

CMS

CERN

LHC

[Large Hadron Collider]

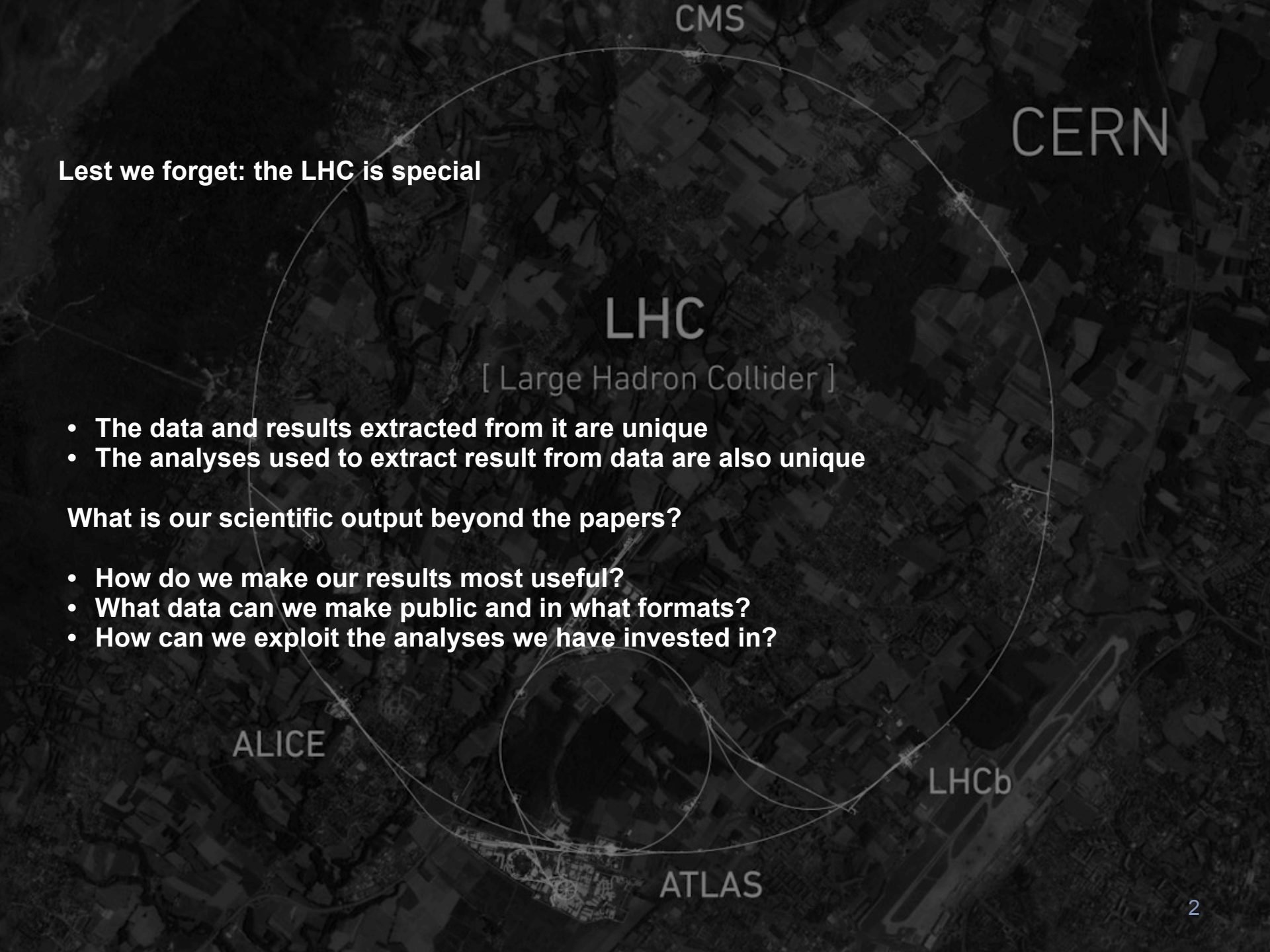
Data and Analysis Preservation

ALICE

L Heinrich
LHCP 2020

LHCb

ATLAS



Lest we forget: the LHC is special

LHC

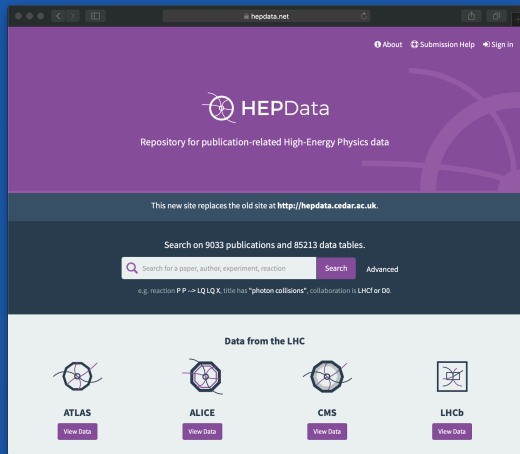
[Large Hadron Collider]

- The data and results extracted from it are unique
- The analyses used to extract result from data are also unique

What is our scientific output beyond the papers?

- How do we make our results most useful?
- What data can we make public and in what formats?
- How can we exploit the analyses we have invested in?

