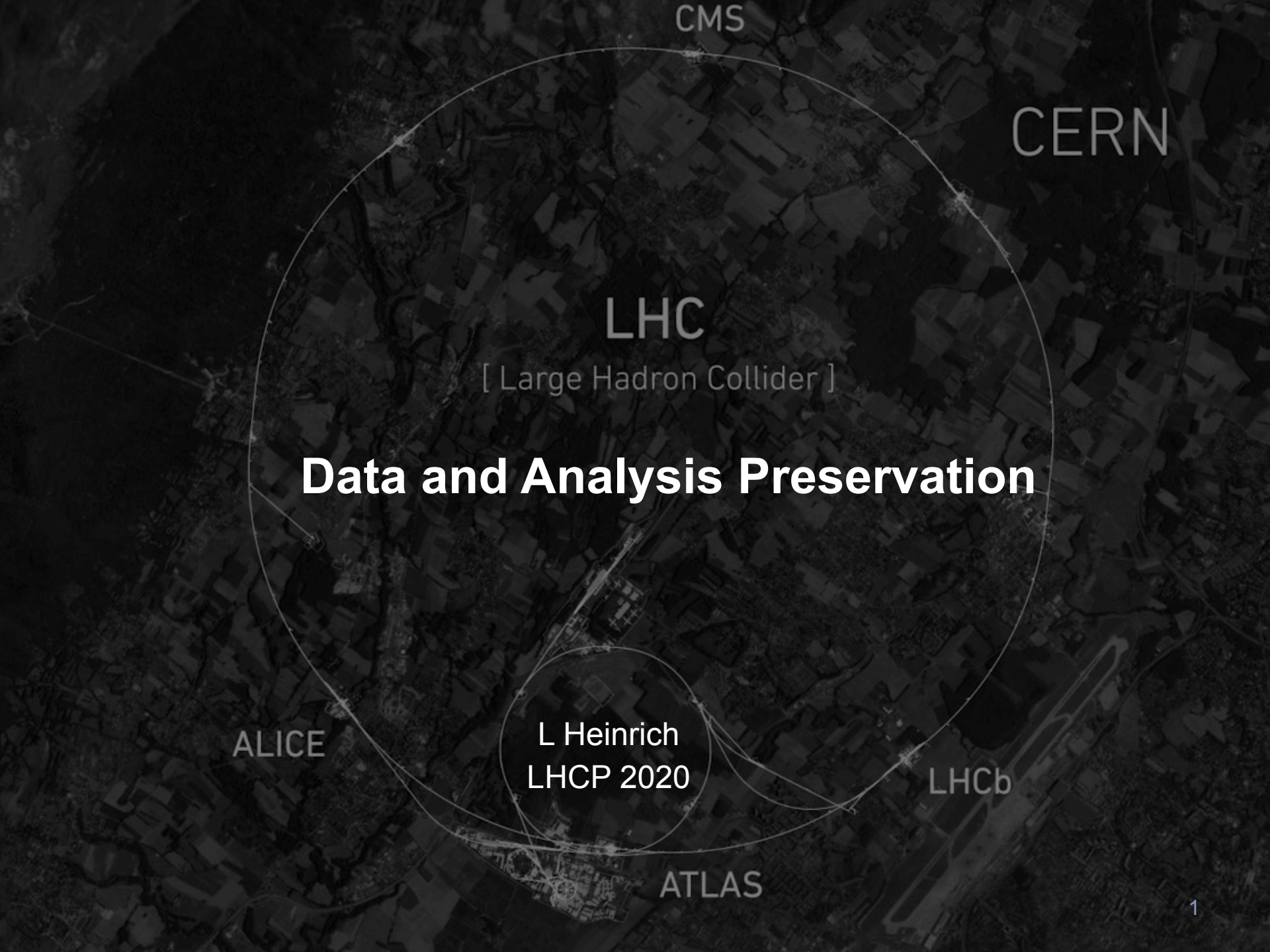
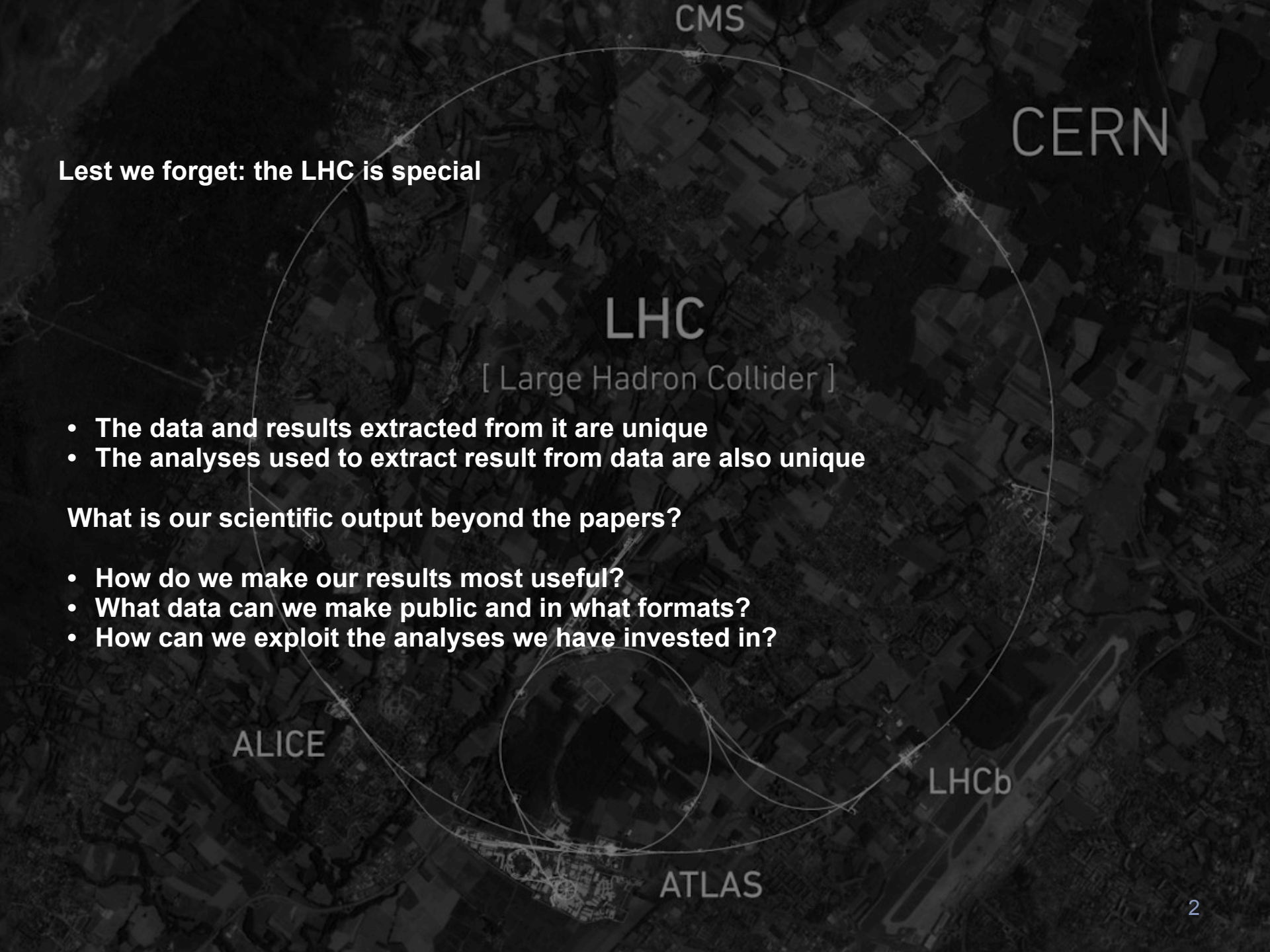


