

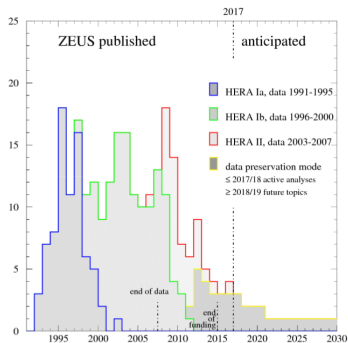


Reusable data, reproducible analyses: today, tomorrow, next decade

Tibor Šimko
CERN

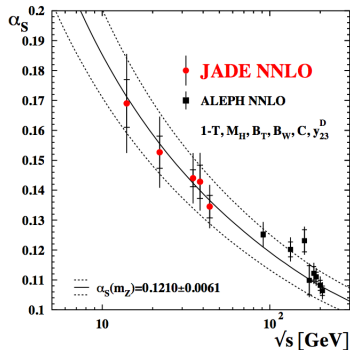
IEEE eScience2018 · Amsterdam, The Netherlands
29 October – 1 November 2018

Long-term value of data!



Achim Geiser <https://indico.cern.ch/event/588219>

Collaborations publish papers even ~ 15 years after data taking ends



DPHEP <https://arxiv.org/abs/1205.4667>

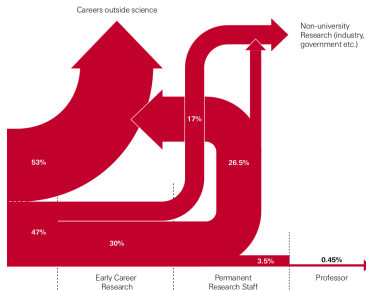
JADE data (1979–1986) still unique even ~ 35 years later

Long-term value of knowledge?



CMS collaboration

Experimental physics done by groups of ~ 3000 physicists

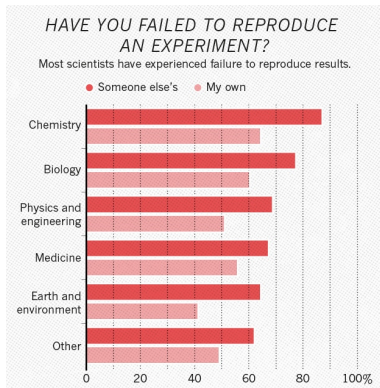
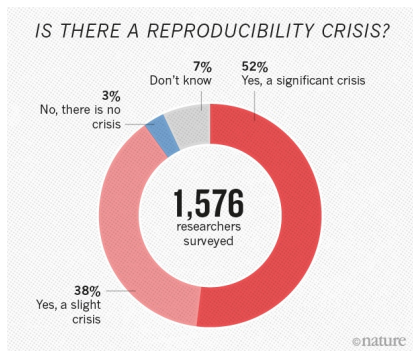


Career after PhD

THE ROYAL SOCIETY

High turnover of young researchers

Reusable and reproducible?



<https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970>

Half of researchers cannot reproduce their own experimental results

FAIR data principles

■ Findable

- persistent identifiers
- rich metadata
- indexed and searchable

■ Accessible

- retrievable by identifiers
- standard protocols
- metadata vs data accessibility

■ Interoperable

- knowledge representation language
- common vocabularies
- references to other metadata and data

■ Reusable

- domain-relevant attributes and community standards
- clear licensing
- provenance tracking

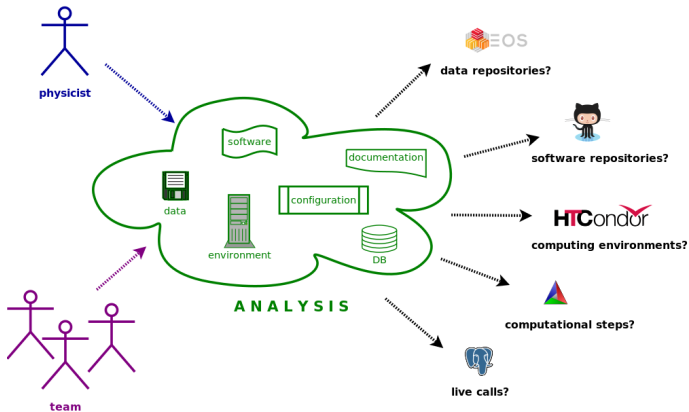
<https://www.nature.com/articles/sdata201618>

Ten rules for repr. comp. research

- 1 For every result, keep track of how it was produced
- 2 Avoid manual data manipulation steps
- 3 Archive the exact versions of all external programs used
- 4 Version control all custom scripts
- 5 Record all intermediate results, when possible in standardized formats
- 6 For analyses that include randomness, note underlying random seeds
- 7 Always store raw data behind plots
- 8 Generate hierarchical analysis output, allowing layers of increasing detail to be inspected
- 9 Connect textual statements to underlying results
- 10 Provide public access to scripts, runs, and results

<https://doi.org/10.1371/journal.pcbi.1003285>

Preserving reusing data analyses



Capturing knowledge and assets of individual physics analyses to facilitate their future reuse

