ATLAS Plans & Status

Lukas Heinrich, Marumi Kado
on behalf of the ATLAS

Open Science Strategy Workshop 2021

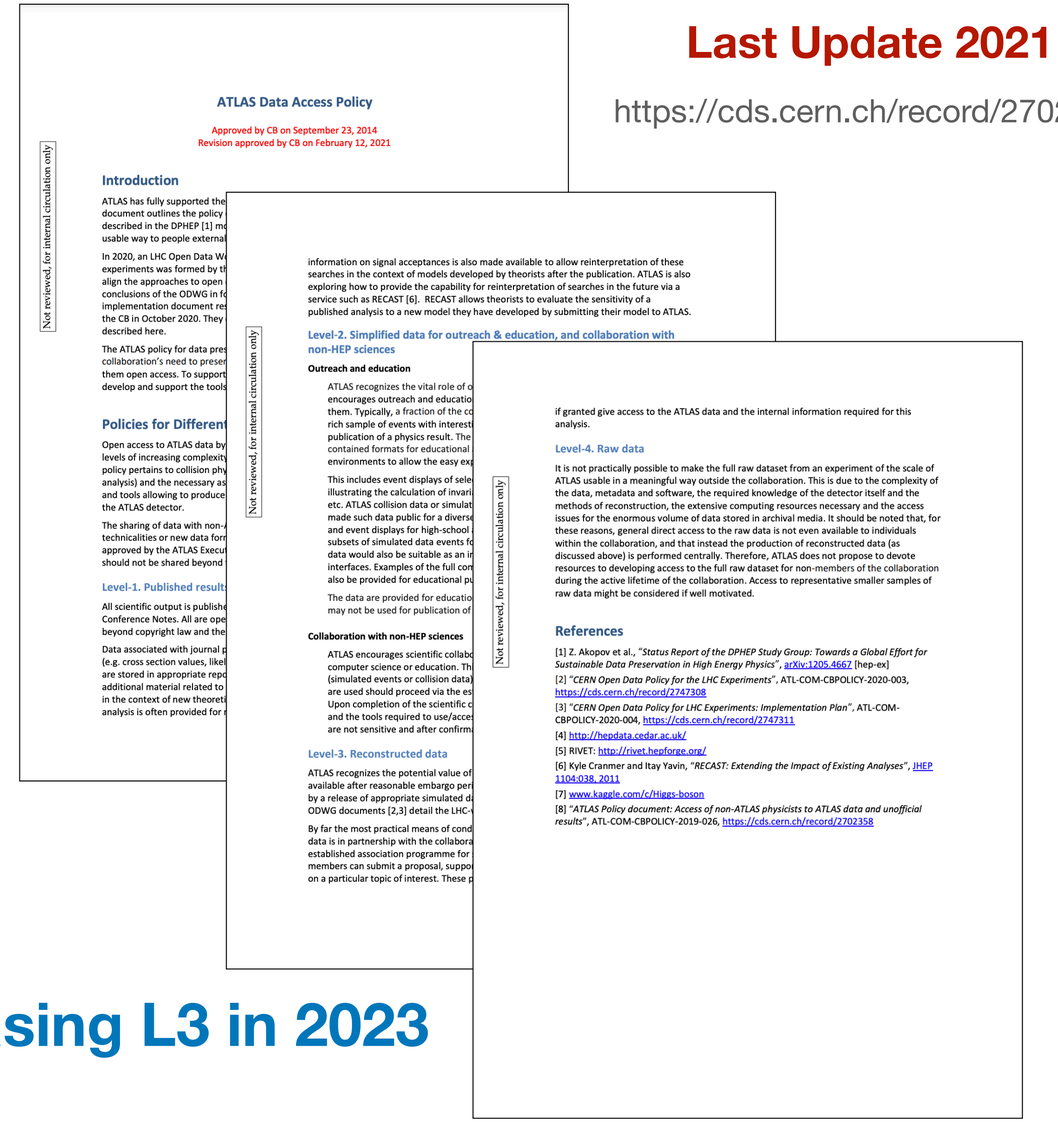
Policy Documents: Access

Data Access in ATLAS organized along DPHEP Access Levels defined in 2015 Status Report (link)

Level 1	Publication
Level 2	Education & Outreach
Level 3	Reconstruction & Research
Level 4	Raw Data

Last Update 2021

https://cds.cern.ch/record/2702390



ATLAS Data Access Policy

Approved by CB on September 23, 2014
Revision approved by CB on February 12, 2021

Not reviewed, for internal circulation only

Introduction

ATLAS has fully supported the
document outlines the policy
described in the DPHEP [1] mo
usable way to people external

In 2020, an LHC Open Data W
experiments was formed by th
align the approaches to open
conclusions of the ODWG in fo
implementation document re
the CB in October 2020. They
described here.