Data and Analysis Preservation

L Heinrich
LHCP 2020



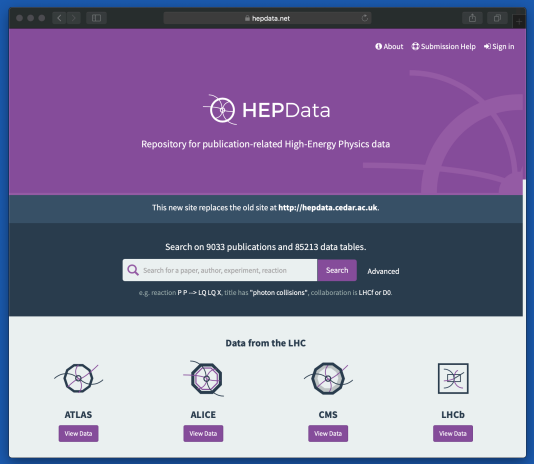
Lest we forget: the LHC is special

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

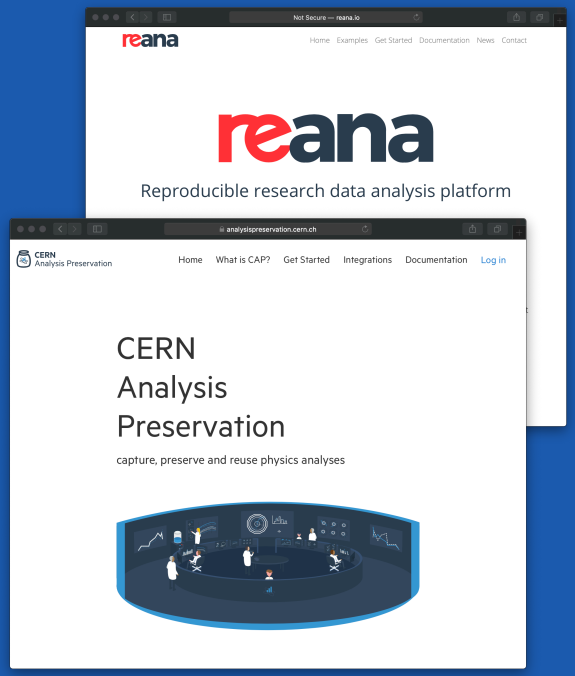
What is our scientific output beyond the papers?

- **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?**

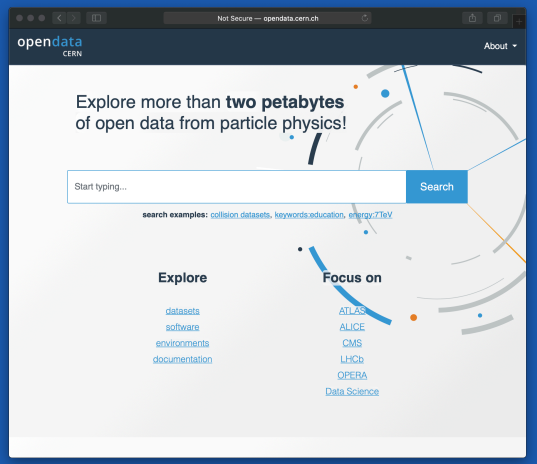
Three broad areas of activity



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

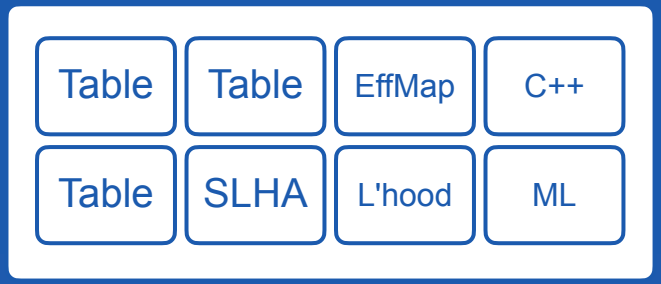
high quality data products for published analyses

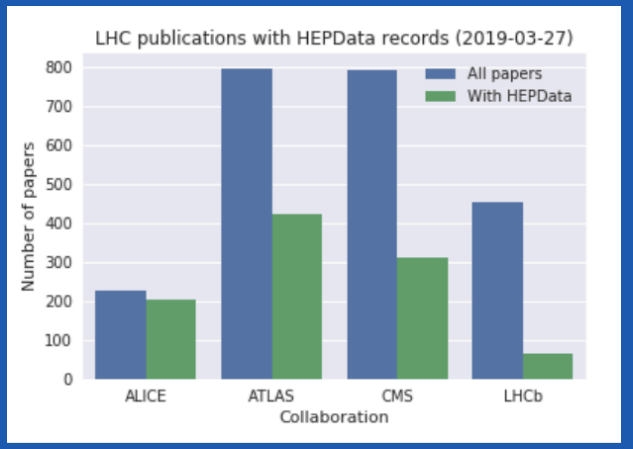
publicly available. All LHC experiments rely on this.