Preserving data 1/2

The screenshot shows the Open Data CERN interface. At the top, there is a search bar and an 'About' link. The main content area displays the title 'Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD)' and its identifier 'i/Mu/Run2010B-Apr21ReReco-v1/AOD, CMS collaboration'. Below this, there is a citation: 'Cite as: CMS collaboration [2014]. Mu primary dataset in AOD format from RunB of 2010 (/Mu/Run2010B-Apr21ReReco-v1/AOD). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.BBMR.C442'. A row of tags includes 'Dataset', 'Online', 'CMS', 'Collision energy 7TeV', 'Accelerator: CERN LHC', and 'Parent Dataset: /Mu/Run2010B-v1/RAW'. The 'Description' section states 'Mu primary dataset in AOD format from RunB of 2010'. The 'Notes' section explains that the dataset contains all runs from 2010 RunB and provides a link to the 'CMS list of validated runs Cert_136033-149442_7TeV_Apr21ReReco_Collisions10_JSON_v2.txt'. The 'Related Datasets' section lists 'i/Mu/Run2010B-v1/RAW'. The 'Characteristics' section reports 'Dataset: 32376291 events 2979 files 3.2 TB in total'. The 'System Details' section shows 'Global tag: FT_R_42_V10A:AB' and 'Recommended release for analysis: CMSSW_4_2_1_patch1'.

How were these data selected?

There are four categories of triggers in the Mu dataset (with significant overlaps):

- 70% inclusive single muon triggers with varying trigger pt threshold 3.5,7.9,11,13,15,17,19,21 GeV plus a few with loosened quality cuts.
- 20% isolated single muon triggers with varying trigger pt threshold 9,11,13,15,17 GeV.
- 10% inclusive dimuon triggers with varying trigger pt threshold 3.5 GeV plus one Z-muon trigger with loosened quality cuts.
- 20% combinations of muon triggers with various pt thresholds 3.5,7.8,9,11 GeV with some EM/e/gamma or hadronic/jet energy deposit with thresholds 6-100 GeV.

How were these data validated?

During data taking all the runs recorded by CMS are certified as good for physics analysis if all subdetectors, trigger, lumi and physics objects (tracking, electron, muon, photon, jet and MET) show the expected performance. Certification is based first on the offline shifters evaluation and later on the feedback provided by detector and Physics Object Group experts. Based on the above information, which is stored in a specific database called Run Registry, the Data Quality Monitoring group verifies the consistency of the certification and prepares a json file of certified runs to be used for physics analysis. For each reprocessing of the raw data, the above mentioned steps are repeated. For more information see:

[CMS data quality monitoring: Systems and experiences](#)

[The CMS Data Quality Monitoring software experience and future improvements](#)

[The CMS data quality monitoring software: experience and future prospects](#)

How can you use these data?

You can access these data through the CMS Virtual Machine. See the instructions for setting up the Virtual Machine and getting started in

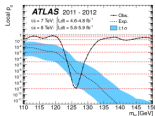
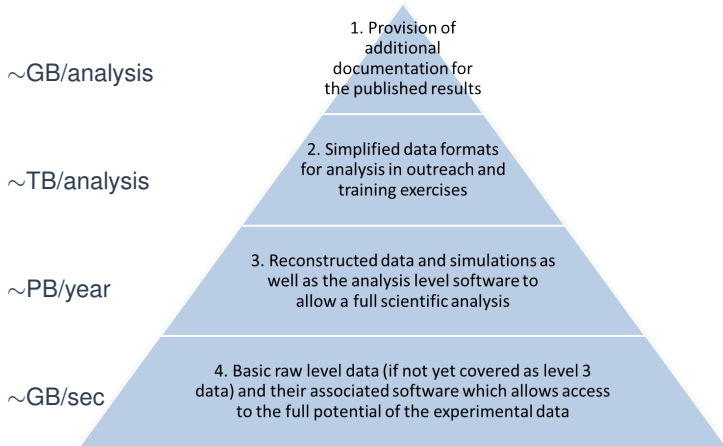
[How to install the CMS Virtual Machine](#)

[Getting started with CMS open data](#)

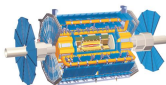
<http://opendata.cern.ch/record/14>

Context information about data selection, validation, use

Preserving data 2/2



analysis



Slicing through LHC data pyramid

Preserving code 1/2

zenodo Research Shared

Settings

1 Flip the switch

2 Create a release

3 Get the badge

institon-cern / decouple

Decouple and recouple — 0.0.1

institon-cern / decouple v1.1.3

Releases

v1.1.3

07a2526 zip tar.gz

```
{
  "name": "Plein, Tllean",
  "affiliation": "Institut für Theoretische Ph",
  "description": "This repository contains the soft",
  "access_right": "open",
  "license": "mit-license",
  "related_identifiers": [
    {
      "identifier": "arXiv:1401.0000",
      "relation": "isCitedBy"
    }
  ]
}
```

.zenodo.json

DOI 10.5281/zenodo.8345

<https://guides.github.com/activities/citable-code>

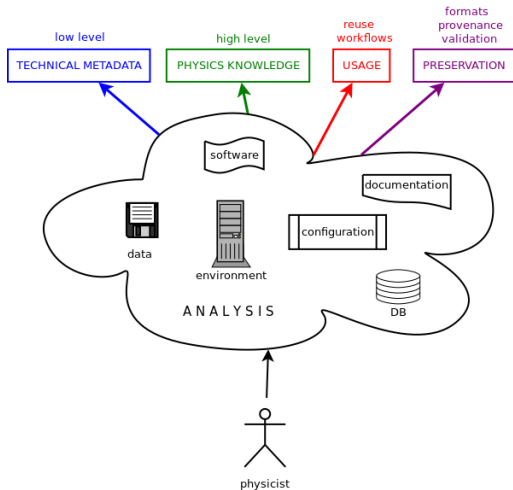
GitHub ↔ Zenodo bridge to automatically preserve releases