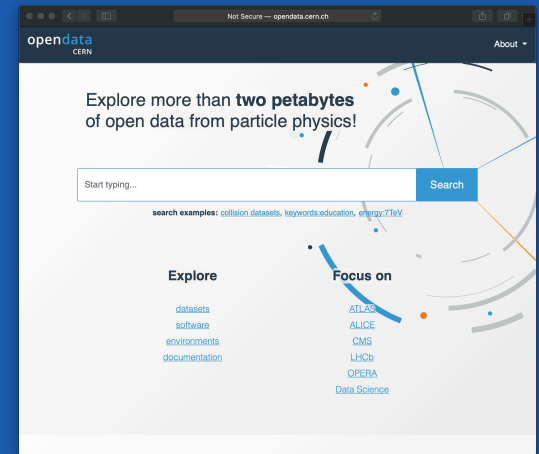
Three broad areas of activity



Analysis Data Products
and Result Preservation



Reproducible Workflows &
Analysis Preservation



Open Data for Outreach,
Education and Research

HepData has been the main vehicle to provide
high quality data products for published analyses

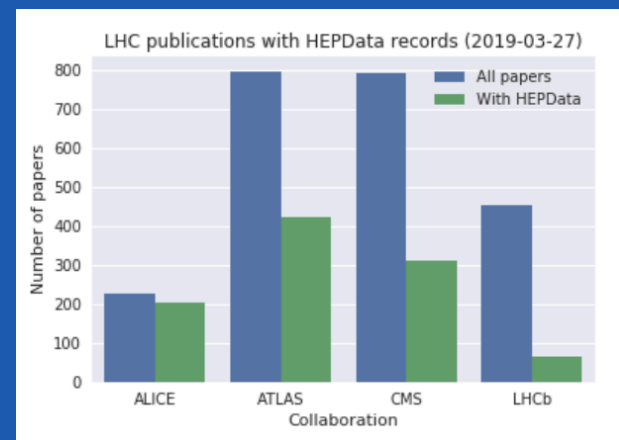
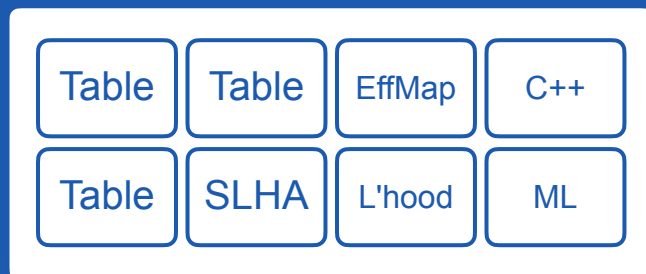
publicly available. All LHC experiments rely on this.

- HepData submission often required for analysis approval

Types of data products expanded from



to broader collection of data



ALICE: 90%
ATLAS: 52%
CMS: 39%
LHCb: 14%

Additional Material helps approximate reimplementation of data analyses w/ e.g. Rivet (can cover also BSM and HI)

```
#include "SimpleAnalysis/AnalysisClass.h"  
#include "SimpleAnalysis/NtupleMaker.h"  
#include "SimpleAnalysis/PDFReweight.h"  
#include <LHAPDF/LHAPDF.h>  
#include "TMath.h"
```

```
DefineAnalysis(EwkOneLeptonTwoBjets2018)  
// Wh->l+bb+met analysis (Run2 data)
```

```
void EwkOneLeptonTwoBjets2018::Init()  
  
{  
    // Define signal/control regions  
    // Define signal regions
```

```
// -*- C++ -*-
#include "Rivet/Analysis.hh"
#include "Rivet/Projections/ChargedFinalState.hh"
#include "Rivet/Tools/Correlators.hh"
#include "Rivet/Tools/AliceCommon.hh"
#include "Rivet/Projections/AliceCommon.hh"

namespace Rivet {

    /// @brief Multiparticle azimuthal correlations pp, pPb, XeXe and PbPb
    class ALICE_2019_I1723697 : public CumulantAnalysis {
    public:

        /// Constructor
        ALICE_2019_I1723697() :
            CumulantAnalysis("ALICE 2019 I1723697") {}
    };
}
```

Confronting Experimental Data with Heavy-Ion Models

RIVET for Heavy Ions

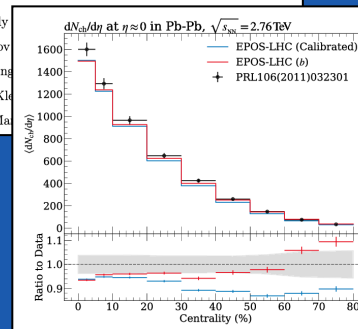
Christian Bierlich,^{1,2} Andy

Peter Harald Lindenov

Jan Fiete Grosse-Oetring

Patrick Kirchgäßer,⁶ Jochen KleChristine O. Rasmussen,² Ma

© 2011 Pearson Education, Inc.