- **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->1l+bb+met analysis (Run2 data)

void EwkOneLeptonTwoBjets2018::Init()

{
  // Define signal/control regions
```

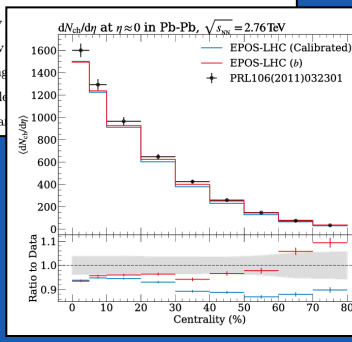
```
// -*- C++ -*-
#include "Rivet/Analysis.hh"
#include "Rivet/Projections/ChargedFinalState.hh"
#include "Rivet/Tools/Correlators.hh"
#include "Rivet/Tools/AliceCommon.hh"
#include "Rivet/Projections/AliceCommon.hh"

namespace Rivet {


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

    /// Constructor
    ALICE_2019_I1723697() :
      CumulantAnalysis("ALICE_2019_I1723697") {}
```

LU TP 20-04
MCNET-20-04

Confronting Experimental Data with Heavy-Ion Models

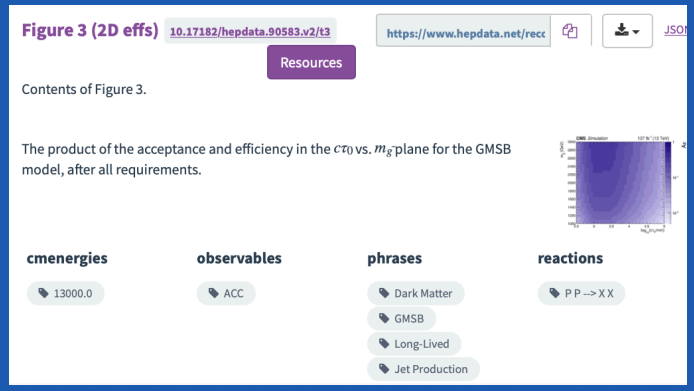
RIVET for Heavy Ions

Christian Bierlich,[1,2] Andy
Peter Harald Lindenov
Jan Fiete Grosse-Oetring
Patrick Kirchgaeßer,[6] Jochen Kl
Christine O. Rasmussen,[2] Ma



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

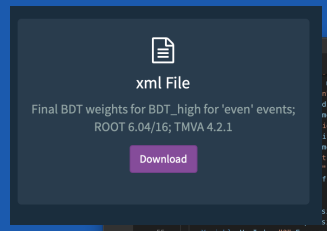
Efficiency Maps:

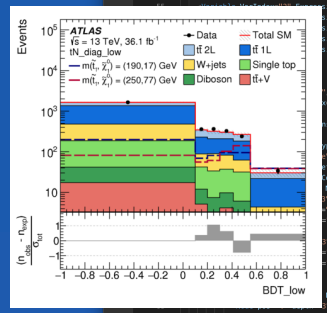


[CMS Record]

ML models uploaded to HepData

[ATLAS Record]



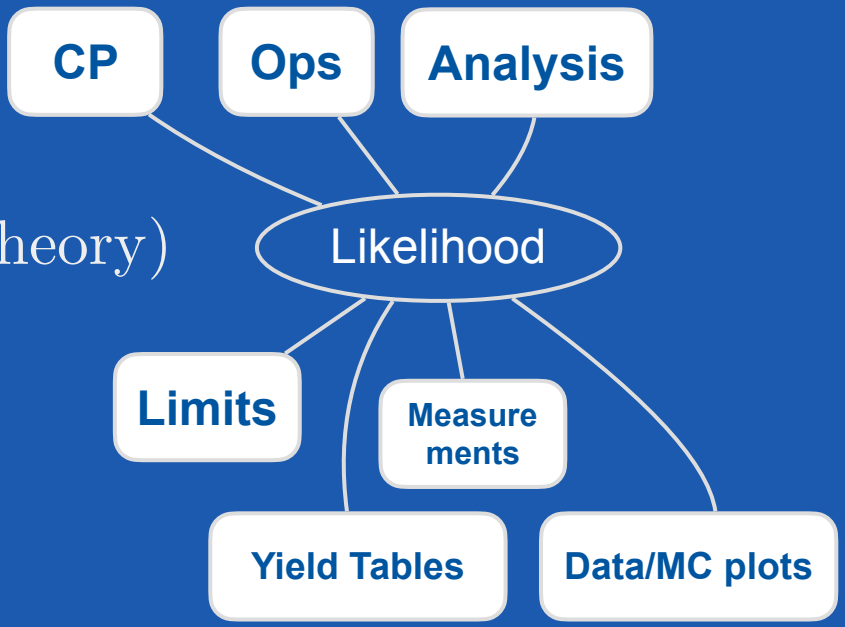


A new Frontier: Public Likelihoods

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

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?