Preserving code 2/2

The screenshot shows the Zenodo website interface. At the top, there is a blue navigation bar with the Zenodo logo, a search bar, and links for 'Upload' and 'Communities'. On the right side of the bar are 'Log in' and 'Sign up' buttons. Below the navigation bar, the main content area displays search results for 44,228 items. On the left side, there are three filter panels: 'Access Right' (listing Open, Closed, Restricted, and Embargoed), 'File Type' (listing various file formats like Pdf, Png, Zip, etc.), and 'Keywords' (listing various biological categories like Taxonomy, Animalia, etc.). The main results area shows a list of software packages, each with a date, version, and 'Software' and 'Open Access' tags. The first result is 'E3SM-Project/acme_diags: v1.5.0' by Zeshawn Shaheen, Jili Chengzhu Zhang, golaz, and Charles Doutriaux. The second is 'jstrube/LCIOjl: Keisha' by Jan Strube and Elliot Saba. The third is 'opensciencegrid/gratia-probe: Syntax Fix' by Chris Green, Philippe Canal, Marco Mambelli, tanyakevshina, edjuist, Brian P Bockelman, Suchandra Thapa, Edgar Fajardo, Derek Wetzelt, Brian Lin, John Weigand, hyunwoofnal, and Mats Rynge. The fourth is 'GerardBalaoro/jQuery-Tourer: Version 1.0.0' by Gerard Balaoro and Claudia Romano. Each result includes a brief description and an upload date of October 28, 2018. A 'View' button is present for each result.

Over 58,000 DOIs for software on Zenodo

Preserving analyses 1/4

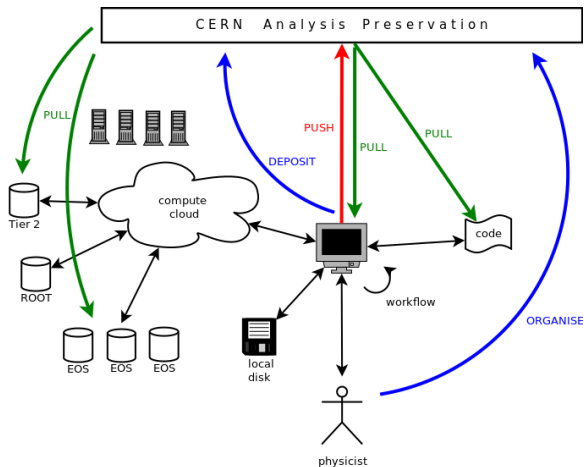


INVENIO

- JSON Schema
- W3C DCAT
- domain-specific fields

Structuring knowledge behind research data analysis

Preserving analyses 2/4



INVENIO

- datasets:
local storage,
cloud storage
- software:
Git, SVN
- information:
DBs, TWiki,
SharePoint
- protocols:
HTTP, XRootD

Taking consistent snapshot of analysis assets at a certain time

Preserving analyses 3/4

The screenshot displays the OpenData CERN search interface. At the top, there is a search bar with the text "Search" and a magnifying glass icon, and an "About" link. Below the search bar, the interface is divided into several sections:

- Filter by type:** A list of categories with counts: Dataset (58), Derived (58), Documentation (4), Activities (4), Environment (1), VM (1), Software (3), Analysis (2), and Tool (1).
- Filter by experiment:** A list of experiments with counts: ALICE (4), ATLAS (97), CMS (1), and LHCb (4).
- Filter by year:** A list of years with counts: 2011 (18) and 2012 (37).
- Filter by file type:** A list of file types with counts: jpg (1) and root (3).
- Filter by keywords:** A list of keywords with counts: education (13), external resource (9), masterclass (66), and teaching (7).

The main search results area shows "Sort by: Best match" and "asc." and "Display: detailed" and "20 results". It indicates "Found 66 results." and lists several ATLAS datasets:

- ATLAS ZPath 2015 Masterclass dataset:** A dataset of 1000 event display files accessible events were recorded in 2012 by the ATLAS det. Buttons: Dataset, Derived, ATLAS.
- ATLAS WPath 2015 Masterclass dataset:** A dataset of 1000 event display files accessible events were recorded in 2011 by the ATLAS det. Buttons: Dataset, Derived, ATLAS.
- ATLAS WPath 2015 Masterclass dataset:** A dataset of 1000 event display files accessible events were recorded in 2011 by the ATLAS det. Buttons: Dataset, Derived, ATLAS.
- ATLAS WPath 2014 Masterclass dataset:** A dataset of 1000 events taken in 2011 by the ATLAS det. Buttons: Dataset, Derived, ATLAS.

On the right side, there is a detailed view of a physics object, "Item #1", under the "Physics Objects" tab. The object is a "tjet" with "Jet type" AK5Cabo, "Jet Corrections" JetCorrections, and "Number" <math>1 <= </math>. The "Selection Criteria" are Loose, Medium, Other, and Tight. The "Discriminator" is Tag, Select Tag, and Value. The "pT Cuts" are Item #1, <math> < > </math>, and GeV. There is a button "Add New Item" at the bottom.

