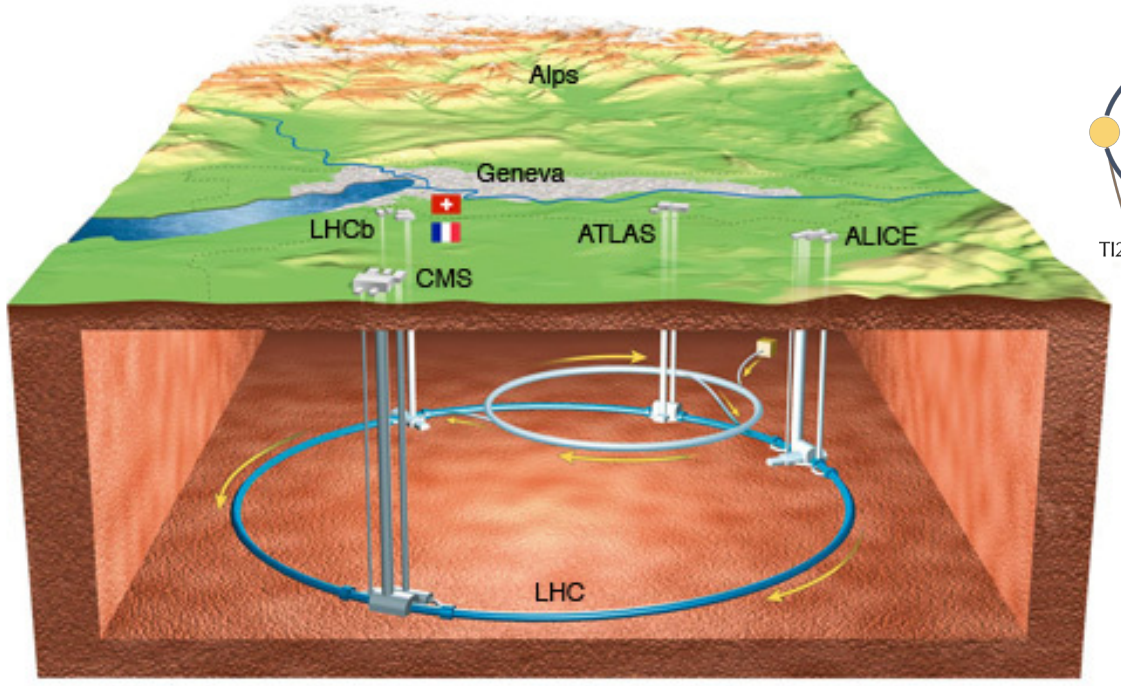
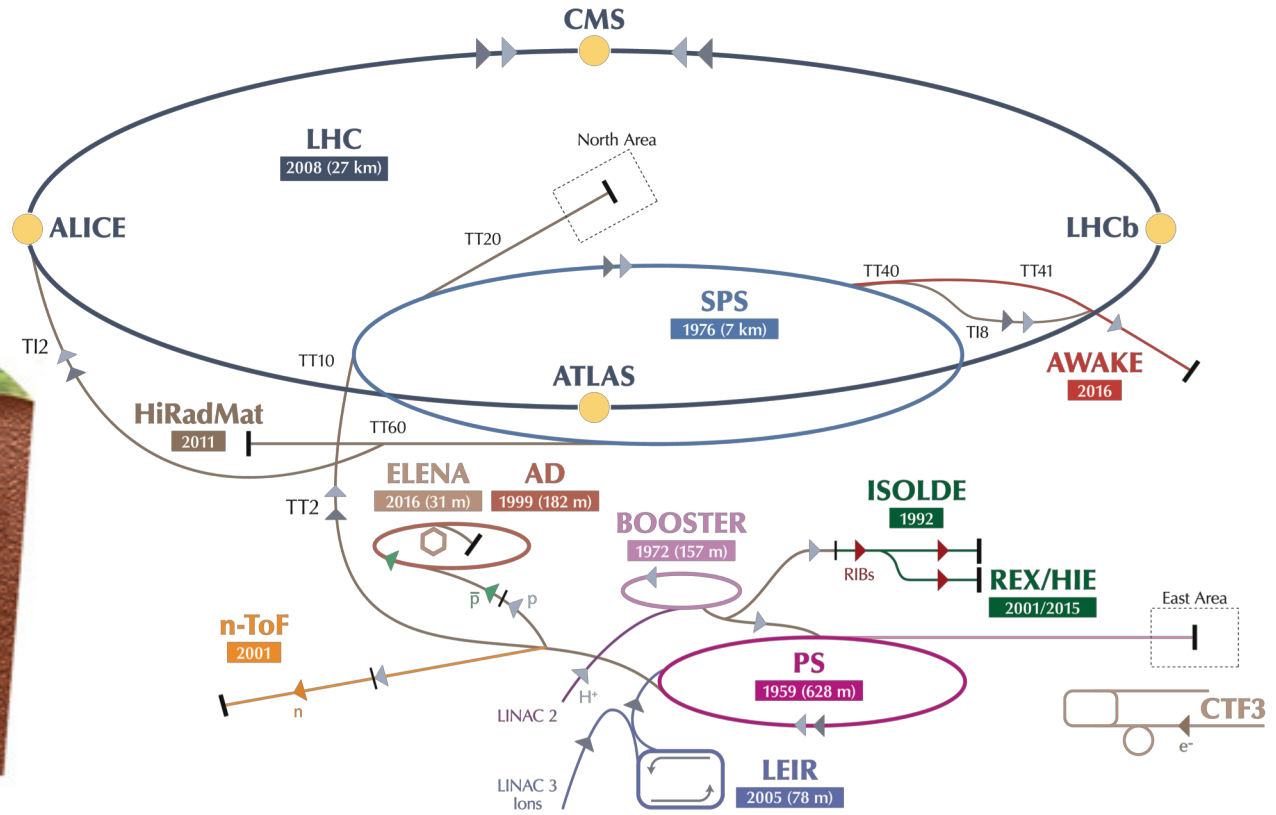
On-Demand Distributed Workflows for Physics Analysis at the CMS Experiment