CERN

Long History:

2000: 1st PHYSTAT

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 sh
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

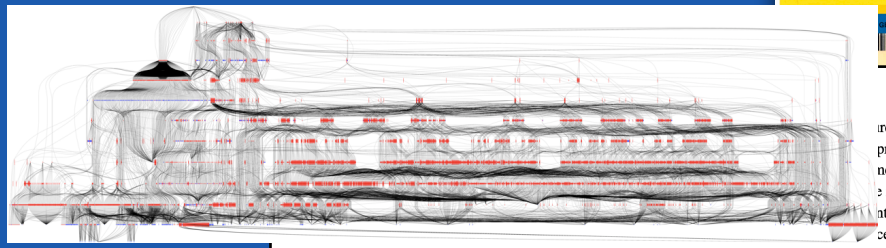
WORKSHOP ON CONFIDENCE LIMITS

CERN, Geneva, Switzerland
17–18 January 2000

PROCEEDINGS

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

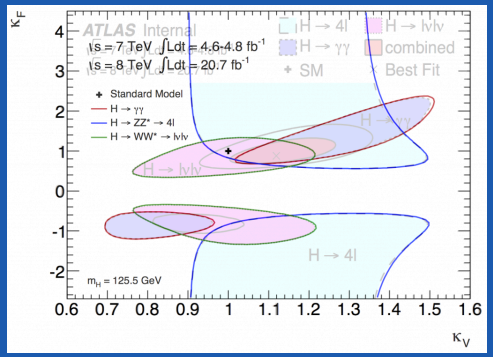
2010: Introduction of Workspaces



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).

2012: ATLAS profile likelihoods



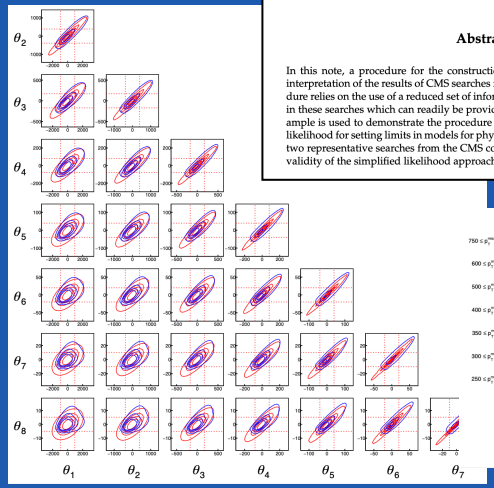
2017: CMS Simplified Likelihoods

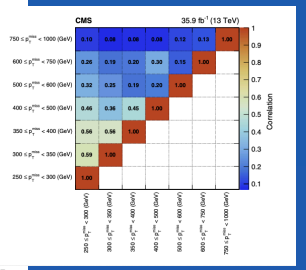
Simplified likelihood for the re-interpretation of public CMS results

The CMS Collaboration

Abstract

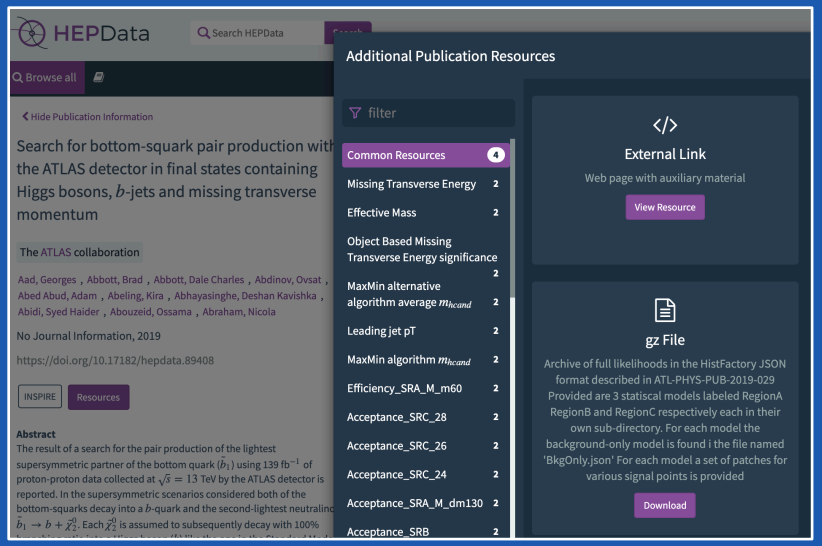
In this note, a procedure for the construction of simplified likelihoods for the re-interpretation of the results of CMS searches for new physics is presented. The procedure relies on the use of a reduced set of information on the background models used in these searches which can readily be provided 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.



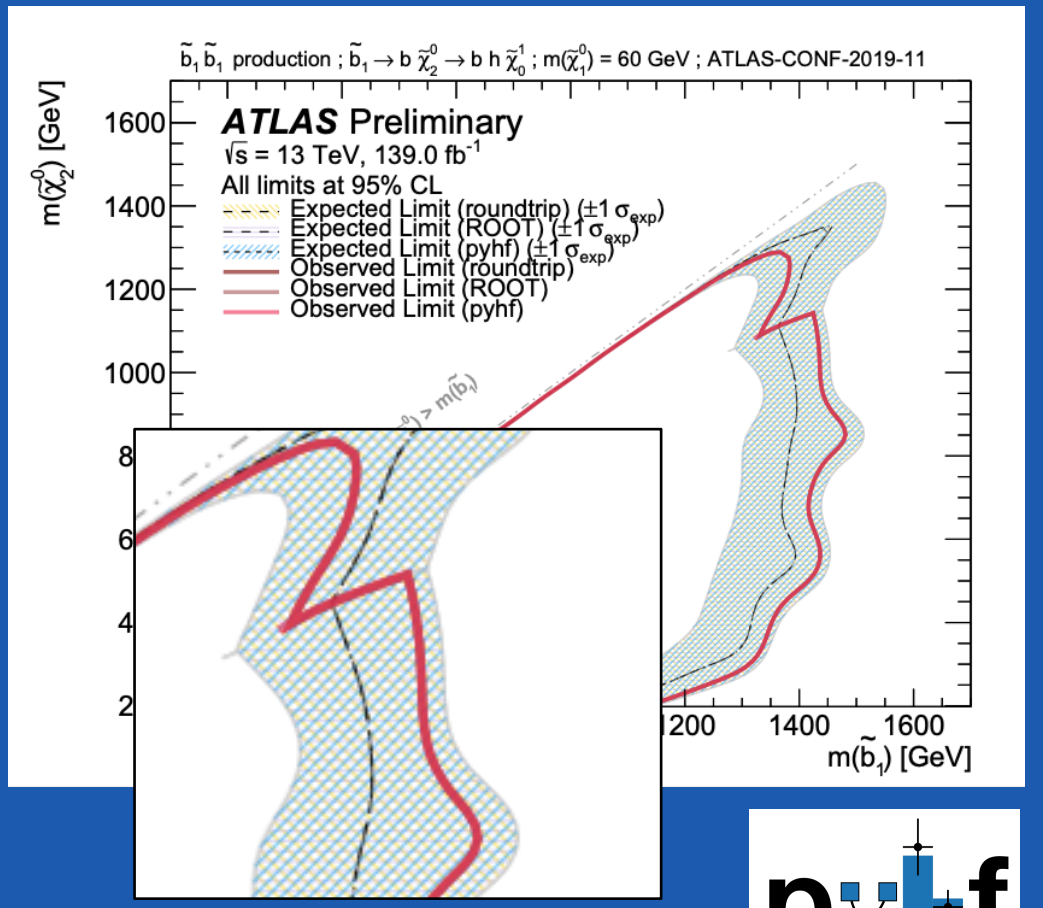


2019: first full likelihood release

- using JSON Schema of HistFactory class of Likelihoods




