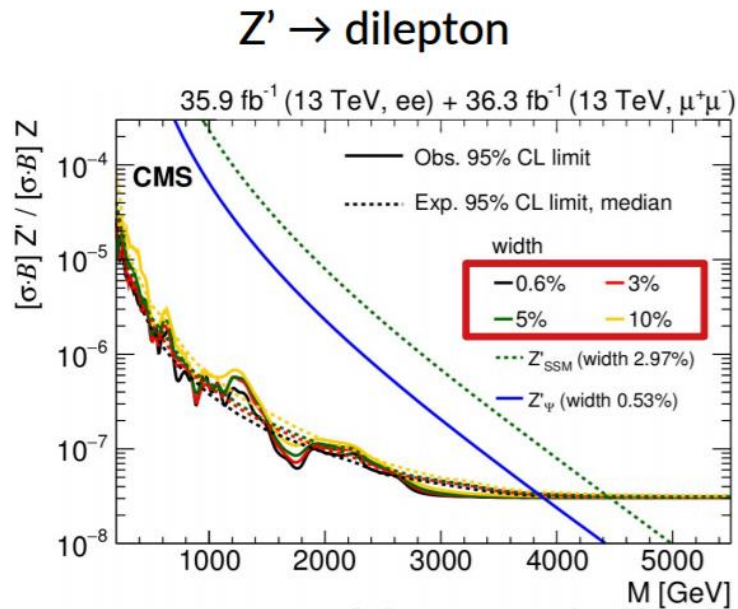


For more details check [Clemens Lange talk](#) at the Analysis Systems Topical Workshop



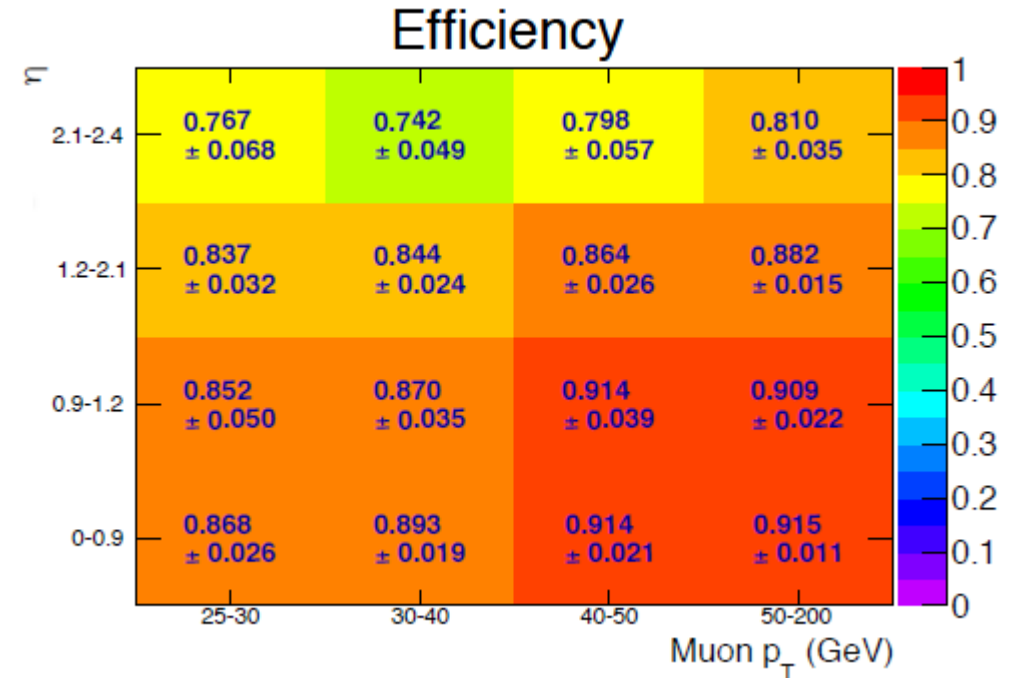
Providing information for reinterpretation

- Information provided in HEPData are mostly figures and tables → Easy part
- Often, results can be formulated generically (e.g. as cross section limit) by parametrizing a few free parameters
 - e.g resonance mass
- This gets more complicated when considering e.g. different resonant widths or production mechanisms