Information discovery through rich search syntax and facets

Preserving analyses 4/4

The screenshot shows the 'CERN Analysis Preservation' web interface. The top navigation bar includes 'Create', 'Search', and a user profile icon. Below the navigation bar, there are 'Share' and 'Save' buttons. The main content area is titled 'Submission Form' and contains a 'Preserve your analysis' section on the left with a text input for 'Analysis Name' and a 'Start Preserving' button. The main form area has several sections: 'Basic Information' with a description and a right-pointing arrow; 'Stripping/Turbo selections [0 items]' with a right-pointing arrow; 'ntuple/userDST-production [0 items]' with a right-pointing arrow; and a 'User Analysis' section with a dropdown arrow.

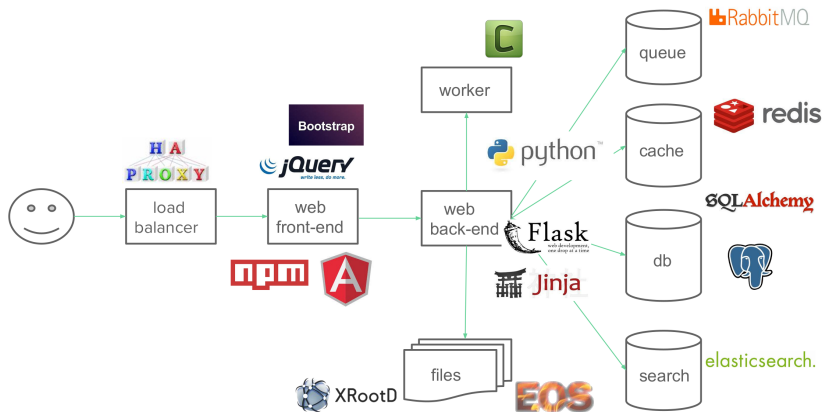
```
$ pip install cap-client
```

```
$ export CAP_SERVER_URL=https://analysispreservation.cern.ch/  
$ export CAP_ACCESS_TOKEN=<your generated access token from server>
```

```
$ cap-client files upload <file path> --pid/-p <existing pid>  
$ cap-client files upload file.json -p 89b593c498874ec8bcafc88944c458a7  
File uploaded successfully.
```

Web based and command-line based deposit workflows

Digital repository technology



Technology stack using **INVENIO** digital repository

FAIR data principles

■ Findable

- persistent identifiers
- rich metadata
- indexed and searchable

■ Accessible

- retrievable by identifiers
- standard protocols
- metadata vs data accessibility

■ Interoperable

- knowledge representation language
- common vocabularies
- references to other metadata and data

■ Reusable

- domain-relevant attributes and community standards
- clear licensing
- provenance tracking

<https://www.nature.com/articles/sdata201618>

Reusing analyses

reana

Reproducible research data analysis platform

Flexible

Run many computational workflow engines.



Scalable

Support for remote compute clouds.



Reusable

Containerise once, reuse elsewhere. Cloud-native.



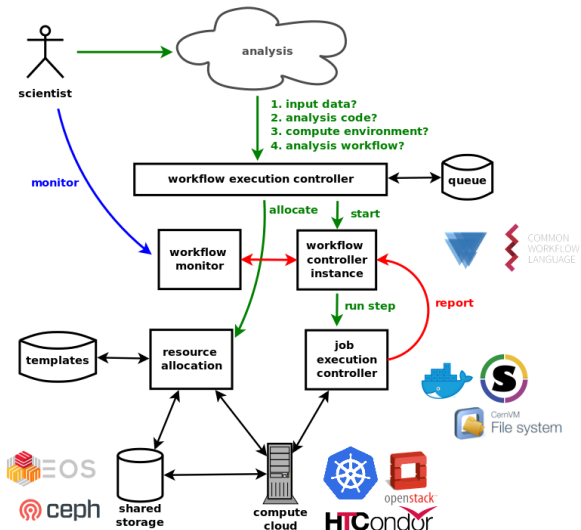
Free

Free Software. MIT licence.
Made with ❤️ at CERN.



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

REANA architecture



REANA technology

The screenshot shows the REANA Hub interface. At the top, there's a navigation bar with 'Pull requests', 'Issues', 'Marketplace', and 'Explore'. Below that, the 'REANA Hub' logo and tagline 'Reproducible research data analysis platform' are visible. A search bar and navigation tabs for 'Repositories', 'People', 'Teams', 'Projects', and 'Settings' are present. The main content area is titled 'Pinned repositories' and lists several repositories with their respective languages, stars, and versions. A sidebar on the right shows 'Top languages' and 'People' sections.

REANA Hub
Reproducible research data analysis platform
http://www.renatahub.io | info@renatahub.io

Repositories 25 | People 15 | Teams 2 | Projects 0 | Settings

Pinned repositories Customize pinned repositories

- reana**
REANA: Reproducible research data analysis platform
Python ★ 17 V7
- reana-client**
REANA command-line client
Python ★ 1 V6
- reana-cluster**
REANA cluster management
Python ★ 1 V4

Search repositories... Type: All Language: All New

reana-workflow-controller
REANA Workflow Controller
Python V5 GPU 2.0 Updated 20 minutes ago

reana-job-controller
REANA Job Controller
Python V3 GPU 2.0 Updated an hour ago

reana-cluster
REANA cluster management
Python ★ 1 V4 GPU 2.0 Updated 3 hours ago

reana-message-broker
REANA Message Broker
Python V4 GPU 2.0 Updated 5 hours ago

reana-client
REANA command-line client
Python ★ 1 V6 GPU 2.0 Updated 5 hours ago

Top languages
Python C Makefile C++ Jupyter Notebook

People 15 >

Invite someone

■ micro-services



■ REST API



Bravado

■ services



■ deployments



Four questions

1 Input data

What is your input data?