```
$> 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
}
```



pyhf differentiable Likelihoods

CERN

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

reana

Reproducible research data analysis platform

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

CERN

Analysis Preservation

capture, preserve and reuse physics analyses

CERN

1. **capture software**

   archive analysis code incl. deps.

2. **capture commands**

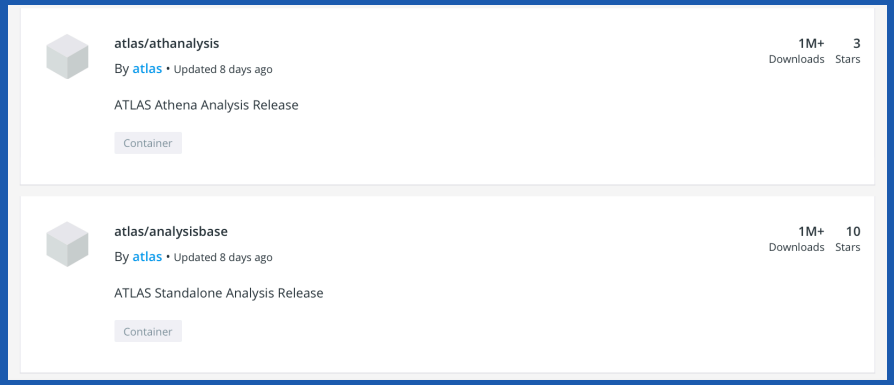
   what do with the captured software

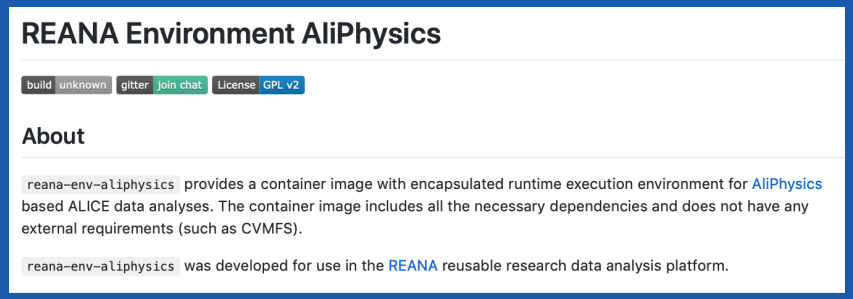
3. **capture workflow**

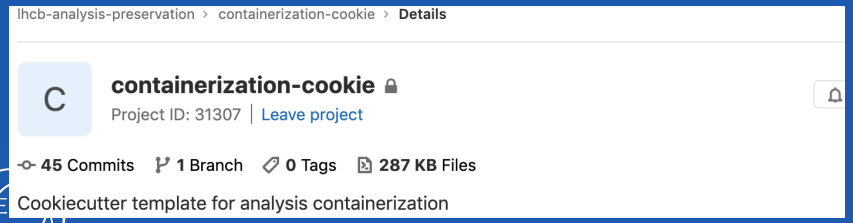
   order of individual steps

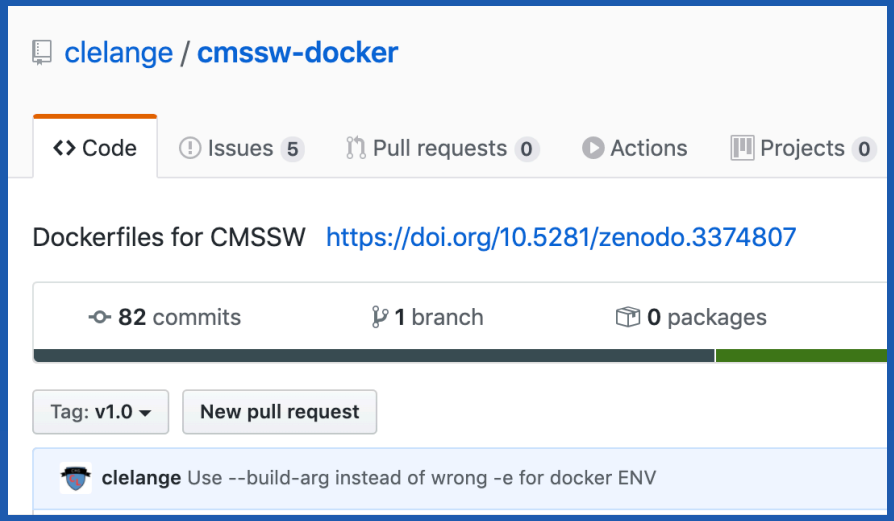
Containers universally seen as suitable techology:

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









1. **capture software**