The ATLAS policy for data pres
collaboration's need to preser
them open access. To support
develop and support the tools

Policies for Different

Open access to ATLAS data by
levels of increasing complexity
policy pertains to collision phy
analysis) and the necessary as
and tools allowing to produce
the ATLAS detector.

The sharing of data with non-
technicalities or new data form
approved by the ATLAS Execu
should not be shared beyond

Level-1. Published result

All scientific output is publishe
Conference Notes. All are ope
beyond copyright law and the

Data associated with journal p
(e.g. cross section values, likel
are stored in appropriate repo
additional material related to
in the context of new theoreti
analysis is often provided for

Not reviewed, for internal circulation only

information on signal acceptances is also made available to allow reinterpretation of these searches in the context of models developed by theorists after the publication. ATLAS is also exploring how to provide the capability for reinterpretation of searches in the future via a service such as RECAST [6]. RECAST allows theorists to evaluate the sensitivity of a published analysis to a new model they have developed by submitting their model to ATLAS.

Level-2. Simplified data for outreach & education, and collaboration with non-HEP sciences

Outreach and education

ATLAS recognizes the vital role of o
encourages outreach and educatio
them. Typically, a fraction of the co
rich sample of events with interest
publication of a physics result. The
contained formats for educational
environments to allow the easy ex

This includes event displays of sele
illustrating the calculation of invari
etc. ATLAS collision data or simulat
made such data public for a diverse
and event displays for high-school
subsets of simulated data events fo
data would also be suitable as an i
interfaces. Examples of the full cor
also be provided for educational pu

The data are provided for educatio
may not be used for publication of

Collaboration with non-HEP sciences

ATLAS encourages scientific collabo
computer science or education. Th
(simulated events or collision data)
are used should proceed via the es
Upon completion of the scientific c
and the tools required to use/acce
are not sensitive and after confirma

Level-3. Reconstructed data

ATLAS recognizes the potential value of
available after reasonable embargo per
by a release of appropriate simulated da
ODWG documents [2,3] detail the LHC-

By far the most practical means of cond
data is in partnership with the collabora
established association programme for
members can submit a proposal, suppo
on a particular topic of interest. These p

Not reviewed, for internal circulation only

if granted give access to the ATLAS data and the internal information required for this analysis.

Level-4. Raw data

It is not practically possible to make the full raw dataset from an experiment of the scale of ATLAS usable in a meaningful way outside the collaboration. This is due to the complexity of the data, metadata and software, the required knowledge of the detector itself and the methods of reconstruction, the extensive computing resources necessary and the access issues for the enormous volume of data stored in archival media. It should be noted that, for these reasons, general direct access to the raw data is not even available to individuals within the collaboration, and that instead the production of reconstructed data (as discussed above) is performed centrally. Therefore, ATLAS does not propose to devote resources to developing access to the full raw dataset for non-members of the collaboration during the active lifetime of the collaboration. Access to representative smaller samples of raw data might be considered if well motivated.

References

[1] Z. Akopov et al., *"Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics"*, arXiv:1205.4667 [hep-ex]

[2] *"CERN Open Data Policy for the LHC Experiments"*, ATL-COM-CBPOLICY-2020-003, https://cds.cern.ch/record/2747308

[3] *"CERN Open Data Policy for LHC Experiments: Implementation Plan"*, ATL-COM-CBPOLICY-2020-004, https://cds.cern.ch/record/2747311

[4] http://hepdata.cedar.ac.uk/

[5] RIVET: http://rivet.hepforge.org/

[6] Kyle Cranmer and Itay Yavin, *"RECAST: Extending the Impact of Existing Analyses"*, JHEP 1104:038, 2011