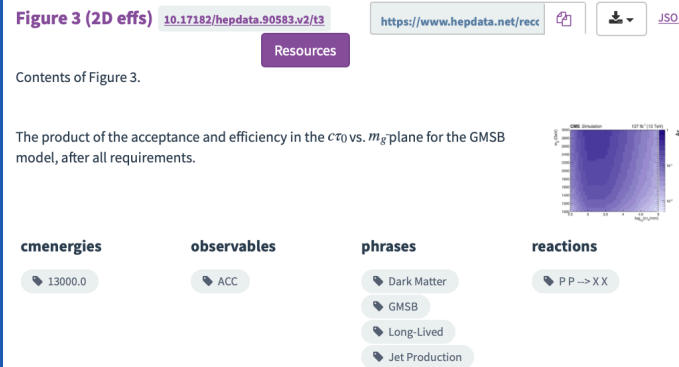


C++ Code Snippets as starting point, or (better) full Rivet Routine

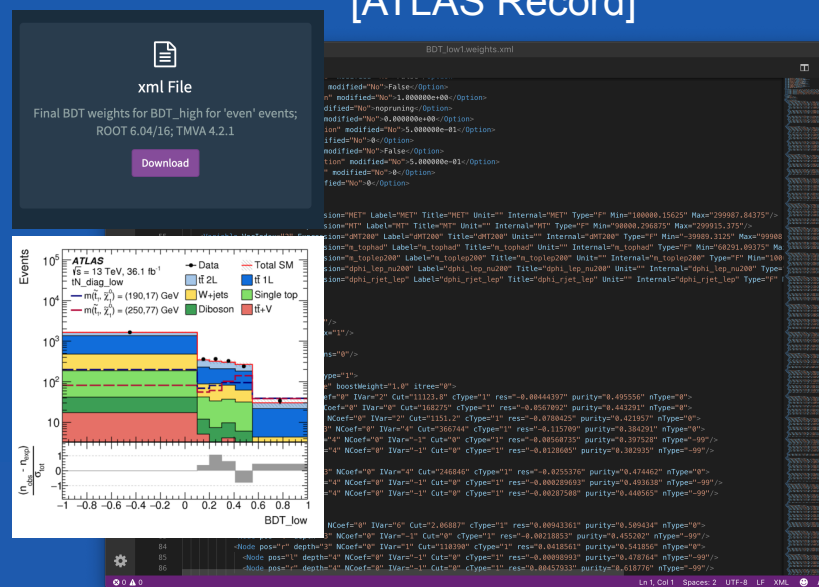
Efficiency Maps:

ML models uploaded to HepData

[ATLAS Record]



[CMS Record]

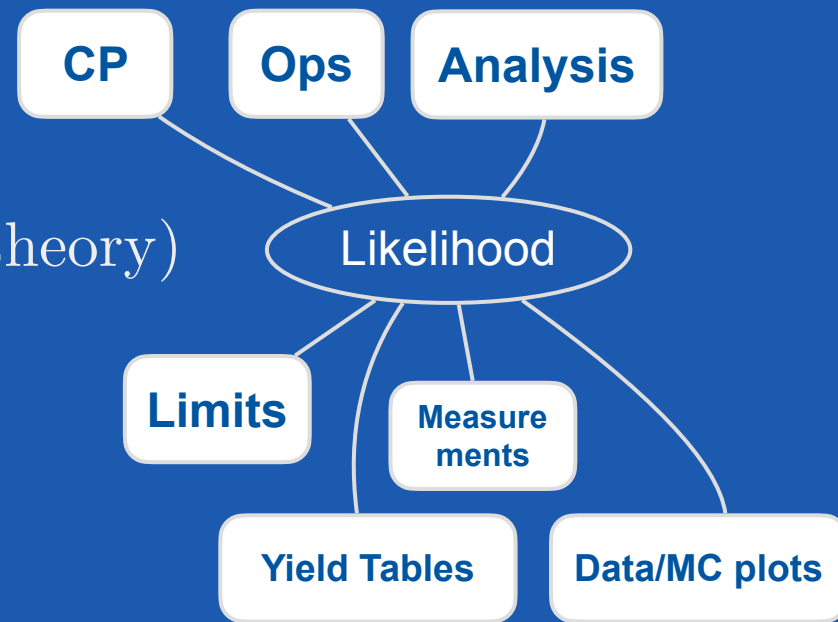


A new Frontier: Public Likelihoods

$$p(\text{theory}|\text{data}) \sim p(\text{data}|\text{theory}) \cdot p(\text{theory})$$

likelihood:
experimentalists

prior:
theorists



The likelihood is the central object in analysis

- the best data product we can provide in principle
- a high-density compressed representation
- encodes detailed systematic uncertainties

Often HepData information (yields, uncertainties...) is used to reconstruct approximate likelihood

What if we just provided it directly?

Long History:

2000: 1st PHYSTAT

2010: Introduction of Workspaces

2012: ATLAS profile likelihoods

2017: CMS Simplified Likelihoods

Massimo Corradi

It seems to me that there is a general consensus that what is really meaningful for is *likelihood*, and almost everybody would agree on the prescription that experiments should provide a likelihood function for these kinds of results. Does everybody agree on this statement, to hoods?

Louis Lyons

Any disagreement? Carried unanimously. That's actually quite an achievement for t

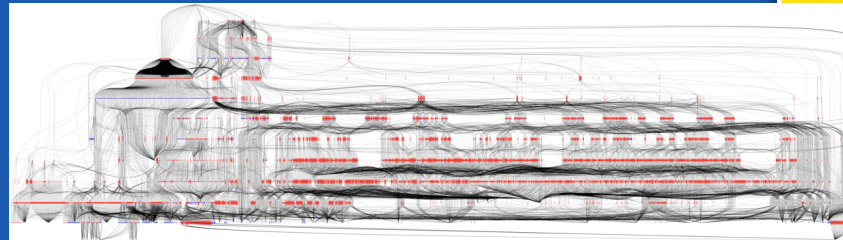
ORGANISATION EUROPÉENNE POUR LA RECHERCHE NUCLÉAIRE
CERN EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