Diyaselis M. Delgado López

Department of Physics
UPRM CMS Research Group

CERN | European Organization for Nuclear Research





Large Hadron Collider is the main 27-kilometre particle accelerator ring at **CERN.**

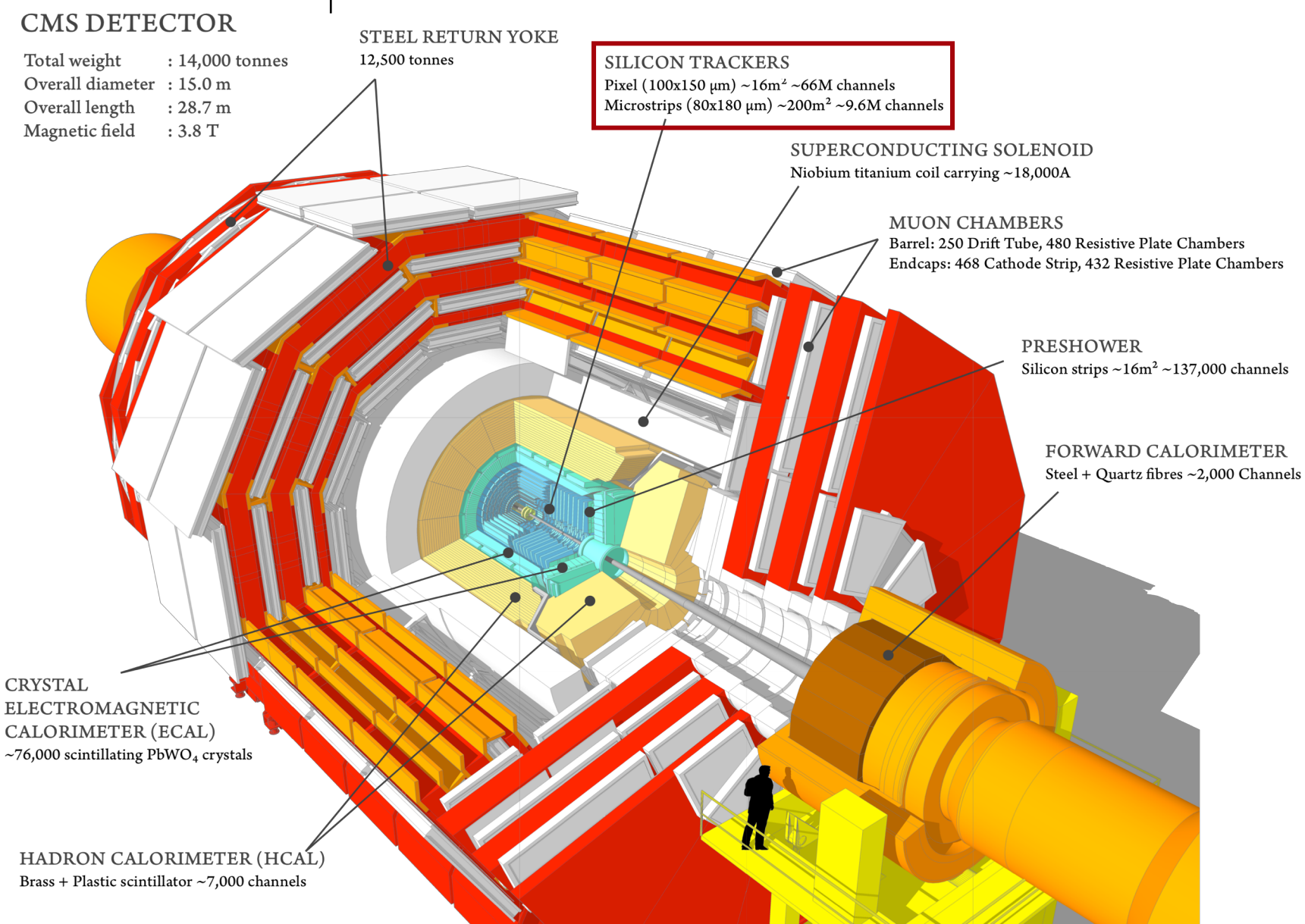
p (protons) ions RIBs (Radioactive Ion Beams) n (neutrons) $\bar{p}$ (antiprotons) e^- (electrons) proton/antiproton conversion proton/RIB conversion

LHC Large Hadron Collider SPS Super Proton Synchrotron PS Proton Synchrotron AD Antiproton Decelerator CTF3 Clic Test Facility

AWAKE Advanced WAKefield Experiment ISOLDE Isotope Separator OnLine REX/HIE Radioactive EXperiment/High Intensity and Energy ISOLDE

LEIR Low Energy Ion Ring LINAC LINear ACcelerator n-ToF Neutrons Time Of Flight HiRadMat High-Radiation to Materials

CMS | Compact Muon Solenoid



CMS is a general purpose detector at the LHC.