[7] www.kaggle.com/c/Higgs-boson

[8] *"ATLAS Policy document: Access of non-ATLAS physicists to ATLAS data and unofficial results"*, ATL-COM-CBPOLICY-2019-026, https://cds.cern.ch/record/2702358

- **As all LHC experiments: L4 not released**
- **Long & successful experience with L1/L2**
- **New since ODWG: ATLAS plant to start releasing L3 in 2023**

Policy Documents: Access

Level 3: ATLAS participated in new CERN Open Data Working group under DRC to develop common LHC-wide policy on Level 3 Data

Two Documents:

CERN Policy: begin to release data & software needed for robust analysis 5 years after close of a run

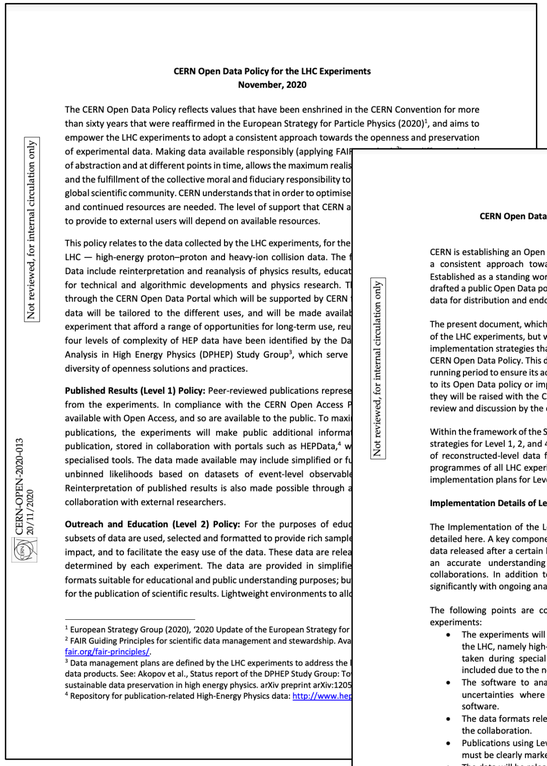
ATLAS Implementation (v. similar to CMS):

- release 25% of data after 5 years
- release 50% of data after 10 years
- remainder at close of experiment
- max 20% of integ. lumi. as Open Data

CERN Policy

ATLAS
CMS
LHCb
ALICE
TOTEM

public internal



CERN Open Data Policy for the LHC Experiments
November, 2020

Not reviewed, for internal circulation only

CERN-OPEN-2020-013
20/11/2020

The CERN Open Data Policy reflects values that have been enshrined in the CERN Convention for more than sixty years that were reaffirmed in the European Strategy for Particle Physics (2020)[1], and aims to empower the LHC experiments to adopt a consistent approach towards the openness and preservation of experimental data. Making data available responsibly (applying FAI
of abstraction and at different points in time, allows the maximum realis
and the fulfillment of the collective moral and fiduciary responsibility to
global scientific community. CERN understands that in order to optimise
and continued resources are needed. The level of support that CERN a
to provide to external users will depend on available resources.

This policy relates to the data collected by the LHC experiments, for the
LHC — high-energy proton–proton and heavy-ion collision data. The
Data include reinterpretation and reanalysis of physics results, educat
for technical and algorithmic developments and physics research. T
through the CERN Open Data Portal which will be supported by CERN
data will be tailored to the different uses, and will be made availab
experiment that afford a range of opportunities for long-term use, reu
four levels of complexity of HEP data have been identified by the Da
Analysis in High Energy Physics (DPHEP) Study Group[3], which serve
diversity of openness solutions and practices.

Published Results (Level 1) Policy: Peer-reviewed publications represe
from the experiments. In compliance with the CERN Open Access P
available with Open Access, and so are available to the public. To maxi
publications, the experiments will make public additional informa
publication, stored in collaboration with portals such as HEPData,[4] w
specialised tools. The data made available may include simplified or fu
unbinned likelihoods based on datasets of event-level observable
Reinterpretation of published results is also made possible through a
collaboration with external researchers.