WORKSHOP ON CONFIDENCE LIMITS

CERN, Geneva, Switzerland
17-18 January 2000

PROCEEDINGS

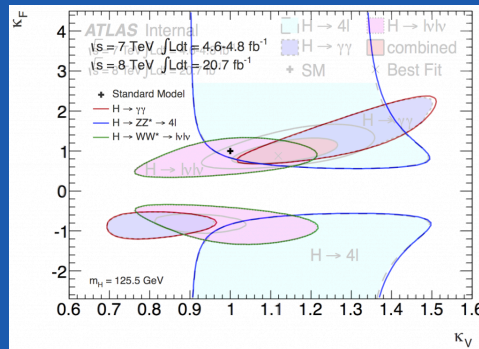
Editors: F. James, L. Lyons, Y. Pivovarov



ement
primar-
nent be
; ingre-
ations).
ce. The

power of the workspace is that it allows one to save data and an arbitrarily complicated model to disk in a ROOT file. These files can then be shared or archived, and they provide all the low-level ingredients necessary for a proper combination in a unified framework. A direct advantage of this is a **digital publishing of the results**.

The **RooWorkspace** class of RooFit provides the low-level functionality for storing the full model and the data and, in addition, it provides a convenient functionality to create easily the model via a string interface (workspace factory).

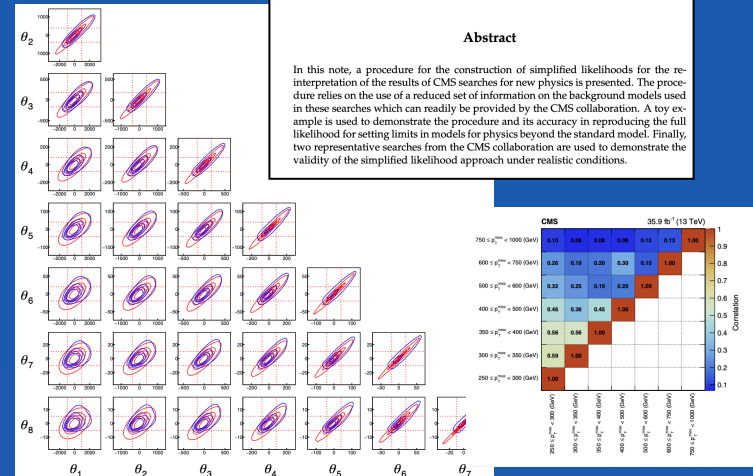


Simplified likelihood for the re-interpretation of public CMS results

The CMS Collaboration

Abstract

In this note, a procedure for the construction of simplified likelihoods for the re-interpretation of the results of CMS searches for new physics is presented. The procedure relies on the use of a reduced set of information on the background models used in these searches which can readily be provided by the CMS collaboration. A toy example is used to demonstrate the procedure and its accuracy in reproducing the full likelihood for setting limits in models for physics beyond the standard model. Finally, two representative searches from the CMS collaboration are used to demonstrate the validity of the simplified likelihood approach under realistic conditions.