Why open data access is needed?

- **Use cases:** Analysis within the collaboration, education, outreach, analysis by external users.
  - **Data:** 158 TB of raw and legacy reconstructed data (AOD).
  - **Software:** CMSSW open source through VM image (CERNVM), which builds the appropriate environment from CVFMS, available in CODP.

Easy access to old data for collaboration members and external users that has educational and scientific value, a societal impact.

DPOA | Data Preservation and Open Access

DPOA

opendata CERN

- is the access point to data produced in research done at CERN.

reana

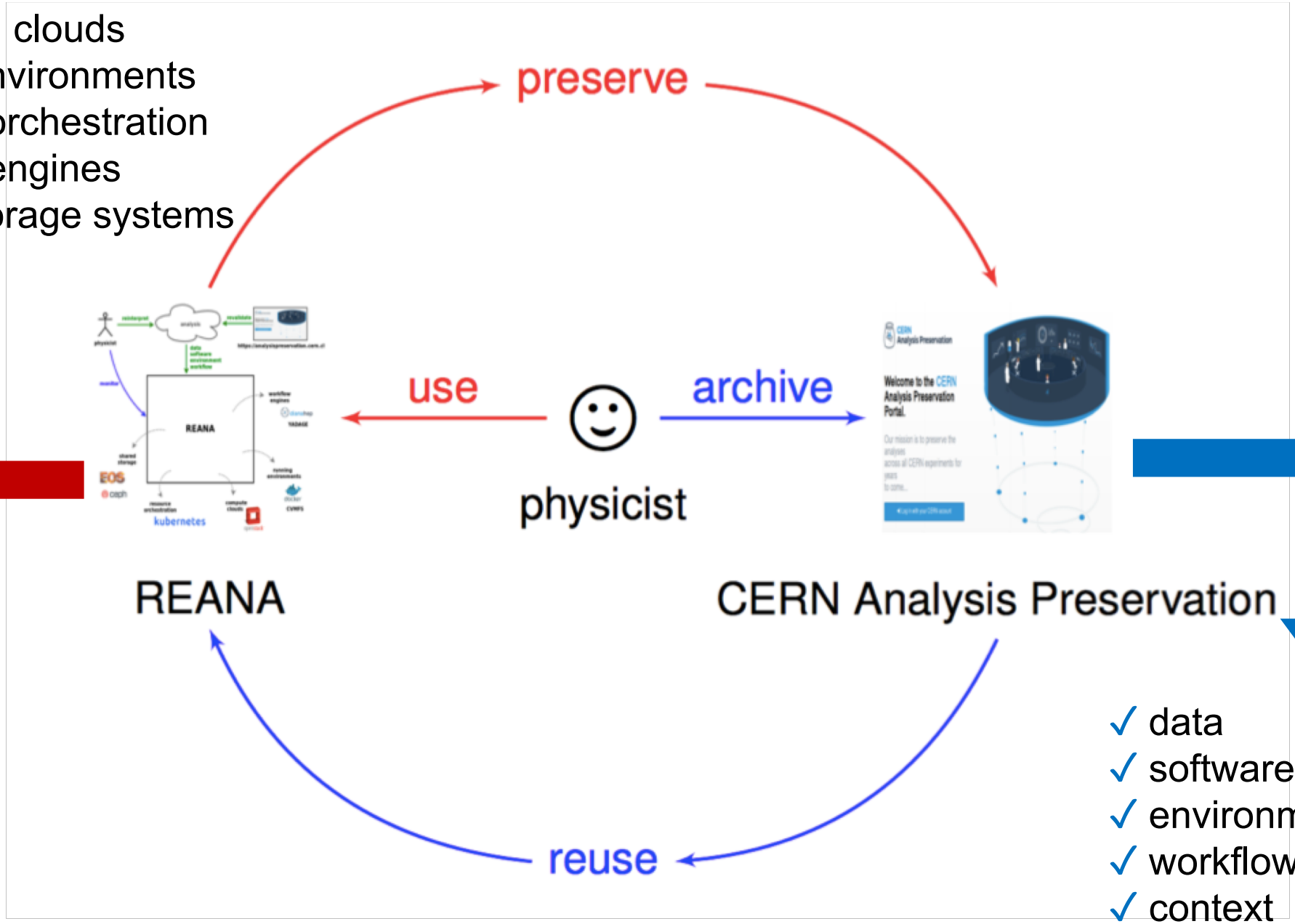
- Reusable and reproducible research data analysis platform.

CERN ANALYSIS PRESERVATION

- Addresses the need for the long-term preservation of the data analysis process.