Outreach and Education (Level 2) Policy: For the purposes of educ
subsets of data are used, selected and formatted to provide rich sample
impact, and to facilitate the easy use of the data. These data are relea
determined by each experiment. The data are provided in simplifie
formats suitable for educational and public understanding purposes; bu
for the publication of scientific results. Lightweight environments to allo

[1] European Strategy Group (2020), '2020 Update of the European Strategy for
[2] FAIR Guiding Principles for scientific data management and stewardship. Ava
fair.org/fair-principles/.
[3] Data management plans are defined by the LHC experiments to address the
data products. See: Akopov et al., Status report of the DPHEP Study Group: To
sustainable data preservation in high energy physics. arXiv preprint arXiv:1205
[4] Repository for publication-related High-Energy Physics data: http://www.hep

CERN Open Data Policy for LHC Experiments: Implementation Plan
November 2020

Not reviewed, for internal circulation only

CERN is establishing an Open Data Policy, which aims to empower LHC experiments to adopt a consistent approach towards the openness and preservation of experimental data. Established as a standing working group, the CERN Open Data Working Group (ODWG)[1] has drafted a public Open Data policy document outlining these approaches for LHC experimental data for distribution and endorsement by the Collaboration Boards of the LHC experiments.

The present document, which is also distributed to and endorsed by the Collaboration Boards of the LHC experiments, but which remains internal to the LHC collaborations[2], describes the implementation strategies that the LHC experiments will follow in compliance with the public CERN Open Data Policy. This document will be reviewed by the ODWG at the end of each LHC running period to ensure its accuracy. If an experiment wishes to make any significant changes to its Open Data policy or implementation strategy as described in the relevant documents, they will be raised with the CERN Director of Research and Computing, who would facilitate review and discussion by the other LHC experiments under the auspices of the ODWG.

Within the framework of the Study Group Levels of HEP Data [1], all experiments follow similar strategies for Level 1, 2, and 4, while the release strategy for Level 3 data varies. The release of reconstructed-level data for scientific analysis by one experiment affects the scientific programmes of all LHC experiments and must therefore be coordinated. The corresponding implementation plans for Level 3 Open Data are therefore detailed in the following.

Implementation Details of Level 3 Open Data

The Implementation of the Level 3 Open Data policy for the four large LHC experiments is detailed here. A key component of the implementation strategy is controlling the fraction of data released after a certain latency period, which is important to ensure sufficient time for an accurate understanding and the scientific exploitation of the data within the collaborations. In addition to this latency, data may be withheld if a release interferes significantly with ongoing analyses.

The following points are common for Level 3 Open Data implementations across the experiments:

- The experiments will release data associated with the main physics programmes of the LHC, namely high-energy proton–proton collisions, and heavy-ion collisions. Data taken during special runs with non-standard beam configurations might not be included due to the need for specialised treatment of such data.
- The software to analyse the data (including application of the main systematic uncertainties where practical) will be made publicly available as open-source software.
- The data formats released will be the same as those used for internal analysis within the collaboration.
- Publications using Level 3 Open Data will not be reviewed by the collaboration, and must be clearly marked as being independent of the collaboration.
- The data will be released through the CERN Open Data Portal [2].

[1] The current composition of this working group can be seen in Appendix 1.
[2] This document is intended to be internal to the LHC Experiments (ALICE, ATLAS, CMS, LHCb and TOTEM), as well as relevant groups in the CERN Research and Computing Sector (IT and RCS-SIS).

1

Policy Documents: Access

ATLAS Policy allows for special purpose datasets to be prepared & released for research policies: accompanied with publication as explanation