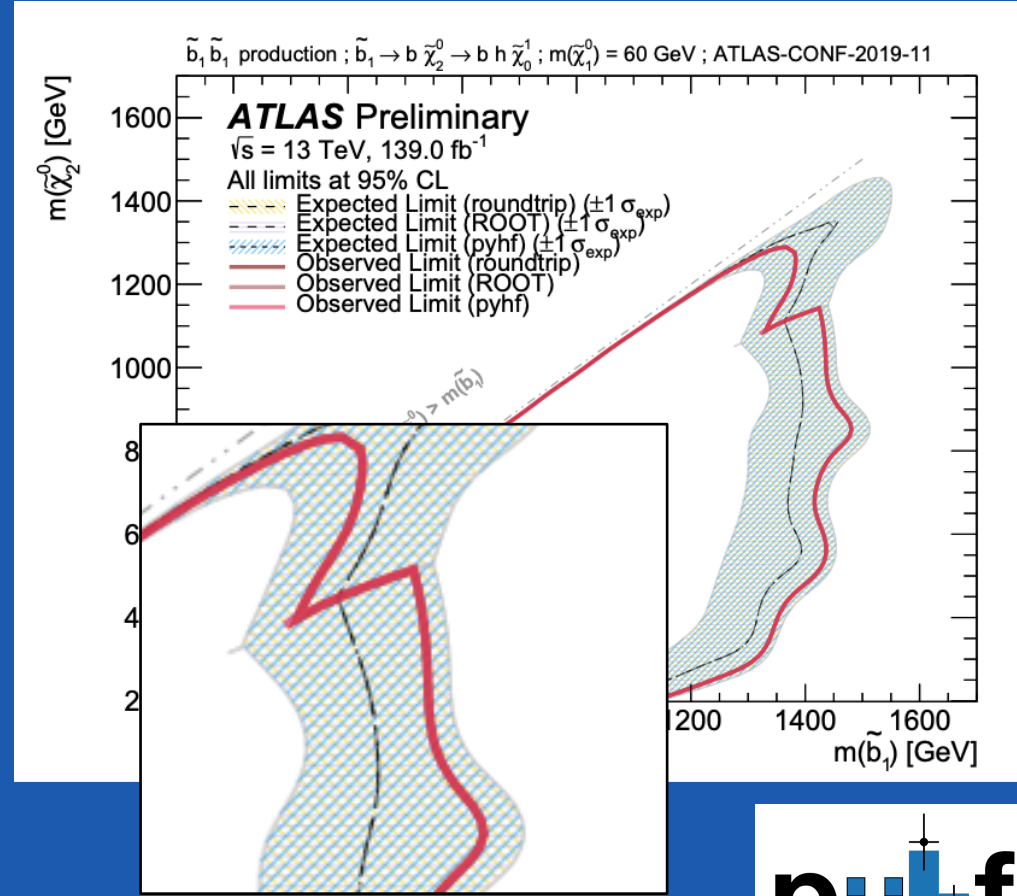
2019: first full likelihood release

- using JSON Schema of HistFactory class of Likelihoods

The screenshot shows the HEPData website interface. The search results for 'bottom-squark pair production' are displayed. The results include a list of publications, a table of common resources, and a section for additional publication resources. The common resources table lists various quantities and their values. The additional resources section includes an external link and a gz file.

Common Resources	Value
Missing Transverse Energy	2
Effective Mass	2
Object Based Missing Transverse Energy significance	2
MaxMin alternative algorithm average m_{Higgs}	2
Leading jet pT	2
MaxMin algorithm m_{Higgs}	2
Efficiency_SRA_M_m60	2
Acceptance_SRC_28	2
Acceptance_SRC_26	2
Acceptance_SRC_24	2
Acceptance_SRA_M_dm130	2
Acceptance_SRB	2

```
$> curl -sL https://doi.org/10.17182/hepdata.89408.v1/r2|tar -xf -
$> pyhf cls RegionA/BkgOnly.json --patch RegionA/patch.sbottom_1300_205_60.json
{
  "CLs_exp": [
    0.09022509053507759,
    0.1937839194960632,
    0.38432344933992,
    0.6557757334303531,
    0.8910420971601081
  ],
  "CLs_obs": 0.24443627759085326
}
```



Internal Reuse:

Efforts by all LHC experiments to foster internal analysis preservation.
Ingredients for AP:

capture software

archive analysis code incl.
dependencies

capture commands

what do with the
captured software

capture workflow

order of individual steps

data assets

input data needed
to run the analysis

Additionally: data

CERN provides infrastructure to
assist experiments

REANA: workflows-as-a-service

CAP: store workflow and
other analysis artifacts
(software, etc)



reana

Reproducible research data analysis platform

CERN
Analysis Preservation

capture, preserve and reuse physics analyses



1. capture software

archive analysis
code incl. deps.

2. capture commands

what do with the
captured software

3. capture workflow

order of individual
steps

Containers universally seen as suitable technology:

all experiments have some
infrastructure to run
experiment / analysis code
in containers

REANA Environment AliPhysics

build unknown glitter join chat License GPL v2

About

`reana-env-aliphysics` provides a container image with encapsulated runtime execution environment for AliPhysics based ALICE data analyses. The container image includes all the necessary dependencies and does not have any external requirements (such as CVMFS).

`reana-env-aliphysics` was developed for use in the REANA reusable research data analysis platform.

lhcb-analysis-preservation > containerization-cookie > Details

C

containerization-cookie

Project ID: 31307 | [Leave project](#)

45 Commits 1 Branch 0 Tags 287 KB Files

Cookiecutter template for analysis containerization



atlas/athanalysis

By atlas • Updated 8 days ago

ATLAS Athena Analysis Release

Container

1M+ Downloads 3 Stars



atlas/analysisbase

By atlas • Updated 8 days ago

ATLAS Standalone Analysis Release

Container

1M+ Downloads 10 Stars

clelange / cmssw-docker

<> Code

Issues 5

Pull requests 0

Actions

Projects 0

Dockerfiles for CMSSW <https://doi.org/10.5281/zenodo.3374807>

82 commits

1 branch

0 packages

Tag: v1.0

New pull request



clelange Use --build-arg instead of wrong -e for docker ENV

1. capture software

archive analysis
code incl. deps.

2. capture commands

what do with the
captured software

3. capture workflow

order of individual
steps

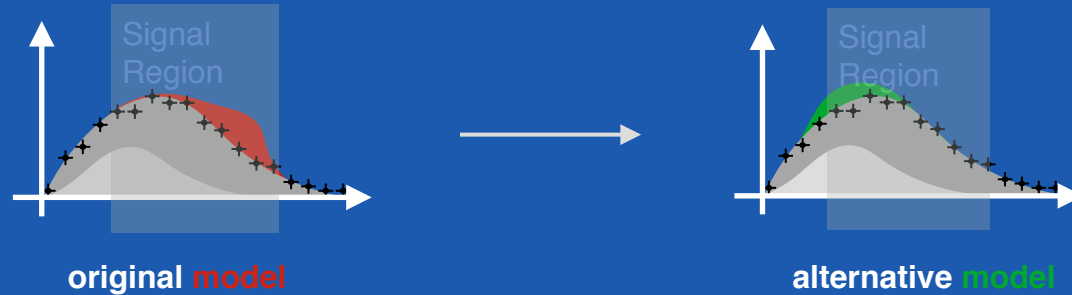
Workflow languages seem to be a good choice:

REANA supports Common Workflow Language, Yadage

- looking into snakemake (LHCb also has snakemake starter kit)