- input files
- input parameters

3 Compute environment

What is your environment?

- operating system
- database calls

2 Analysis code

Which code analyses it?

- software frameworks
- user code

4 Analysis workflow

Which steps did you take?

- single command
- complex workflows

Simple example

```
Region,1500,1600,1700,1750,1800,1850,1900,1950,1999,2008,2010,2012,2050,2150
World,100,100,100,100,100,100,100,100,100,100,100,100,100,100,100
Africa,18.8,19.7,15.5,13.4,10.9,8.8,8.1,8.8,12.8,14.5,14.8,15.2,19.8,23.7
Asia,53.1,58.4,63.9,63.5,64.9,64.1,57.4,55.6,60.8,60.4,60.4,60.3,59.1,57.1
Europe,18.3,19.1,18.3,20.6,20.8,21.9,24.7,21.7,12.2,10.9,10.7,10.5,7,5.3
Latin America and the Caribbean,8.5,1.7,1.5,2.2,5.3,4.5,6.6,8.5,8.6,8.6,8.6,9.1,9.4
Northern America,0.7,0.5,0.3,0.3,0.7,2.1,5.6,8.5,1.5,5.5,4.4,4.1
Oceania,0.7,0.5,0.4,0.3,0.2,0.2,0.4,0.5,0.5,0.5,0.5,0.5,0.5,0.5,0.5
```

1 input: CSV file

```
FROM centos:7
RUN yum install -y epel-release
RUN yum install -y \
    gcc \
    python-devel \
    python-pip
RUN pip install ipython==5.0.0 jupyter==1.0.0
ADD world_population_analysis.ipynb /code/
ADD World_historical_and_predicted_populations_in_percentage.csv /code/
WORKDIR /code
CMD ["jupyter", "nbconvert", "world_population_analysis.ipynb"]
```

3 environment: CentOS7, IP5

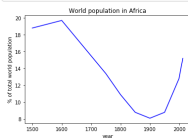
```
In [2]: # define input parameters
input_file = "../data/World_historical_and_predicted_populations_in_percentage.csv"
output_file = "../results/plot.png"
region = 'Africa'
year_min = 1500
year_max = 2012

In [3]: # read input data file
df = pd.read_csv(input_file)

In [4]: # add index
df = df.set_index("Region", drop=False)


In [5]: # select region and years based on input parameters
dfs = df.loc[region, str(year_min):str(year_max)]
dft = pd.DataFrame({'year': dfs.index.astype(int), 'percentage': dfs.values}, columns=['year', 'percentage'])

In [6]: # create output plot and save it to a file
plot = plt.plot(dft['year'], dft['percentage'], color='blue')
plt.title('World population in {}'.format(region))
plt.xlabel('year')
plt.ylabel('% of total world population')
plt.savefig(output_file)
```

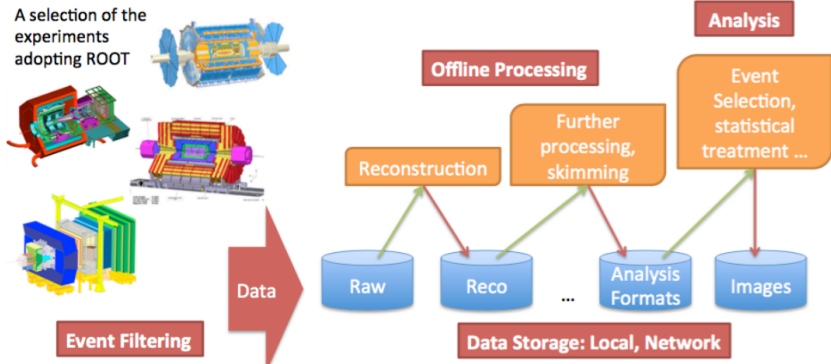


2 code: Jupyter notebook

4 workflow: papermill ...

 <https://github.com/reanahub/reana-demo-worldpopulation>

HEP data analyses



D. Krücker *et al* <https://indico.desy.de/indico/event/18343>

Targeting both data production and data analysis stages

Data production

Validation code for reprocessing AOD from 2011 MinimumBias RAW sample

Lasilä-Perini, Kati (2017). Validation code for reprocessing AOD from 2011 MinimumBias RAW sample. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.9581BGV4

Software **14.4 kB** **1 file** **14.4 kB** in total

Description

This code for validation reproduces AOD format in this record from the 2011 MinimumBias RAW sample. It only contains the configuration file to be run on 2011 OpenData VM using CMS5W_5_3_32. No compilation is required. The configuration file has been slightly modified from the original one to take into account the OpenData computing environment (global tag, input file, commenting out unnecessary steps). Note that the code in this category is not meant to be a pedagogical example but is a validation tool.