Existing materials

- For standard objects (such as leptons), selection efficiencies are provided:
 - Example: CMS SUSY object efficiencies
 - Could be used in Delphes-based analysis
- Can be used to some extent also for more complex objects (e.g. substructure)
- However, for uncommon objects that rely on specialised reconstruction (e.g. displaced vertices) things get tricky



Full reinterpretation workflow already tested for some analysis: i.e SUSY with H(bb)

See [Andreas Albert talk](#) for more details



In practice: Simplified likelihood

Statistical interpretation typically uses likelihood approach

Free parameters θ = independent physical uncertainty sources

$$\mathcal{L}(\mu, \theta) = \mathcal{P}(\text{data} | \mu \cdot s(\theta) + b(\theta)) \cdot p(\tilde{\theta} | \theta)$$

Simplified implementation

Free parameters θ = deviation from central value in each bin

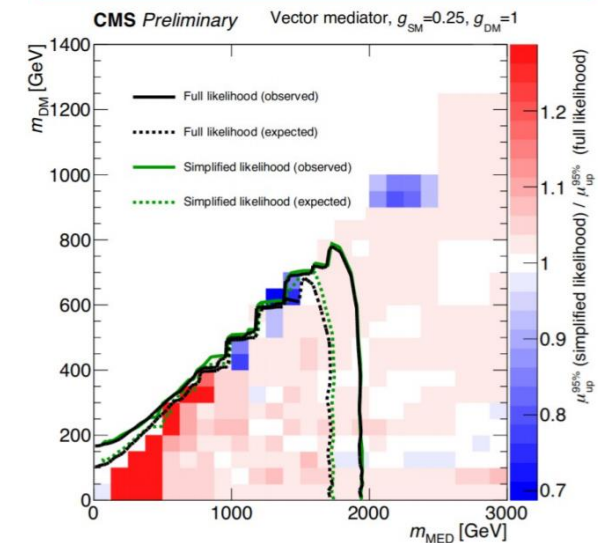
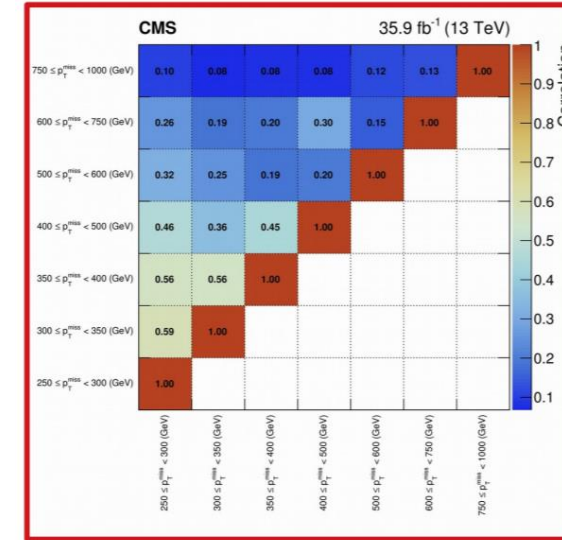
$$\mathcal{L}_S(\mu, \theta) = \prod_{i=1}^N \frac{(\mu \cdot s_i + b_i + \theta_i)^{n_i} e^{-(\mu \cdot s_i + b_i + \theta_i)}}{n_i!} \cdot \exp\left(-\frac{1}{2} \theta^T \mathbf{V}^{-1} \theta\right)$$

(bins)

Inputs for fully defined likelihood:

- You calculate signal yields s_i in each bin
- We provide:
 - BG yields b_i + uncertainties
 - data yields n_i
 - Covariance matrix \mathbf{V}

<https://cds.cern.ch/record/2242860?ln=en>



Preserved additional internal material

- Datasheets used for the signal extraction are preserved for a large number of analyses
 - Based on our internal statistical combination tool *Higgs Combine*
 - Including event distributions, templates and systematic variations
 - Allows reusing results for statistical combination and summary plots
- Furthermore, our internal analysis management tools stores:
 - Link to analysis twiki (mainly for review)
 - Link to (pre-)approval presentations
 - Analysis note(s) (i.e. internal documentation)
 - The actual Paper
 - Metadata (DOI, arXiv, HEPData, Rivet, CDS, data set)
- Could also think about an approach using an analysis description language
 - Pursued by some people within CMS

Ideally:

Having all this information accesible just from one place, indexed, searchable (final state, triggers, datasets...) → **CERN Analysis Preservation portal**



CERN Analysis Preservation Portal

<https://analysispreservation.cern.ch/>



The **CERN Analysis Preservation Portal (CAP)** is a service for the four LHC collaborations at CERN developed to address the need for the long-term preservation of all the digital assets and associated knowledge in the data analysis process, in order to enable future reproducibility of research results.