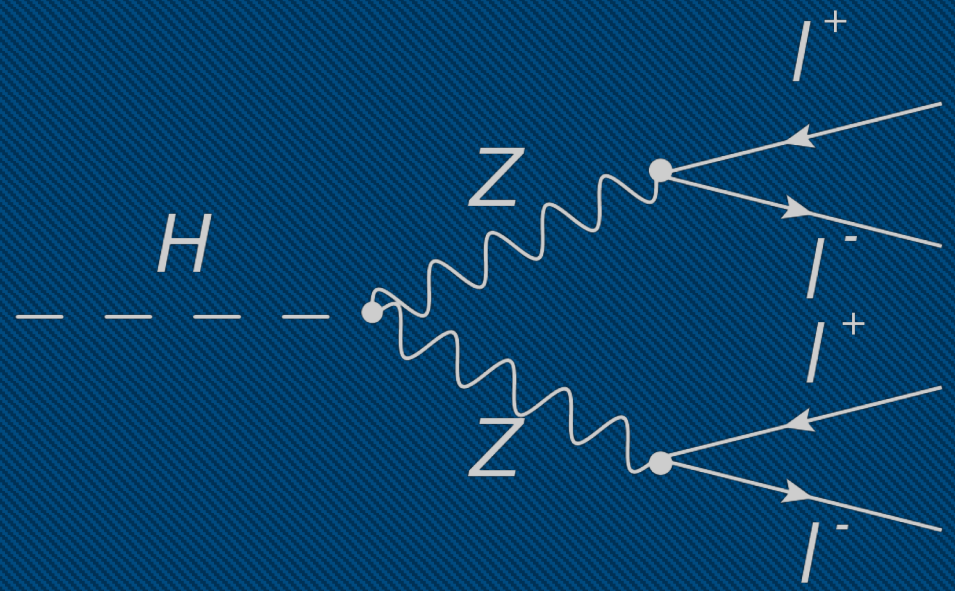
CASE STUDIES IN REPRESENTATION OF CMS

✓ computing clouds
✓ running environments
✓ resource orchestration
✓ workflow engines
✓ shared storage systems



✓ data
✓ software
✓ environment
✓ workflow
✓ context
✓ documentation

opendata
CERN

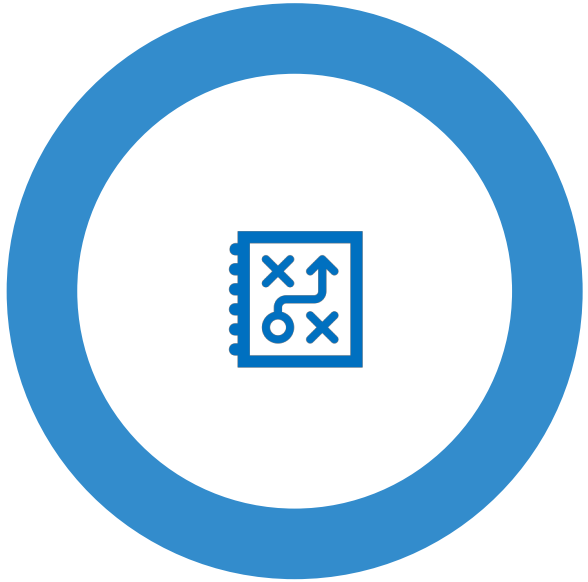


Higgs-to-four-lepton Analysis using 2011-2012 data

Preservation and Reproducibility of the Analysis

CMS Open Data:
Inputs and CMS software environment

CAP:
Analysis structure and metadata



ReANA:
Analysis workflow, commands, and output

Structure the Analysis

1 Input data

What is your input data?

- input files
- input parameters

2 Analysis code

Which code analyses it?

- software frameworks
- user code

3 Compute environment

What is your environment?

- operating system
- database calls

4 Analysis workflow

Which steps did you take?