Use with

Use this with the following dataset:

/MinimumBias/Run2011A-v1/RAW

Characteristics

Dataset: 1 files **14.4 kB** in total

System Details

Use this code with the CMS Open Data VM environment for 2011 open data
Software release: CMS5W_5_3_32
CMS VM Image, for 2011 and 2012 CMS open data

Example code for production of flat jet tuple using 2011 data

Zenalev, Oleksandr; Haapalehto, Mattias (2017). Example code for production of flat jet tuple using 2011 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.JT93.129

Software **17.0 kB** **1 file** **17.0 kB** in total

Description

This code is a CMS5W module producing flat tuples from 2011A Jet data, set up at Research level, i.e. It requires university-student-level programming experience. Minimal acquaintance with Linux and the ROOT analysis package (<https://root.cern.ch/>) as well as a basic text editor is needed.

Use with

Use this with 2011 jets primary dataset and QCD MC dataset (see detailed instructions in the readme file and code).

```
jet/Run2011A-12Oct2013-v1/AOD
/QCD_Pt-80to120_TuneZ2_7TeV_gyNtA6/Summer11LegDR-PU_S13_START53_LV6-v1/AODSIM
```

Notes

The content of the resulting root file is described in readme

Characteristics

Dataset: 1 files **17.0 kB** in total

System Details

Use this code with the CMS Open Data VM environment
Software release: CMS5W_5_3_32

Reconstruction and flat jet tuple production from CMS Open Data

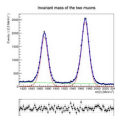
Data analyses

```
$ cat inputs/names.txt  
Jane Doe  
Joe Bloggs  
$ reana-client start  
$ reana-client download  
$ cat outputs/greetings.txt  
Hello Jane Doe!  
Hello Joe Bloggs!
```

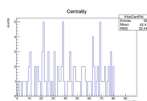
"Hello world"



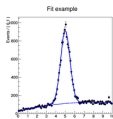
Parametrised jupyter notebook



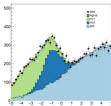
LHCb rare charm decay search



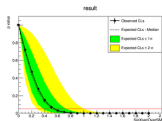
ALICE LEGO train test run



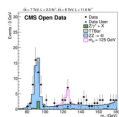
ROOT/RooFit physics analysis



ATLAS BSM search



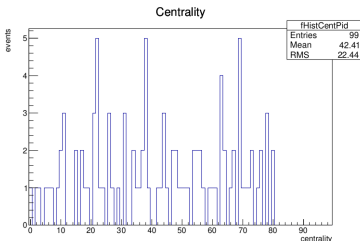
ATLAS RECAST



CMS Higgs-to-four-leptons

Several physics analysis examples

Example: ALICE analysis train



<https://github.com/reanahub/reana-demo-alice-lego-train-test-run/>

Using ALICE production software for train test run and validation

Example: LHCb rare charm decay

Physics Letters B 756 (2016) 285–292



Search for $D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^-$ and $D_{(s)}^+ \rightarrow \pi^- \mu^+ \mu^-$ decays

LHCb Collaboration

ARTICLE INFO
Article history:
Received 16 April 2015
Received in revised form 1 June 2015
Accepted 4 June 2015
Available online 7 June 2015
Editor: M. Bona

ABSTRACT

A search for rare charm $D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^-$ and $D_{(s)}^+ \rightarrow \pi^- \mu^+ \mu^-$ decays is performed using precision calibration data corresponding to an integrated luminosity of $36.1 \pm 0.4 \text{ fb}^{-1}$ recorded by the LHCb experiment in 2011. No signal is observed and the 90% (95%) confidence level (CL) limits on the branching fractions are found to be:
 $\mathcal{B}(D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^-) < 7.3 (8.3) \times 10^{-6}$,
 $\mathcal{B}(D_{(s)}^+ \rightarrow \pi^- \mu^+ \mu^-) < 4.1 (4.6) \times 10^{-6}$,
 $\mathcal{B}(D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^-) < 2.2 (2.5) \times 10^{-6}$,
 $\mathcal{B}(D_{(s)}^+ \rightarrow \pi^- \mu^+ \mu^-) < 1.2 (1.4) \times 10^{-6}$.
These limits are the most stringent to date.

© 2015 LHCb. Published by Elsevier B.V. <http://dx.doi.org/10.1016/j.phlet.2015.05.028>

1. Introduction

Flavour-changing neutral current (FCNC) processes are rare within the Standard Model (SM) as they cannot occur at tree level. At the loop level, they are suppressed by the GIM mechanism [1] but the cancellation was established in $D^+ \rightarrow \pi^+ \mu^+ \mu^-$ and $D^+ \rightarrow \pi^+ \mu^+ \mu^-$ decays with branching fractions of the order 10^{-10} and 10^{-9} , respectively [2,3]. In contrast to the B meson system, where the very high scale of the top quark in the loop renders the suppression, the GIM cancellation is absent and it is crucial to search for expected branching fractions for $D \rightarrow \mu^+ \mu^-$ processes in the $(3-3) \times (3-3)$ gauge [4–6]. This suppression provides a unique opportunity to search for FCNC charm decays and to probe the coupling of top-type quarks to electroweak processes, as discussed in [7,8,10,9].