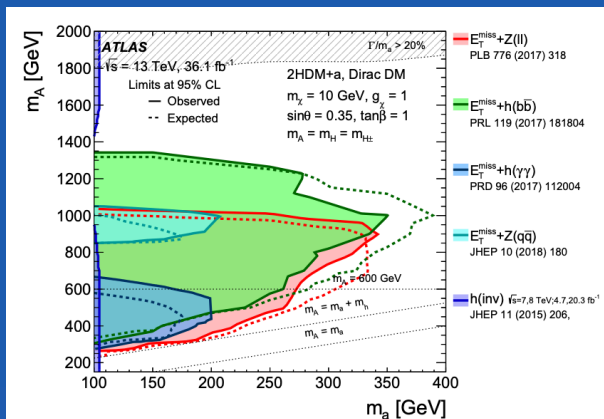
The image displays three overlapping screenshots of the REANA GitHub repository, each showing a different analysis example. The top-left screenshot shows the 'REANA example - ALICE LEGO train test run' page, which includes a 'build passing' badge and a 'license: GPL-2.0' badge. The top-right screenshot shows the 'REANA example - CMS Higgs-to-four-leptons' page, which includes a 'build: unknown' badge and a 'license: MIT' badge. The bottom-center screenshot shows the 'REANA example - ATLAS RECAST' page, which includes a 'build: passing' badge and a 'license: MIT' badge. Each screenshot also shows a 'README.rst' file with an 'About' section and an 'Analysis structure' section. The 'About' section for the ATLAS RECAST example includes a mathematical equation:
$$D_{(s)}^+ \rightarrow \pi^+ \mu^+ \mu^-$$
. The 'Analysis structure' section for the ATLAS RECAST example includes a list of inputs: '1. Input data' and 'The analysis takes the following inputs:'. The 'Input data' section for the ATLAS RECAST example includes a list of inputs: '• dxaodi input ROOT file'.

Major use-case for internal re-use: reinterpretation

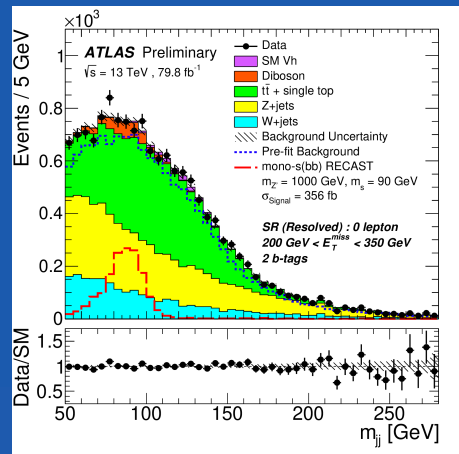


ATLAS: require analyzers to preserve analysis that at least reinterpretation w/ REANA is possible → realization of RECAST (docker images, scripts, workflows)

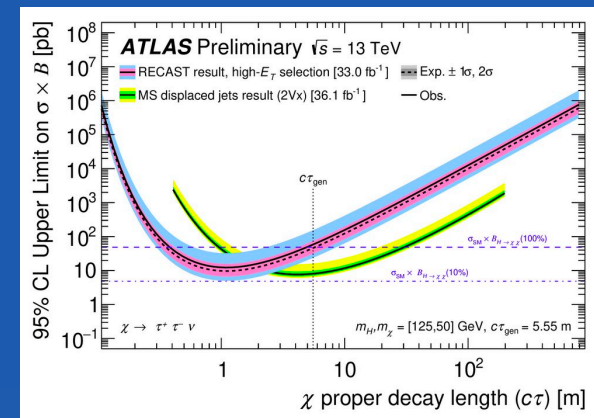
New scientific results based on this (rather technical) requirement



arxiv:1903.01400



ATL-PHYS-PUB-2019-032



ATL-PHYS-PUB-2020-007

CERN Analysis Preservation Examples

CERN Analysis Preservation

BETA

Doc

Files | Data | Source Code

LBZLcDOK.tur.gz (419 KB)

...

BASIC INFORMATION

ANALYSIS NAME

MEASUREMENT

PROPOSERS

NAME

NAME

ORCID

NAME

NAME

STATUS

INSTITUTES INVOLVED

KEYWORDS

STRIPPING/TURBO SELECTIONS

TYPE OF DATASET

CUSTOM NAME

STRIPPING/TURBO LINE

BOOKKEEPING LOCATIONS

TYPE OF DATASET

CUSTOM NAME

STRIPPING/TURBO LINE

BOOKKEEPING LOCATIONS

NTUPLE/USERDEF-PRODUCTION

CUSTOM NAME

INPUT DATASET

PLATFORM

DAVINCI VERSION

OUTPUT EOS LOCATION

LBZLcDOK branching fraction

LBZLcDOK branching fraction

BR(LbZLcD0aK⁰) and BR(LbZLcD0⁰aK⁰) with respect to LBZLcS⁰

Sebastian Neubert

Hartun Tschal

0900-0001-8476-8188

Alexsio Pflüci

Nicola Sedmore

1 - in preparation

8.2

Pentapequarks

real_data

data 2012

XZ(LbD0SDXNPBeauty2CharmLine

AJCHaCaliburn12/BBeam3000GeV-VeloClosed-#MagIs/Real Data/Rec14/5tripping21/1900000000/BHADRON.MDST

AJCHaCaliburn12/BBeam3000GeV-VeloClosed-#MagDown/Real Data/Rec14/5tripping21/1900000000/BHADRON.MDST

real_data

data 2011

XZ(LbD0SDXNPBeauty2CharmLine

AJCHaCaliburn11/BBeam3100GeV-VeloClosed-#MagDown/Real Data/Rec14/5tripping21/1700000000/BHADRON.MDST

AJCHaCaliburn11/BBeam3100GeV-VeloClosed-#MagIs/Real Data/Rec14/5tripping21/1700000000/BHADRON.MDST

2011 samples

data 2011

uB6_64-uB6-qc162-opt

u4x5

Area/0b3wag/Bar04/EXOTICSLbZLcDOK/Impiles/5tripping21

USER ANALYSIS

Copyright 2018 © CERN. Created & Hosted by CERN. Powered by Invenio Software.