   archive analysis code incl. deps.

2. **capture commands**

   what do with the captured software

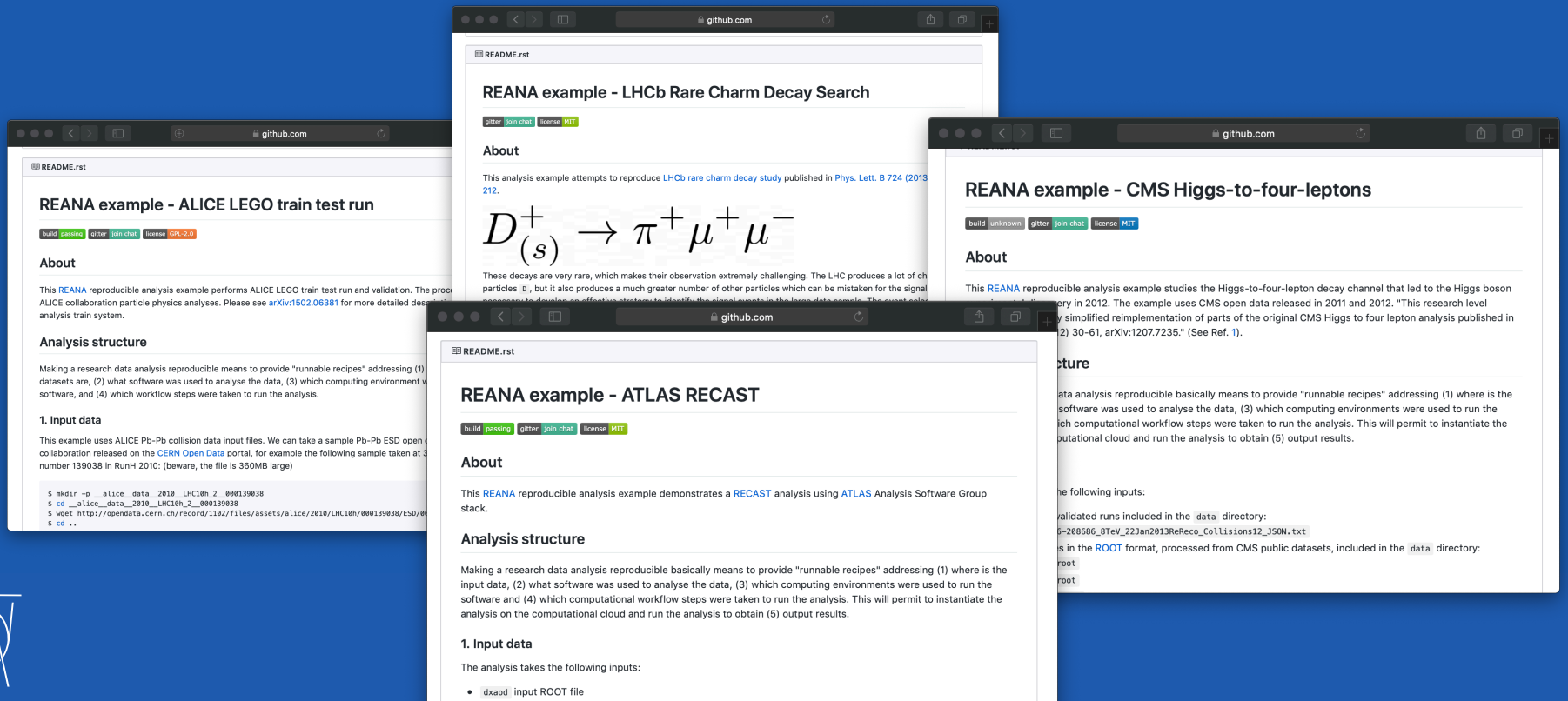
3. **capture workflow**

   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)

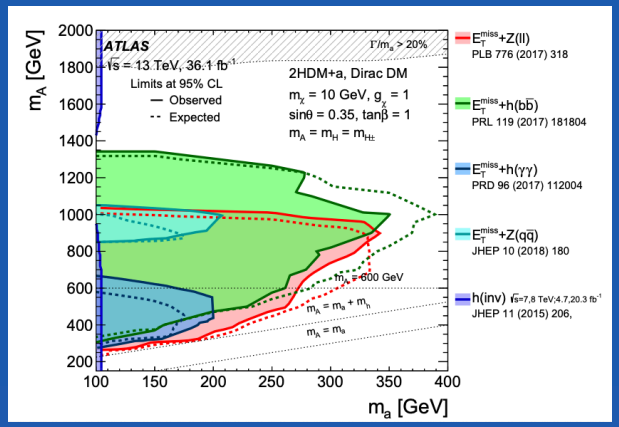


Major use-case for internal re-use: reinterpretation

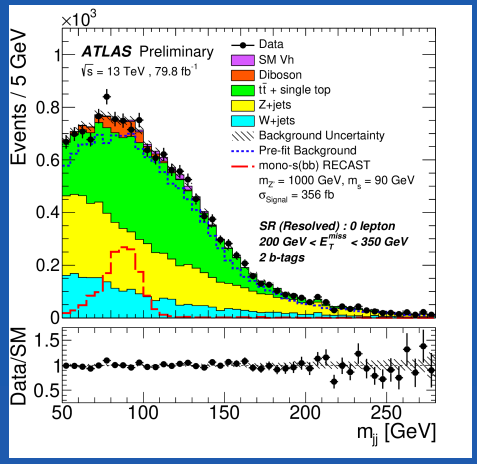
original model alternative model

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

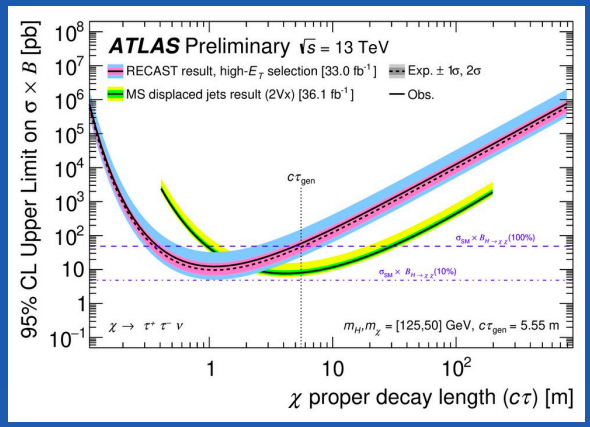
New scientific results based on this (rather technical) requirement



arxiv:1903.01400

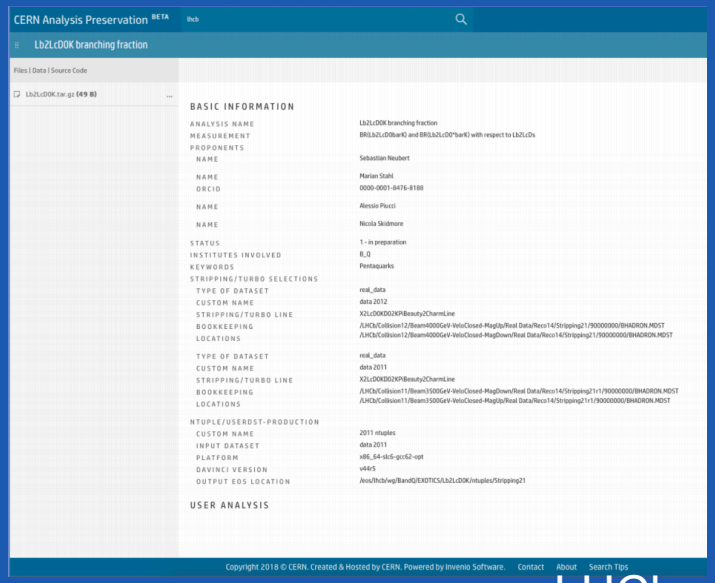


ATL-PHYS-PUB-2019-032

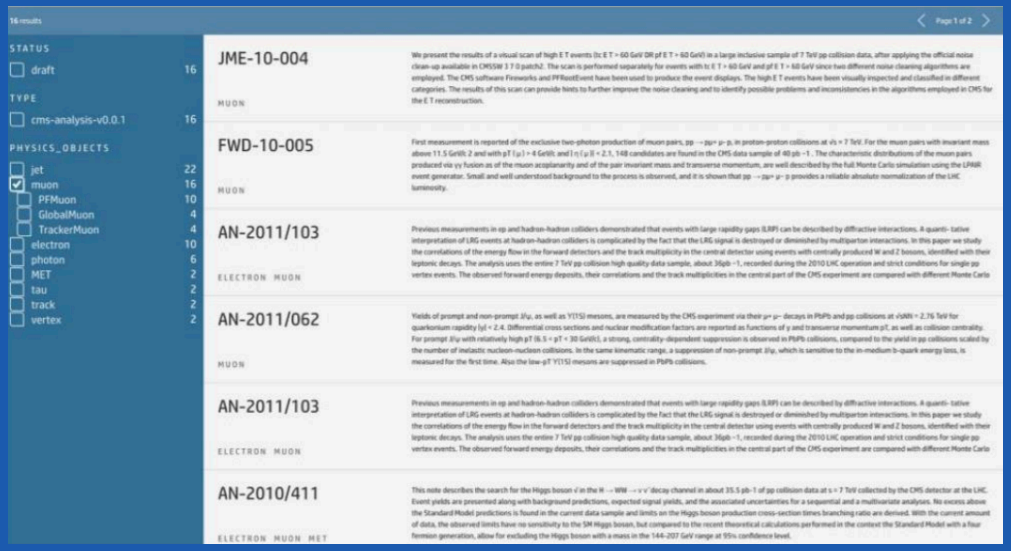


ATL-PHYS-PUB-2020-007

CERN Analysis Preservation Examples



LHCb



CMS

Focus on:

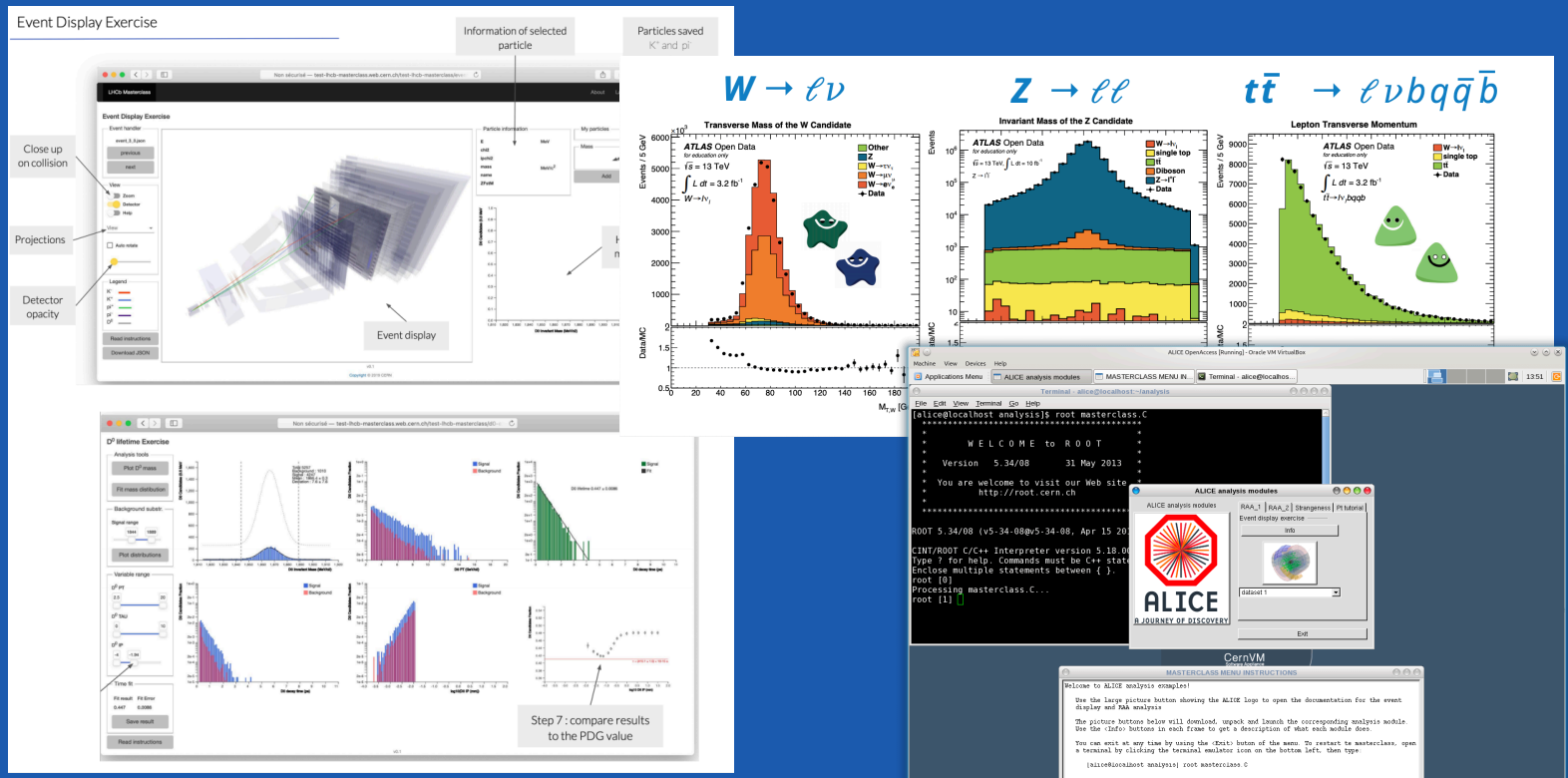
- **ease of use for analysis teams**