The decay $D \rightarrow \pi^+ \mu^+ \mu^-$, although not a FCNC process, provides the most sensitive channel shown in [9,10]. This can be used to motivate a general $D \rightarrow \pi^+ \mu^+ \mu^-$ search as an example of a precision flavour physics process, also supported by a direct [10]. Nevertheless it is useful to distinguish between FCNC and weak annihilation contributions. Note that throughout this letter, the inclusion of charm quark processes is implied.

Many variations of the top such as supersymmetric models with R-parity violation or models involving a fourth quark generation, introduce additional diagrams that a priori need not be

suppressed in the same manner as the SM contribution [11]. The most stringent limit published to date is $\mathcal{B}(D^+ \rightarrow \pi^+ \mu^+ \mu^-) < 3.6 \times 10^{-6}$ [10] CL by the Belle collaboration [9]. The Belle collaboration places the most stringent limit on the D^+ weak annihilation decay with $\mathcal{B}(D^+ \rightarrow \pi^+ \mu^+ \mu^-) < 2.4 \times 10^{-6}$ [10].

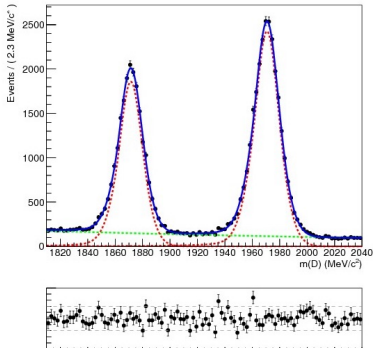
Logic similar to that of the Belle experiment, such as $D^+ \rightarrow \pi^+ \mu^+ \mu^-$ [shown in Eq. (10)] are forbidden in our SM because they are only local through top quark mixing facilitated by a non-zero particle such as a flavour violation [11]. The most stringent limit on the analogous decay at 90% CL is $\mathcal{B}(D^+ \rightarrow \pi^+ \mu^+ \mu^-) < 3 \times 10^{-6}$ and $\mathcal{B}(D^+ \rightarrow \pi^+ \mu^+ \mu^-) < 1.4 \times 10^{-6}$ [10] by the Belle Collaboration [11]. A recent decay set the most stringent limit on SM decays in general with $\mathcal{B}(D^+ \rightarrow \pi^+ \mu^+ \mu^-) < 1.0 \times 10^{-6}$ CL by the LHCb collaboration [12].

This letter presents the results of a search for $D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^-$ and $D_{(s)}^+ \rightarrow \pi^- \mu^+ \mu^-$ decays using pp collision data, corresponding to an integrated luminosity of $36.1 \pm 0.4 \text{ fb}^{-1}$ recorded by the LHCb experiment. The signal channels are examined in the control channels $D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^0$ and $D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^+$, which have branching fraction predictions of $\mathcal{B}(D^+ \rightarrow \pi^+ \mu^+ \mu^0) = 1.58 \pm 0.10 \times 10^{-6}$ and $\mathcal{B}(D^+ \rightarrow \pi^+ \mu^+ \mu^+) = 0.28 \pm 0.14 \times 10^{-6}$ [13].

2. The LHCb detector and trigger

The LHCb detector [14] is a single-arm forward spectrometer covering the pseudorapidity range $2 < \eta < 5$, designed for the study of particles containing b or c quarks. The detector includes

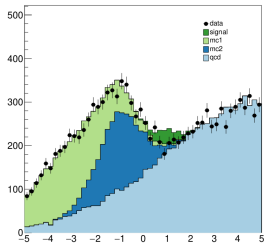
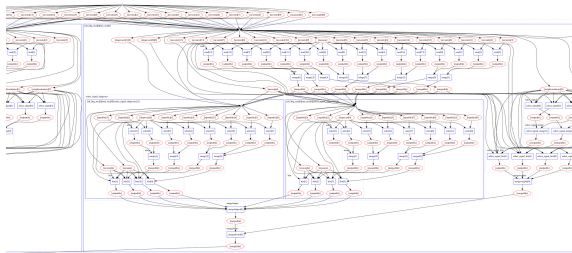
Invariant mass of the two muons



<https://github.com/reanahub/reana-demo-lhcb-d2pimumu/>

Search for $D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^-$ and $D_{(s)}^+ \rightarrow \pi^- \mu^+ \mu^+$ decays

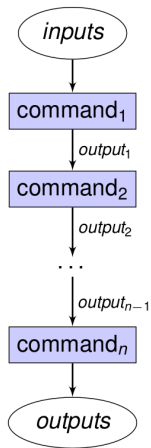
Example: BSM search



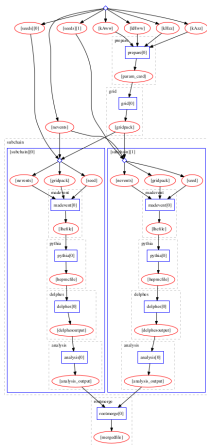
<https://github.com/reanahub/reana-demo-bsm-search/>

Complex computational workflows typical in particle physics analyses

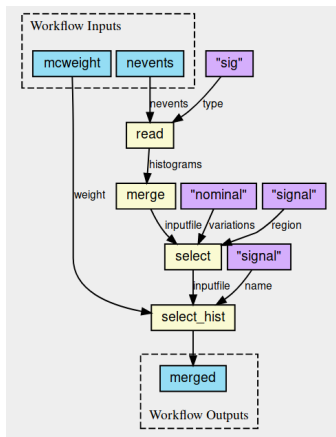
Computational workflows



Serial



Yadage



CWL

Running workflows

1

Structure your analysis

```
version: 0.2.0
code:
  files:
    - code/mycode.py
inputs:
  files:
    - inputs/mydata.csv
parameters:
  myparameter: myvalue
environments:
  - type: docker
    image: johndoe/mypython:1.0
workflow:
  type: cwl
  file: workflow/myworkflow.cwl
outputs:
  files:
    - outputs/myplot.png
```