Both Bulk Level-3 data and Special Purpose datasets are/will be released on

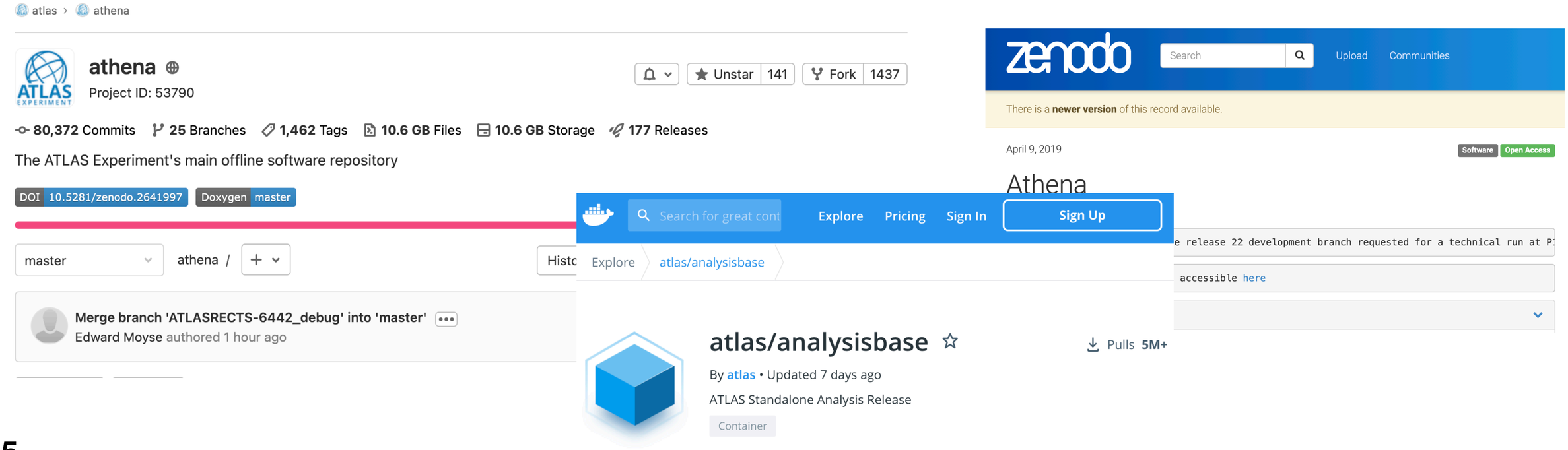
CERN Open Data Portal

→ **crucial service for us for a long-term OD program**
→ **first datasets under new policy released**
→ **starting to preparer for bulk release in 2023**

Policy Documents: Access

Focus of ATLAS Level-3 approach:

- **release format with best-known calibration (pre-processing): PHYSLITE**
- **release software that allows full systematic analysis as Open Source**
- **we release data & tools we use internally to enable best possble analysis**
  responsibility of users: make use of the available resources



Likelihoods - the best Information we have

Likelihood of the Measurement

The best, most precise, most reusable data product for existing data we produce at the LHC

Massimo Corradi

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

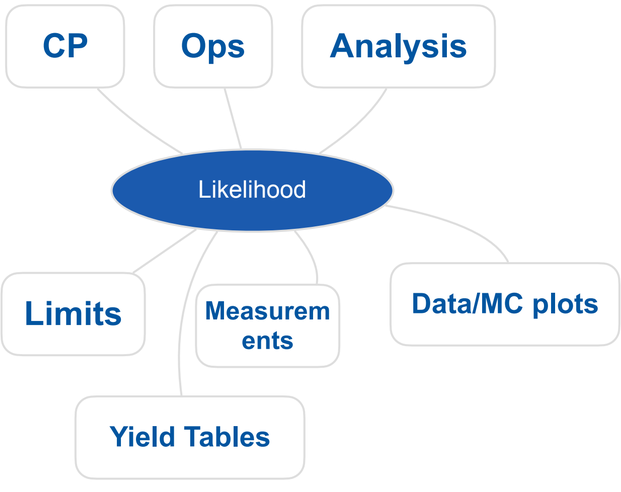
Louis Lyons

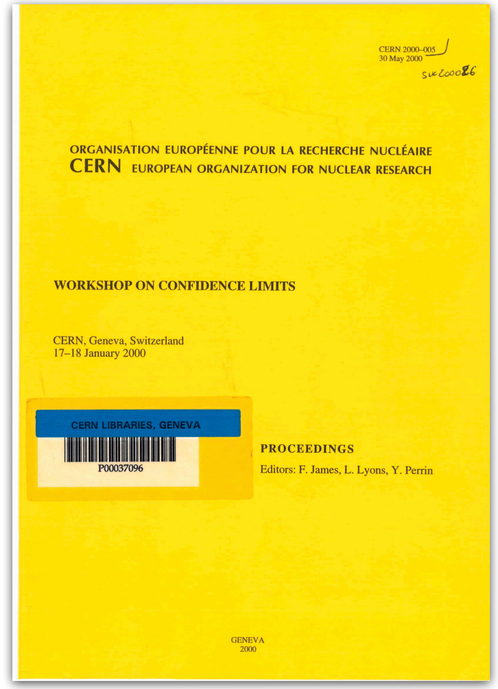
Any disagreement ? Carried unanimously. That's actually quite an achievement for this Workshop.