- single command
- complex workflows

Structure the Analysis

1 Input data

2011 – 2012
RAW data and
MC simulations

2 Analysis code

Plot background
and processed data

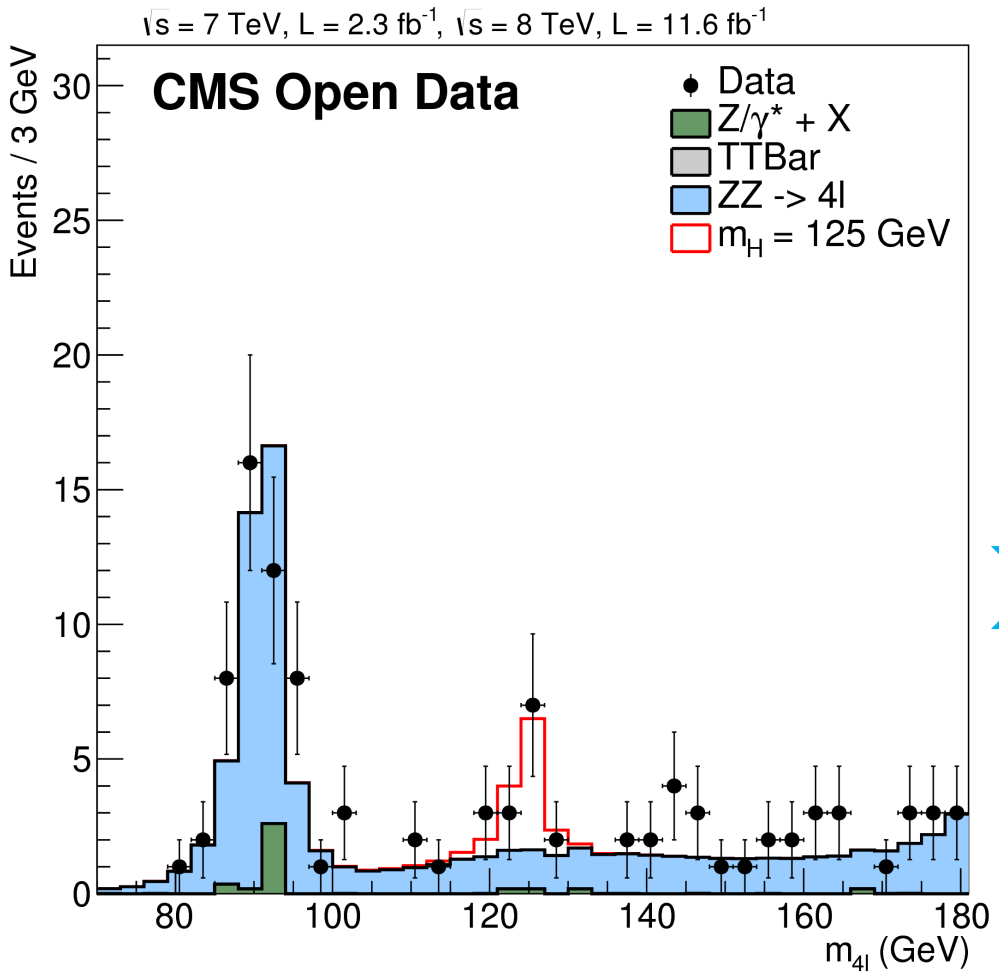
3 Compute environment

CMSSW image
from Docker
containers

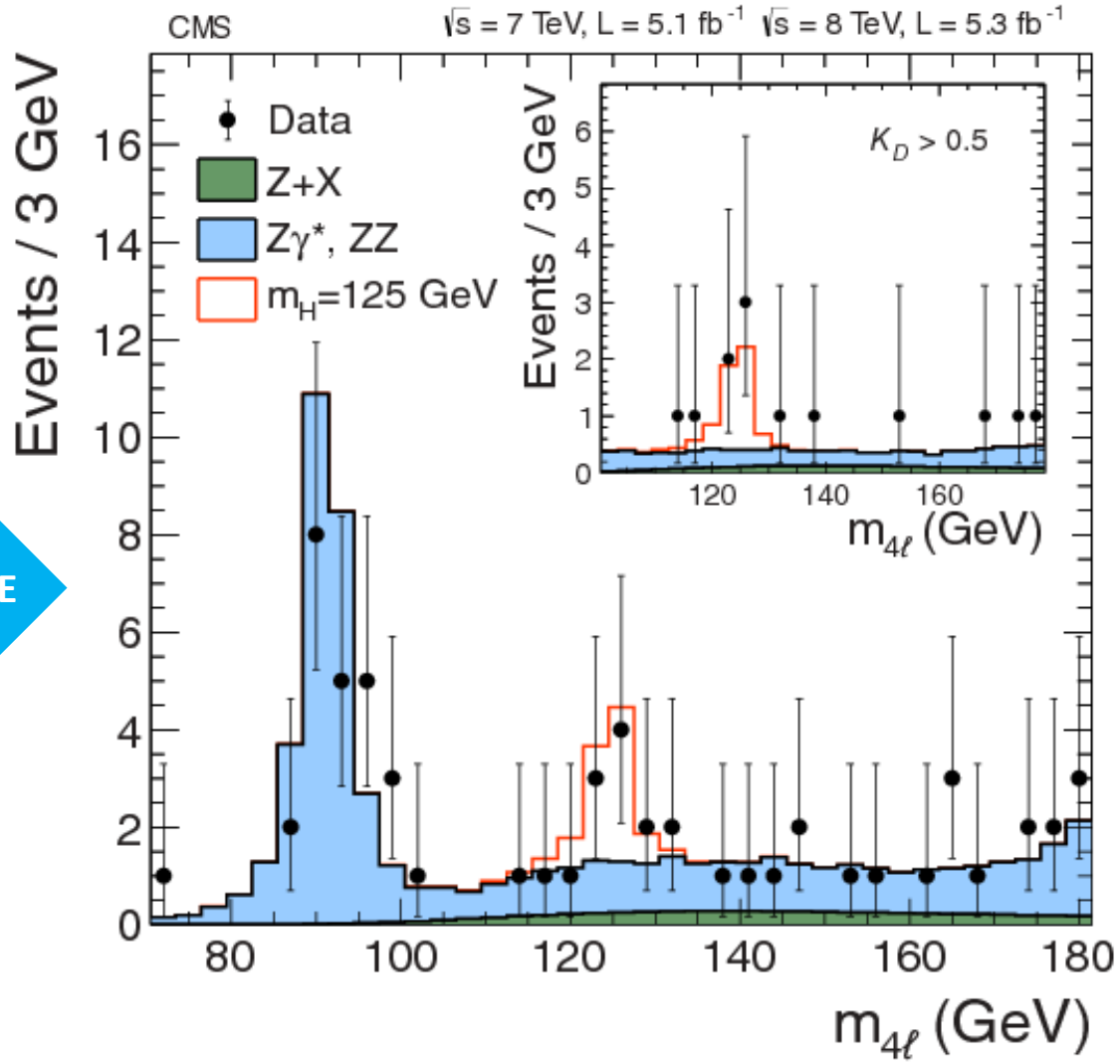
4 Analysis workflow

CWL (Common
Workflow Language)

Higgs discovery: simplified reimplementation



COMPARE



opendata
CERN

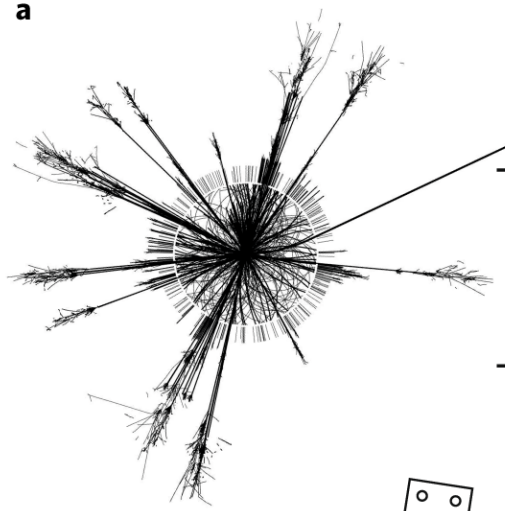
Reprocessing AOD from 2010 -2012 RAW samples for Machine Learning