Further details on Pamfilo's talk



CERN Analysis Preservation Portal





CERN Analysis Preservation ^{BETA}

- Run by CERN Scientific Information Services + help from experiments
- Current status:
 - Not yet in use for ongoing analyses: still in beta phase → Will be proposed to wide usage for the current analyses as soon as all the interface/features are fully tested and ready
 - Being populated with “old” and private production analyses within the DPOA group
- What is already in place?
 - Imports information from CMS analysis management system (CADI)
 - Allows to add/edit information through APIs/CLIs
 - Storing information in addition to what can be directly imported from CADI:
 - ✓ Input data sets (linked to CMS data set search)
 - ✓ Triggers
 - ✓ JSON files (data quality)
 - (Full) Workflows
 - ✓ Additional documentation
 - ✓ Details on statistical treatment (CMS statistics committee questionnaire, data cards...)



How?

CERN Analysis Preservation ^{BETA} Search  Create 

- ALICE Analysis
- Statistics Questionnaire
- CMS Analysis
- Test schema.
- ATLAS Analysis
- LHCb Analysis

PUBLISHED IN COLLABORATION	SHARED WITH YOU	LATEST FROM YOU
B2G-16-007	EXO-12-011	B2G-16-007
B2G-16-006	EXO-12-010	B2G-16-006
B2G-16-005	EXO-12-009	B2G-16-005
B2G-16-004	EXO-12-007	B2G-16-004
...

Vanilla mode

Full reproducibility mode please turn this mode on if you want to capture additional information about main and auxiliary measurements, systematic uncertainties, background estimates, final state particles

Basic Information
Please provide some information relevant for all parts of the Analysis here

Information from CADI database
Automatically taken from CADI, based on CADI ID

Input Data
Please list all datasets and triggers relevant for your analysis here

N-tuples Production [0 items]
Provide details on the intermediate n-tuples production

Additional Resources
Add any useful additional documentation on the analysis

Statistical Treatment

Full reproducibility mode

Full reproducibility mode please turn this mode on if you want to capture additional information about main and auxiliary measurements, systematic uncertainties, background estimates, final state particles

Basic Information
Please provide some information relevant for all parts of the Analysis here

Information from CADI database
Automatically taken from CADI, based on CADI ID

Input Data
Please list all datasets and triggers relevant for your analysis here

N-tuples Production [0 items]
Provide details on the intermediate n-tuples production

Auxiliary Measurements [0 items]
Provide details on auxiliary measurements used in the analysis

Background Estimation [0 items]
Details on the background estimation methods

Final Results
Please provide information necessary to generate final plots and tables for your analysis.

Main Measurements Workflows [0 items]
Please provide information about the main measurements of your analysis

Systematic Uncertainties [0 items]
Details on the systematic uncertainties

Additional Resources
Add any useful additional documentation on the analysis

Statistical Treatment

Two different analysis preservation modes:

- **Vanilla mode:** Basic assets preservation
- **Full reproducibility mode:** Great level of detail → Aiming for reuse

Autocomplete

Input Data

Please list all datasets and triggers relevant for your analysis here



Primary Datasets +

Monte Carlo Signal Datasets +

Monte Carlo Background Datasets +

Official JSON files +


Search interface showing a search bar with a magnifying glass icon and a close button (X). Below the search bar, a dropdown menu displays a list of file paths starting with `/DoubleEG/CMSSW_7_4_10_patch2-74X_dataRun2_HLT_frozen_v1_resub_RelVal_doubEG2015C-v1/`. The paths include `MINIAOD`, `RAW`, `RECO`, `DQMIO`, and `ALCARECO`. Below the list, there is a "Triggers" section with a plus sign (+). At the bottom, there are "OK" and "Remove" buttons.