$$p(\text{theory}\,|\,\text{data}) = \frac{p(\text{data}\,|\,\text{theory})\,p(\text{theory})}{p(\text{data})}$$

- **identified 20 years ago as key Open Science component**

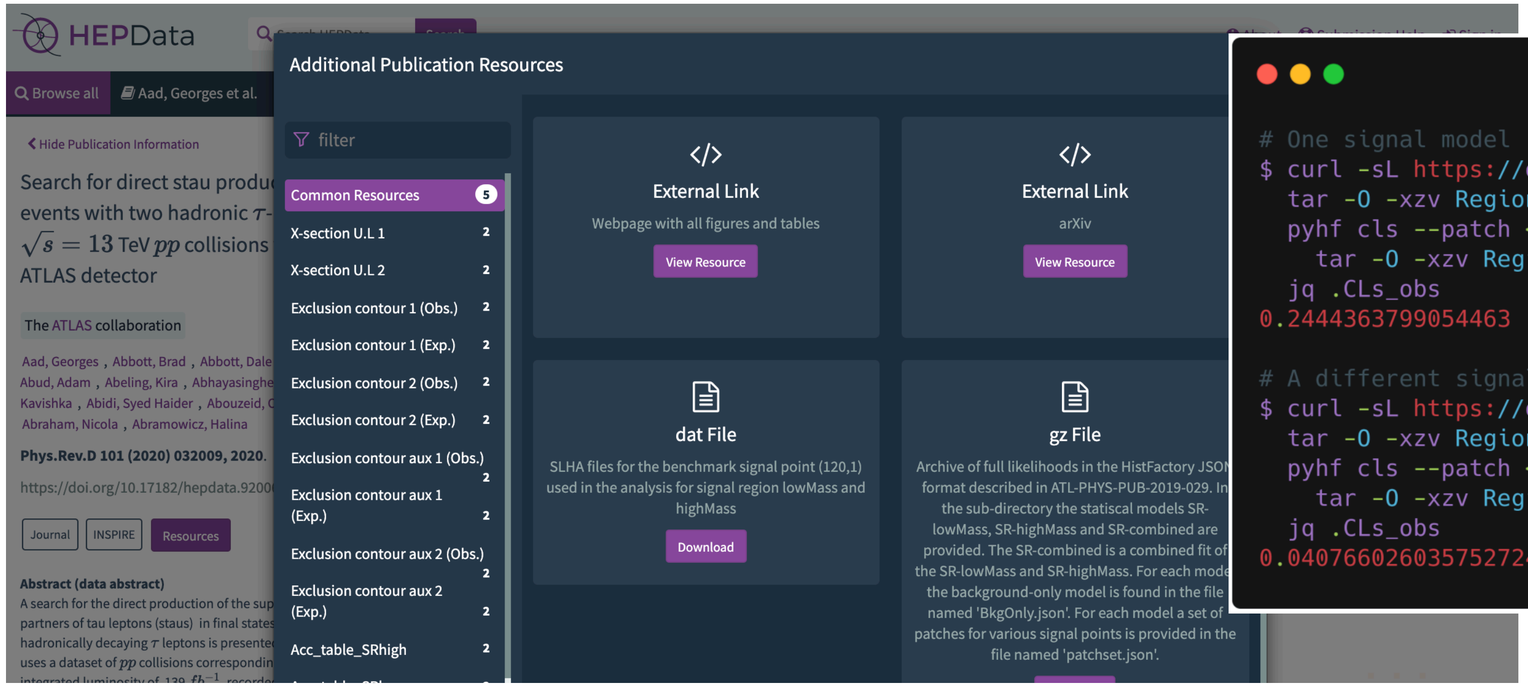
- **encodes almost all details of the published result (systematics, correlations, ...)**
- **enables later re-use by combination (global fits)**
- **directly usable by researches at a high level (complements L3)**





Likelihoods - the best Information we have

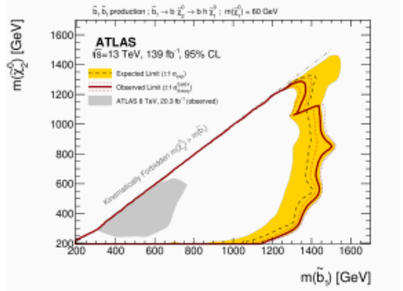
ATLAS releases likelihood as citable, resuable data items