RAW

RECO

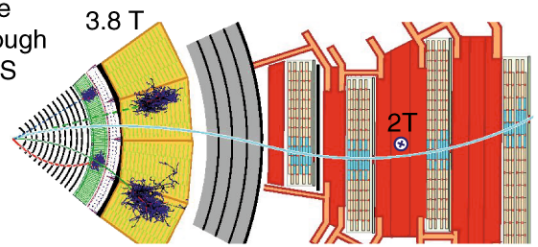
AOD

a



Collision data processing

Transverse slice through CMS



Electromagnetic calorimeter

Monte Carlo simulations

Data storage

b

Physics analysis

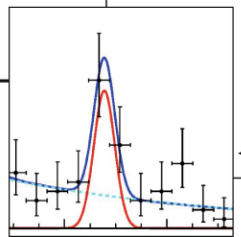
Create signal model

Apply data selection

Create background model

Fit model to data

Physics results



c

CERN analysis preservation

d

opendata CERN

CERN Open Data

Publications

arXiv.org

CERN Open Data Portal Validation

- The objective is to compare the outputs of reprocessed AOD files for 2010-2012 RAW samples with CMS Open Data VM results.

- ReANA, reconstructs the data files and processes to deliver an output plot.

2010 DATASETS

/MinimumBias/Run2010B-v1/RAW
/Electron /Run2010B-v1/RAW
/Mu /Run2010B-v1/RAW
/Jet /Run2010B-v1/RAW

2011 DATASETS

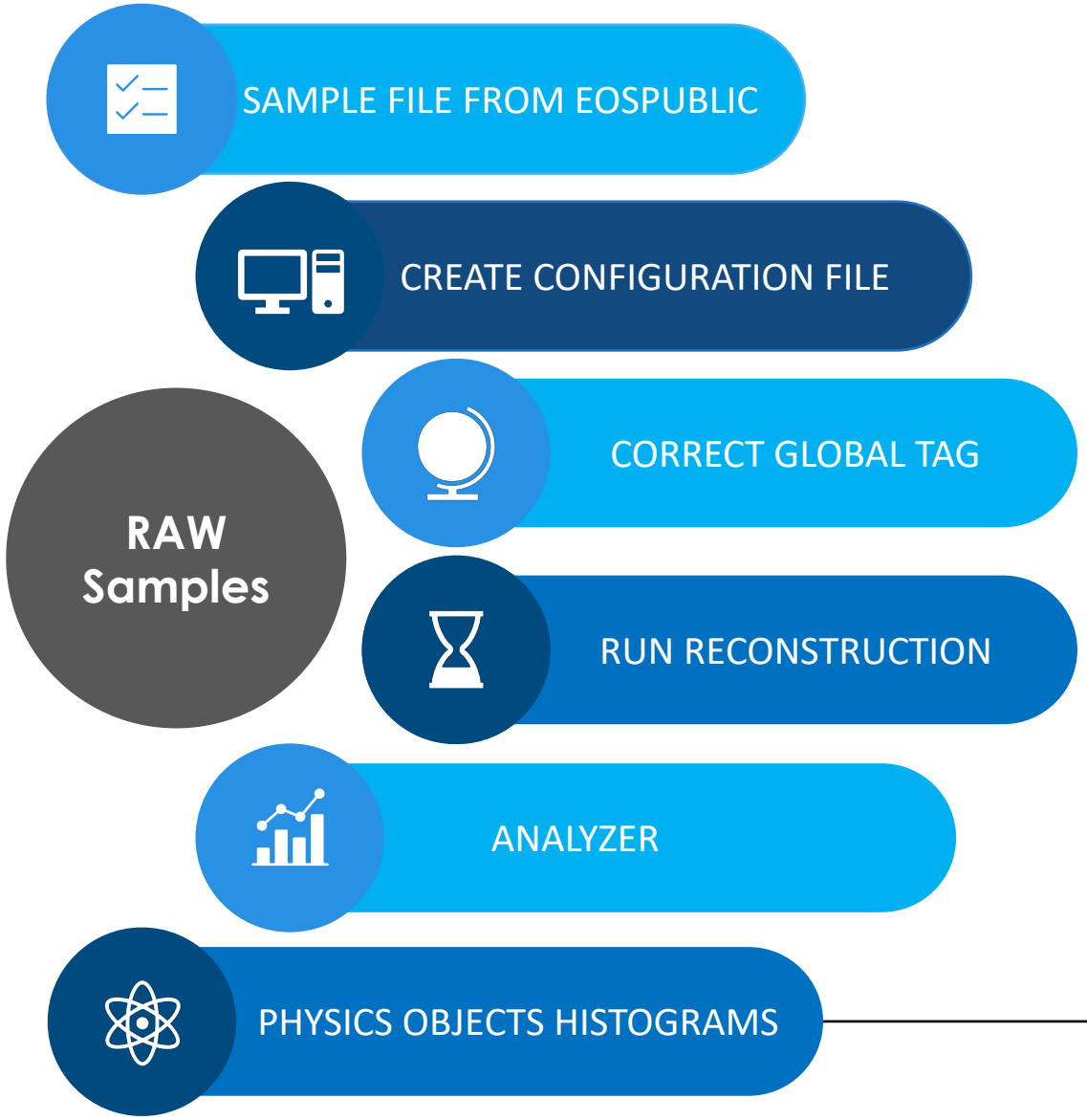
/DoubleElectron/Run2011A-v1/RAW
/SingleElectron /Run2011A-v1/RAW
/DoubleMu /Run2011A-v1/RAW
/SingleMu /Run2011A-v1/RAW
/Jet /Run2011A-v1/RAW