  e.g. auto-complete, automatic ingestion, command line clients

- **ease of use for users**
  - **discoverability / search**
  - **integration with REANA**

Open Data

All LHC Experiments have Open Data Programs

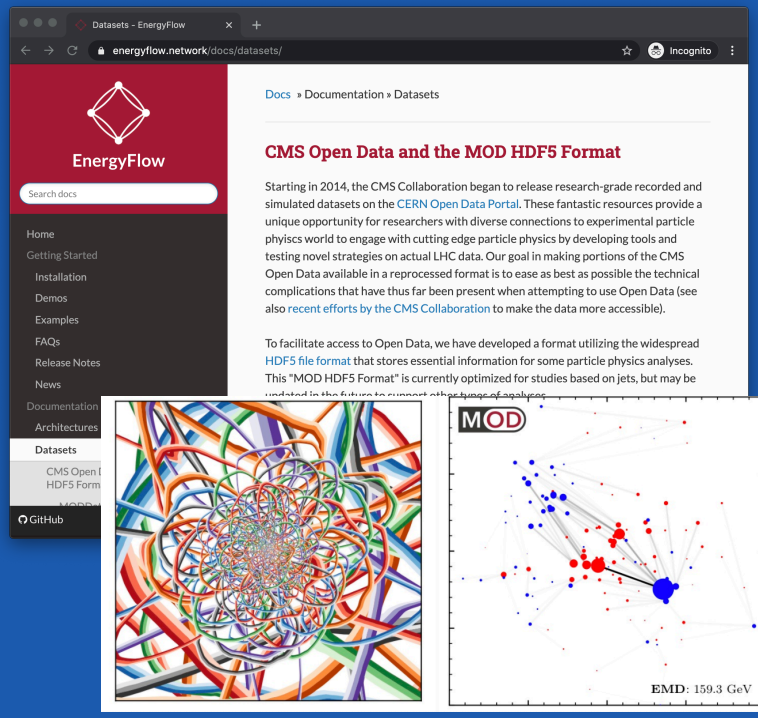
- integrated into CERN Open Data Portal
- ATLAS, LHCb, ALICE so far focused mainly on Outreach & Education



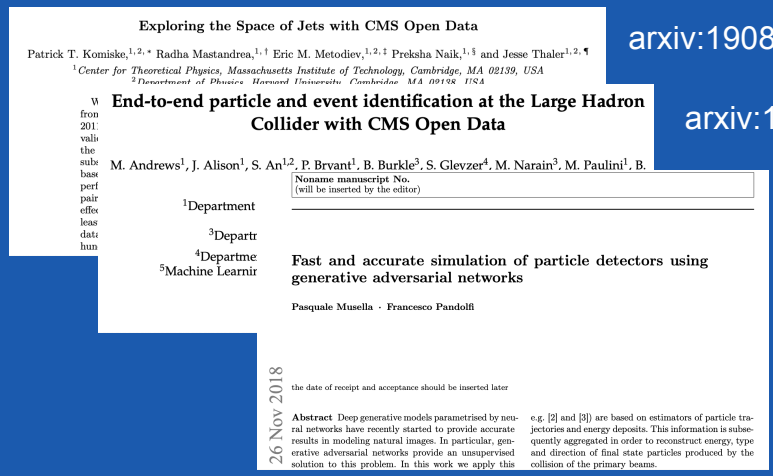
CMS has more expansive Open Data Program for Research

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,†] Eric M. Metodiev,[1,2,‡] Preksha Naik,[1,§] and Jesse Thaler[1,2,¶]

[1]Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Department of Physics, Harvard University, Cambridge, MA 02138, USA

arxiv:1908.08542

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