[Contact](#)
[About](#)
[Search Tips](#)

About Search Tips

LHCb

16 results		
STATUS		
<input type="checkbox"/> draft	16	JME-10-004
TYPE		
<input type="checkbox"/> cms-analysis-v0.0.1	16	MUON
PHYSICS_OBJECTS		
<input checked="" type="checkbox"/> jet	22	
<input type="checkbox"/> PFMuon	10	
<input type="checkbox"/> GlobalMuon	4	
<input type="checkbox"/> TrackerMuon	4	
<input type="checkbox"/> electron	10	AN-2011/103
<input type="checkbox"/> photon	6	
<input type="checkbox"/> MET	2	ELECTRON MUON
<input type="checkbox"/> tau	2	
<input type="checkbox"/> track	2	
<input type="checkbox"/> vertex	2	AN-2011/062
		MUON
		AN-2011/103
		ELECTRON MUON MET
		AN-2010/411

CMS

Focus on:

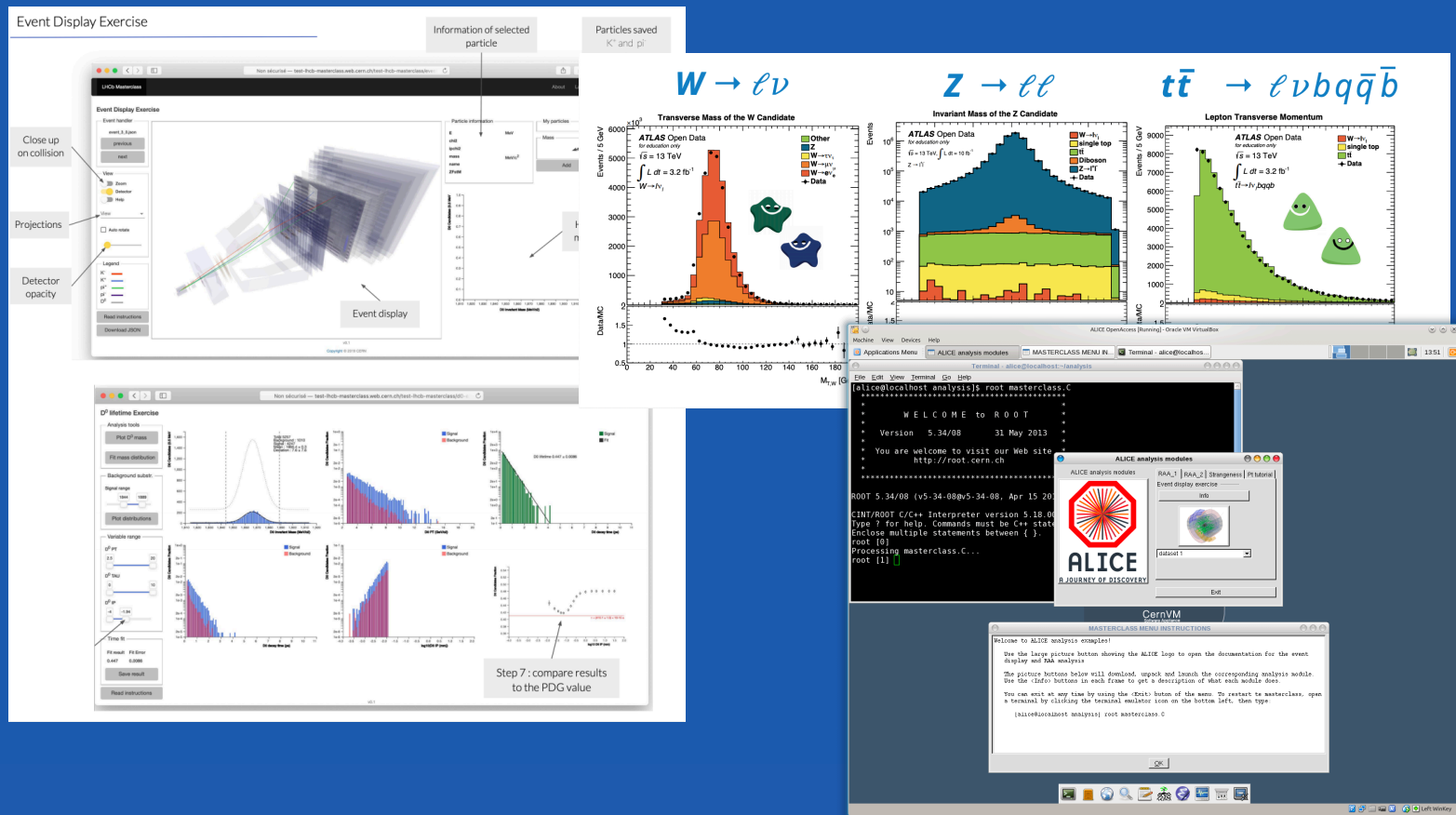
- ease of use for analysis teams
 - e.g. auto-complete, automatic ingestion, command line clients
- ease of use for users
 - discoverability / search
 - integration with REANA