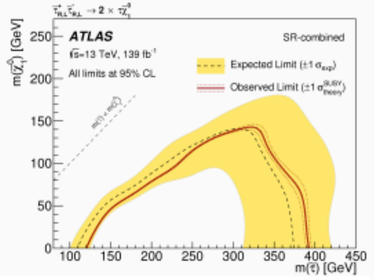

```
# One signal model
$ curl -sL https://doi.org/10.17182/hepdata.89408.v1/r2 | \
  tar -O -xzv RegionA/BkgOnly.json | \
  pyhf cls --patch <(curl -sL https://doi.org/10.17182/hepdata.89408.v1/r2 | \
    tar -O -xzv RegionA/patch.sbottom_1300_205_60.json) | \
  jq .CLs_obs
0.2444363799054463

# A different signal model
$ curl -sL https://doi.org/10.17182/hepdata.89408.v1/r2 | \
  tar -O -xzv RegionA/BkgOnly.json | \
  pyhf cls --patch <(curl -sL https://doi.org/10.17182/hepdata.89408.v1/r2 | \
    tar -O -xzv RegionA/patch.sbottom_1300_230_100.json) | \
  jq .CLs_obs
0.040766026035752724
```

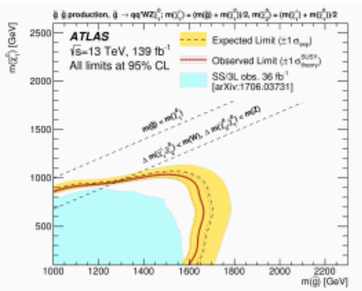
DOI 10.17182/hepdata.89408.v2



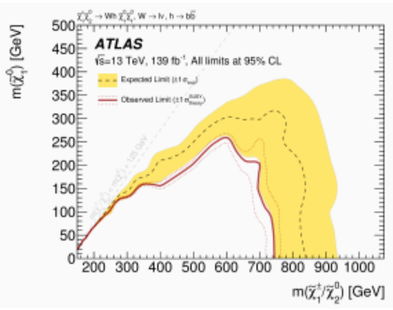
DOI 10.17182/hepdata.92006.v2



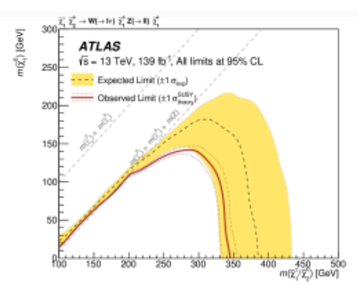
DOI 10.17182/hepdata.91214.v3



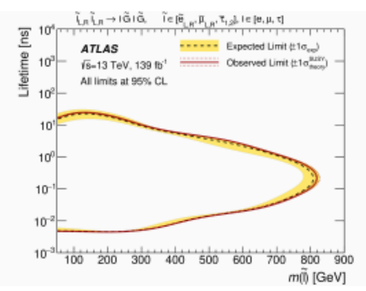
DOI 10.17182/hepdata.90607.v3



DOI 10.17182/hepdata.91127.v2



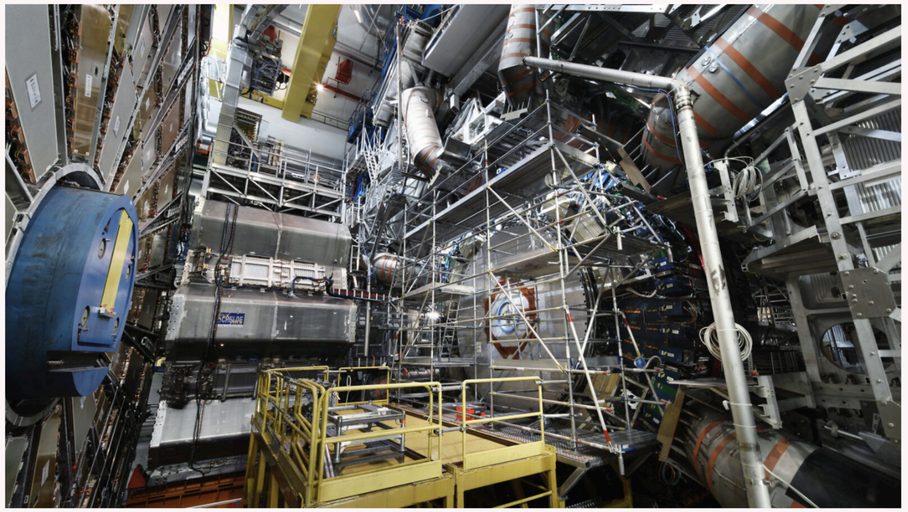
DOI 10.17182/hepdata.98796.v2



Likelihoods - the best Information we have

symmetry topics follow +

A joint Fermilab/SLAC publication



Courtesy of CERN

ATLAS releases ‘full orchestra’ of analysis instruments

01/14/21 | By Stephanie Melchor

The ATLAS collaboration has begun to publish likelihood functions, information that will allow researchers to better understand and use their experiment’s data in future analyses.

Meyrin, Switzerland, sits serenely near the Swiss-French border, surrounded by green fields and the beautiful Rhône river. But a hundred meters beneath the surface, protons traveling at nearly the speed of light collide and create spectacular displays of subatomic fireworks inside the experimental detectors of the Large Hadron Collider at CERN.

3

Updates > News > New open release streamlines interactions with theoretical physicists

News

Tags:
open data

New open release streamlines interactions with theoretical physicists

The ATLAS Collaboration has released the first open likelihoods from an LHC experiment.

12th December 2019 | By Katarina Anthony

Explore ATLAS open likelihoods on the HEPData platform. (Original image: Ahmet Anil Sen/Behance)