Indexed - Searchable

Elasticsearch cluster for indexing and information retrieval needs

Signal Event Selection

- Physics Objects +
- Vetos +
- Processing Steps +

+ 

Systematic Uncertainties [0 items]
Details on the systematic uncertainties

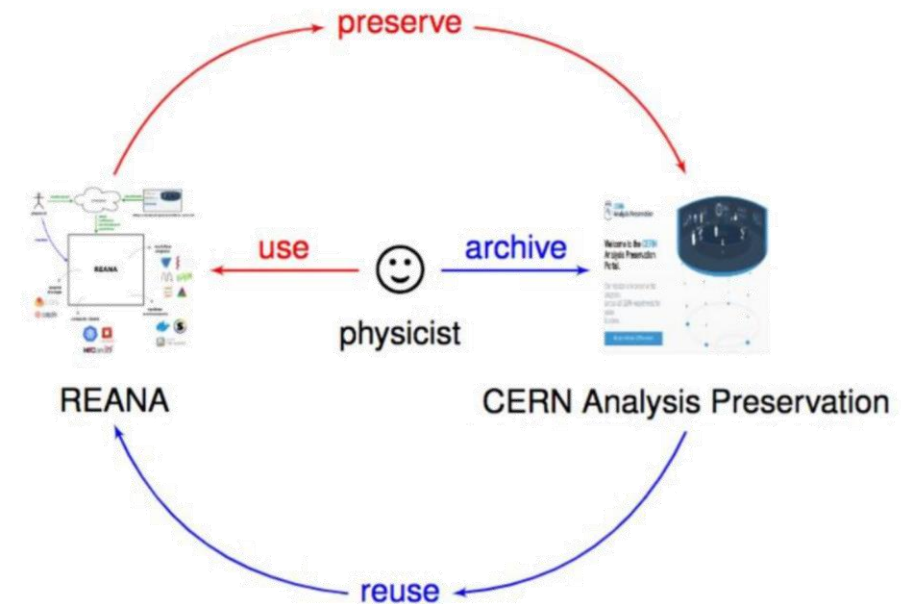
16 results Page 1 of 2

STATUS	TYPE	PHYSICS_OBJECTS	RESULTS
<input type="checkbox"/> draft 16	<input type="checkbox"/> cms-analysis-v0.0.1 16	<input type="checkbox"/> jet 22 <input checked="" type="checkbox"/> muon 16 <input type="checkbox"/> PFMuon 10 <input type="checkbox"/> GlobalMuon 4 <input type="checkbox"/> TrackerMuon 4 <input type="checkbox"/> electron 10 <input type="checkbox"/> photon 6 <input type="checkbox"/> MET 2 <input type="checkbox"/> tau 2 <input type="checkbox"/> track 2 <input type="checkbox"/> vertex 2	JME-10-004 MUON We present the results of a visual scan of high E T events (E T > 60 GeV DR p E T > 60 GeV) in a large inclusive sample of 7 TeV pp collision data, after applying the official noise clean-up available in CMS SW 3.7.0 patch2. The scan is performed separately for events with E T > 60 GeV and p E T > 60 GeV since two different noise clearing algorithms are employed. The CMS software Fireworks and PFRooTEvent have been used to produce the event displays. The high E T events have been visually inspected and classified in different categories. The results of this scan can provide hints to further improve the noise clearing and to identify possible problems and inconsistencies in the algorithms employed in CMS for the E T reconstruction.
			FWD-10-005 MUON First measurement is reported of the exclusive two-photon production of muon pairs, pp → γγ μ+ μ-, in proton-proton collisions at √s = 7 TeV. For the muon pairs with invariant mass above 11.5 GeV/c and with p T (μ) > 4 GeV/c and η(μ) < 2.5, 148 candidates are found in the CMS data sample of 40 pb-1. The characteristic distributions of the muon pairs produced via γγ fusion as of the muon acoplanarity and of the pair invariant mass and transverse momentum, are well described by the full Monte Carlo simulation using the LPAIR event generator. Small and well understood background to the process is observed, and it is shown that pp → γγ μ+ μ- provides a reliable absolute normalization of the LHC luminosity.
			AN-2011/103 ELECTRON MUON Previous measurements in ep and hadron-hadron colliders demonstrated that events with large rapidity gaps (LRG) can be described by diffractive interactions. A quantitative interpretation of LRG events at hadron-hadron colliders is complicated by the fact that the LRG signal is destroyed or diminished by multiparton interactions. In this paper we study the correlations of the energy flow in the forward detectors and the track multiplicity in the central detector using events with centrally produced W and Z bosons, identified with their leptonic decays. The analysis uses the entire 7 TeV pp collision high quality data sample, about 36pb-1, recorded during the 2010 LHC operation and strict conditions for single pp vertex events. The observed forward energy deposits, their correlations and the track multiplicities in the central part of the CMS experiment are compared with different Monte Carlo
			AN-2011/062 MUON Yields of prompt and non-prompt B/s, as well as Υ(1S) mesons, are measured by the CMS experiment via their μ+ μ- decays in PbPb and pp collisions at √NN = 2.76 TeV for quarkonium rapidity y < 2.4. Differential cross sections and nuclear modification factors are reported as functions of y and transverse momentum pT, as well as collision centrality. For prompt B/s with relatively high pT (6.5 < pT < 30 GeV/c), a strong, centrality-dependent suppression is observed in PbPb collisions, compared to the yield in pp collisions scaled by the number of inelastic nucleon-nucleon collisions. In the same kinematic range, a suppression of non-prompt B/s, which is sensitive to the in-medium b-quark energy loss, is measured for the first time. Also the low-pT Υ(1S) mesons are suppressed in PbPb collisions.