Open Data

All LHC Experiments have Open Data Programs

- integrated into CERN Open Data Portal
- ATLAS, LHCb, ALICE so far focused mainly on Outreach & Education



CMS has more expansive Open Data Program for Research

We see external eco-system developing

- Workshop in October [link]

Number of Papers appearing on e.g.
Machine Learning methods for LHC

Exploring the Space of Jets with CMS Open Data
Patrick T. Komiske^{1,2,*}, Radha Mastandrea^{1,1}, Eric M. Metodiev^{1,2,1}, Preksha Naik^{1,1} and Jesse Thaler^{1,2,1}

arxiv:1908.08542

arxiv:1910.07029

arxiv:1805.00850

End-to-end particle and event identification at the Large Hadron Collider with CMS Open Data

M. Andrews¹, J. Alison¹, S. An^{1,2}, P. Brvant¹, B. Burkle³, S. Glevzer⁴, M. Narain³, M. Paulini¹, B.

¹Department

³Departme

⁴Departme

⁵Machine Learn

Noname manuscript No.
(will be inserted by the editor)

Fast and accurate simulation of particle detectors using generative adversarial networks

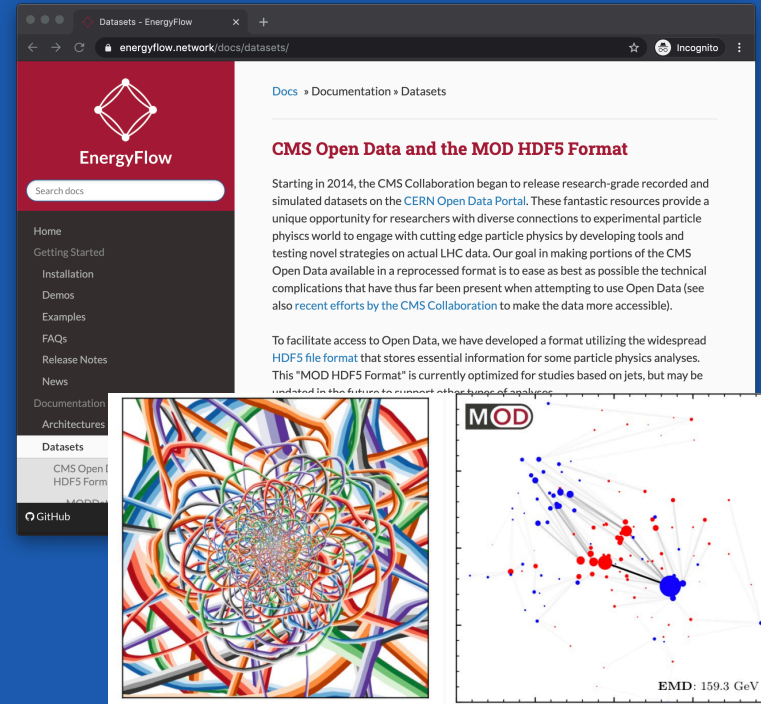
Pasquale Musella · Francesco Pandolfi

26 Nov 2018

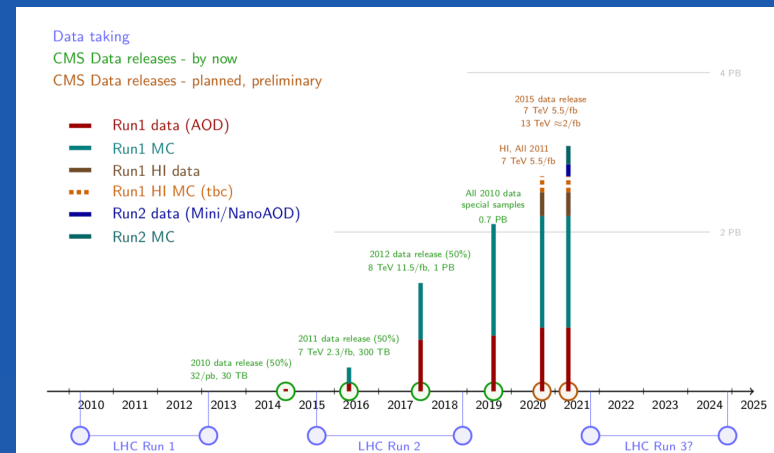
the date of receipt and acceptance should be inserted later

Abstract Deep generative models parametrised by neural networks have recently started to provide accurate results in modeling natural images. In particular, generative adversarial networks provide an unsupervised solution to this problem. In this work we apply this

e.g. [2] and [3]) are based on estimators of particle trajectories and energy deposits. This information is subsequently aggregated in order to reconstruct energy, type and direction of final state particles produced by the collision of the primary beams.



Release Schedule being finalized for coming years



Conclusion:

LHC Experiments have strong analysis and data preservation programs

Technological Progress helps drive fidelity of preservation

- e.g. containers

Cyberinfrastructure for HEP is crucial for adoption

- HepData, CAP, COD, REANA, RECAST

See use of the data products we put out

- Many users of HepData products
- First set of RECAST / internal analysis preservation publications
- Growing number of Open-Data based research
- Strong Outreach & Education based on Open Data