[more](#)

2

Select a REANA cluster...

```
$ export
REANA_SERVER_URL=https://reana.cern.ch/
```

...or install your own

```
# install kubectl 1.9.1 and minikube
0.23.0
$ sudo dpkg -i kubectl*.deb minikube*.deb
$ minikube start --kubernetes-
version="v1.6.4"
# install reana-cluster utility
$ mkvirtualenv reana-cluster
$ pip install reana-cluster
# deploy new cluster and check progress
$ reana-cluster init
$ reana-cluster status
# set environment variables for reana-
client
$ eval ${reana-cluster env}
```

[more](#)

3

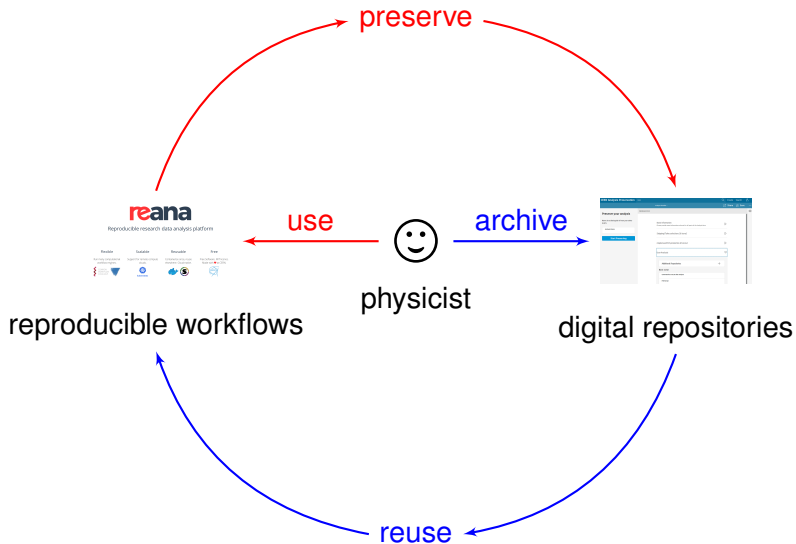
Run your analysis

```
# install reana-client
$ mkvirtualenv reana-client -p /usr/bin
/python2.7
$ pip install reana-client
$ reana-client ping
# create new workflow
$ export REANA_WORKON=$(reana-client
workflow create)
# upload runtime code and inputs
$ reana-client code upload ./code/*
$ reana-client inputs upload ./inputs/*
# start workflow and check progress
$ reana-client workflow start
$ reana-client workflow status
# download outputs
$ reana-client outputs list
$ reana-client outputs download
myplot.png
```

[more](#)

Rich command-line user client

Reproducibility \rightleftharpoons Preservation



Conclusions

■ reusable data, reproducible analyses

- structure knowledge → JSON schema
- capture assets → push and pull
- actionable processes → “workflow is the new data”

■ leverage on synergies

- diverse scientific domains → DAG
- industry standards → containers, “distributed Linux”

■ overcoming challenges




- technological → scaling out, O(10GB) containers, O(10k) steps
- sociological → culture change, smooth integration into daily work

CERN IT D. Kousidis, R. Maciulaitis, J. Okraska, D. Rodriguez, T. Šimko · **CERN SIS** S. Dallmeier-Tiessen, S. Feger, P. Fokianos, A. Lavasa, S. van de Sandt, I. Tsanaktsidis, A. Trzcinska · **ALICE** Y. Foka, M. Gheata, C. Grigoras, M. Zimmermann · **ATLAS** K. Cranmer, L. Heinrich, A. Sanchez Pineda, D. Rousseau, F. Socher · **CMS** H. Bittencourt, A. Calderon, E. Carrera, A. Geiser, A. Huffman, C. Lange, K. Lassila-Perini, L. Lloret, T. McCauley, A. Rao, A. Rodriguez Marrero · **LHCb** S. Amerio, C. Burr, B. Couturier, S. Neubert, C. Parkes, S. Roiser, A. Trisovic · **OPERA** G. De Lellis, S. Dmitrievsky · **CERN CernVM** J. Blomer · **CERN EOS** L. Mascetti, H. Rousseau · **CERN Kubernetes** R. Rocha · **CERN OpenShift** A. Lossent, A. Peon

References






CERN Open Data

-  <http://opendata.cern.ch>
-  <http://github.com/cernopendata>
-  [cernopendata](#)






CERN Analysis Preservation

-  <http://analysispreservation.cern.ch>
-  <http://github.com/cernanalysispreservation>
-  [analysispreserv](#)






REANA

-  <http://www.reanahub.io>
-  <http://github.com/reanahub>
-  [reanahub](#)






Invenio

-  <http://inveniosoftware.org>
-  <http://github.com/inveniosoftware>
-  [inveniosoftware](#)



Zenodo

-  <https://zenodo.org>
-  <http://github.com/zenodo>
-  [zenodo_org](#)