M. Andrews[1], J. Alison[1], S. An[1,2], P. Bryant[1], B. Burkle[3], S. Glevzer[4], M. Narain[3], M. Paulini[1], B.

[1]Department

[3]Departr

[4]Departme

[5]Machine Learnir

arxiv:1910.07029`

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

Fast and accurate simulation of particle detectors using generative adversarial networks

Pasquale Musella · Francesco Pandolfi

the date of receipt and acceptance should be inserted later

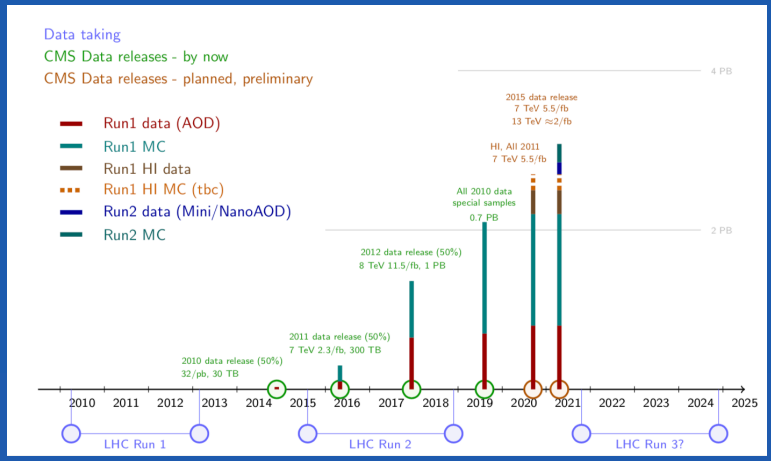
Abstract Deep generative models parametrised by neural networks have recently started to provide accurate results in modeling natural images. In particular, generative adversarial networks provide an unsupervised solution to this problem. In this work we apply this

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.

26 Nov 2018

arxiv:1805.00850`

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

CERN