2012 DATASETS

/DoubleElectron/Run2012B-v1/RAW
/SingleElectron /Run2012B-v1/RAW
/DoubleMuParked/Run2012B-v1/RAW
/SingleMu /Run2012B-v1/RAW
/JetHT/Run2012B-v1/RAW

*RELEASED TO OPEN DATA PORTAL THIS YEAR

Data Reconstruction Process through REANA



Output of Analysis

Reprocessed output ROOT file

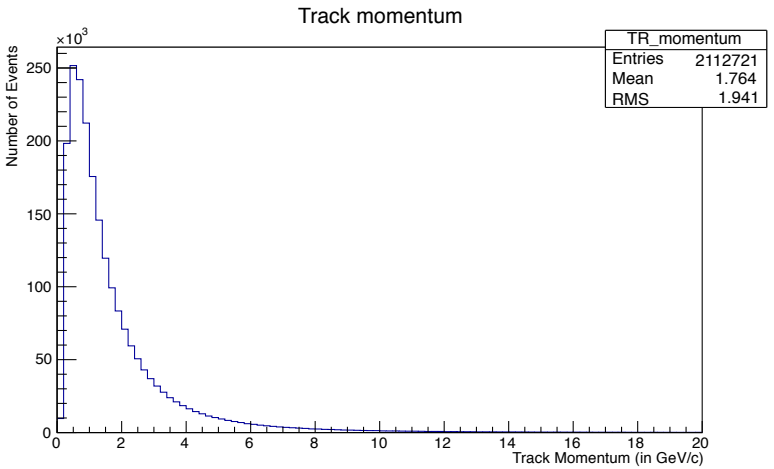
- AOD data format
- Sizeable (~2 GB)

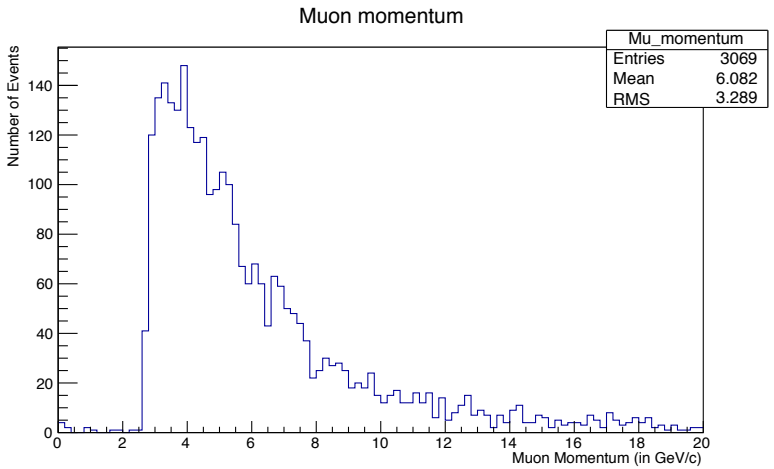
Physics Objects Histograms

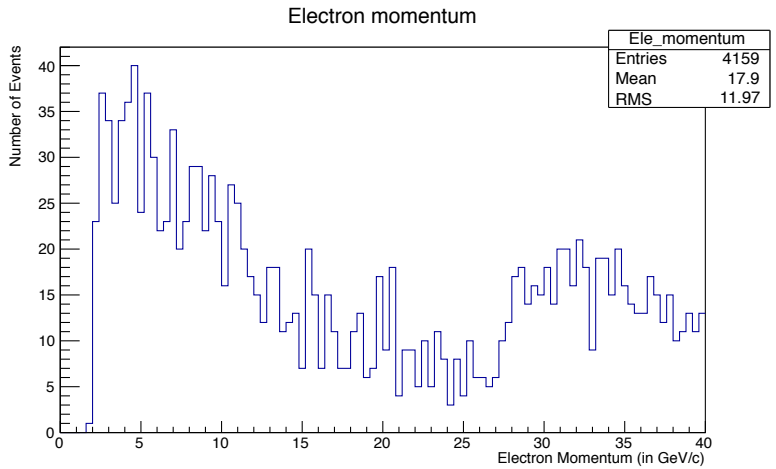
- Loops over different physics objects:
  - Tracks
  - Electrons
  - Muons
  - Photons
  - Jets
  - Taus
  - Missing et
- Fills histograms with p, pt, eta and phi of these objects