			AN-2011/103 ELECTRON MUON Previous measurements in ep and hadron-hadron colliders demonstrated that events with large rapidity gaps (LRG) can be described by diffractive interactions. A quantitative interpretation of LRG events at hadron-hadron colliders is complicated by the fact that the LRG signal is destroyed or diminished by multiparton interactions. In this paper we study the correlations of the energy flow in the forward detectors and the track multiplicity in the central detector using events with centrally produced W and Z bosons, identified with their leptonic decays. The analysis uses the entire 7 TeV pp collision high quality data sample, about 36pb-1, recorded during the 2010 LHC operation and strict conditions for single pp vertex events. The observed forward energy deposits, their correlations and the track multiplicities in the central part of the CMS experiment are compared with different Monte Carlo
			AN-2010/411 ELECTRON MUON MET This note describes the search for the Higgs boson h in the H → WW → νν decay channel in about 35.5 pb-1 of pp collision data at √s = 7 TeV collected by the CMS detector at the LHC. Event yields are presented along with background predictions, expected signal yields, and the associated uncertainties for a sequential and a multivariate analyses. No excess above the Standard Model predictions is found in the current data sample and limits on the Higgs boson production cross-section times branching ratio are derived. With the current amount of data, the observed limits have no sensitivity to the SM Higgs boson, but compared to the recent theoretical calculations performed in the context the Standard Model with a four fermion generation, allow for excluding the Higgs boson with a mass in the 144-207 GeV range at 95% confidence level.

REANA

<http://www.reanahub.io/>

- REANA is a Reproducible research data analysis platform
- Based on docker containers
- Encapsulates the analysis environment, software and workflow in order to run transparently for the user
- Goal:
 - Integrate it with the CAP service
 - Working on dedicated templates



Some more info

- Four people working (as service work) filling CAP → 2015/2016 analyses
- CMSSW is containerised → Any analysis using mostly CMSSW should be quick to containerise
- Working examples:

[reana-demo-cms-h4l](#)

REANA example - CMS Higgs-to-four-leptons analysis

● C++ 🍷 15 ★ 1 🚫 0 🐛 0 Updated 12 days ago

[reana-demo-cms-reco](#)

REANA example - CMS reconstruction

● C++ 🍷 MIT 🍷 6 ★ 0 🚫 4 🐛 0 Updated on 20 Sep

[reana-demo-cms-dimuon-mass-spectrum](#)

REANA example - CMS dimuon mass spectrum

● C++ 🍷 MIT 🍷 4 ★ 0 🚫 0 🐛 0 Updated on 27 Sep

- Working on a jet energy correction workflow
- If you have suggestions for a particular analysis to be containerised/reanified talk to us!

Conclusions

- The **CMS Collaboration is making an effort to preserve analyses** and make them reinterpretable
- **Several tools exist** already:
 - Public webpages (including additional material, efficiencies, simplified likelihoods)
 - HEPData (Rivet)
 - CERN Analysis Preservation Portal
 - Open Data Portal/effort
- The **next big step will be preserving the implementation** (CAP-Reana integration)
 - Software containers make this easier
 - Current big challenge is to find a practical workflow preservation approach

Thanks for your attention