Run2011A: SingleElectron

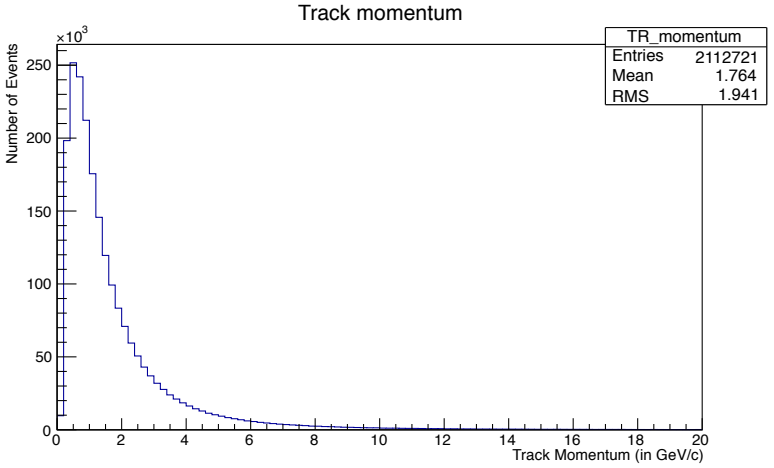
Reconstruction from RAW samples

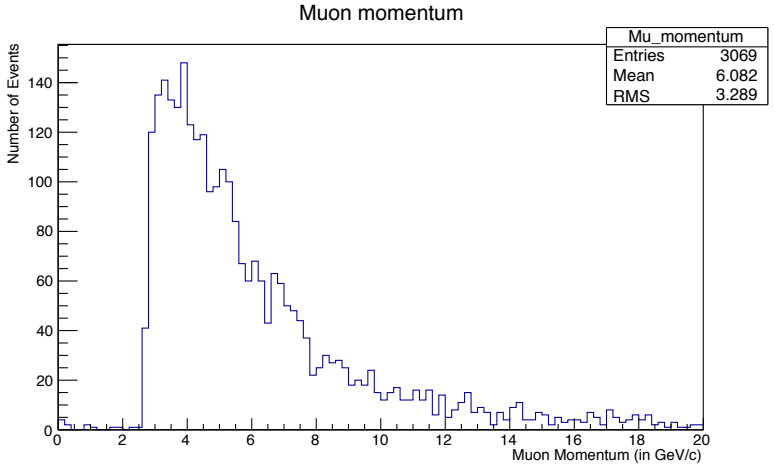


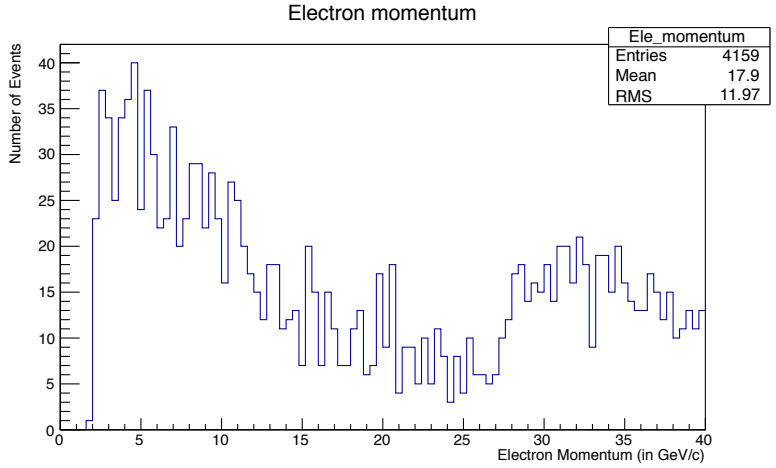




CMS Open Data AOD files





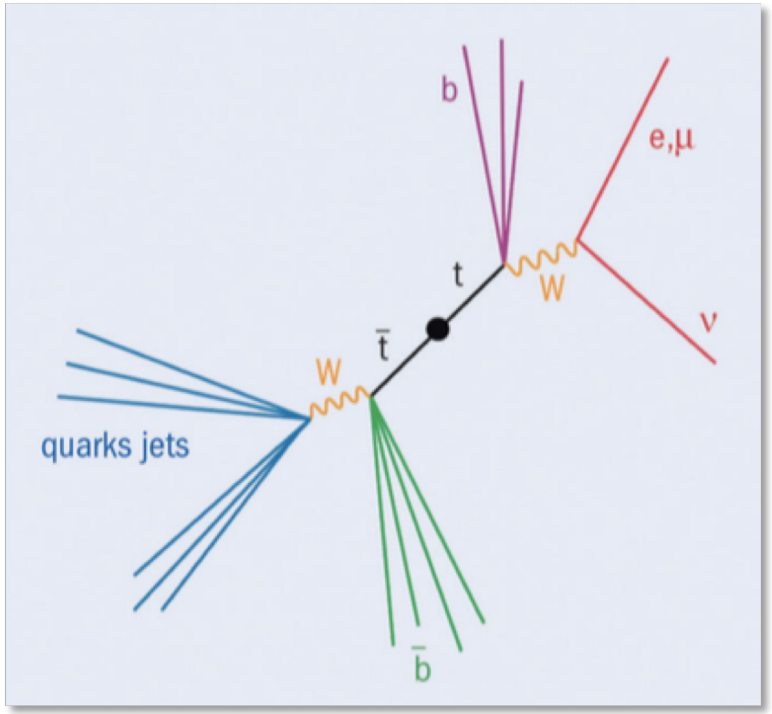


Top quark mass measurement from b-jet energy spectrum

CMS Data Analysis School

- With ReANA, the workflows execute the following steps:

1. Processes the original and simulated collision data to select top-pair events that decay in the eμ channel.
2. Fits and calibrates the b jet energy peak.
3. Compare the results with the standard top-quark mass measurements performed with 8 TeV data.



Summary

H→4l decay analysis example is fully reproducible.

All RAW samples were reconstructed successfully, and have one-to-one match with the original AOD.

Development for ReANA cluster

Workflow implementation for ongoing IA examples

Next Steps

- Correct problems reading condition data for data reconstruction.
- UPRM CMS Research Group: Analysis for DM/Supersymmetry