What if you could test a new theory against LHC data? Better yet, what if the expert knowledge needed to do this was captured in a convenient format? This tall order is now on its way from the ATLAS Collaboration, with the first open release of full analysis likelihoods from an LHC experiment.

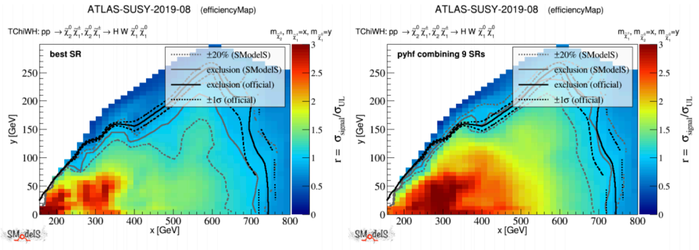


Figure 3: Validation of the TChiWH ($pp \to \tilde{\chi}_2^0\tilde{\chi}_1^\pm$, $\tilde{\chi}_2^0 \to h\tilde{\chi}_1^0$, $\tilde{\chi}_1^\pm \to W\tilde{\chi}_1^0$) result from the ATLAS electroweakino search [23], on the left using the best SR, on the right using the full likelihood.

in accuracy one can reach with using pyhf and full likelihoods. The ATLAS collaboration is at the beginning of a huge effort to provide full statistical models for new analyses. The first analyses published already show how this can help theorists make more trustful reinterpretations. The importance of such likelihood information for, e.g., global fits, has also been emphasised in [2].

4. Conclusions

We presented an interface of SModelS to pyhf that enables the use of the full likelihoods provided by ATLAS in the form of pyhf JSON files. The SModelS database was extended by efficiency map results with the corresponding JSON files of three new ATLAS SUSY analyses [16, 17, 23] for full Run 2 luminosity (139 fb^{-1}).

The new version, SModelS v1.2.4, is publicly available from `https://smodels.github.io/` and can readily be employed for physics studies. We congratulate ATLAS to the important move of making full likelihood information available in digital format and are looking forward to including more such data in future updates of SModelS.

This completes the work started in contribution 15 of [9] for SModelS; the MadAnalysis 5 interface to pyhf should become available in the upcoming MadAnalysis 5 v1.9 release.

Last but not least we note that the technical discussions with the pyhf team are handled via github’s issue tracking system, see e.g. `https://github.`

ATLAS first LHC experiment to push this Open Data frontier

Analysis Preservation

Beyond the collision data, ATLAS produced many other types of data
Importantly, the analyses prepared to study the collision data is itself data worth preserving

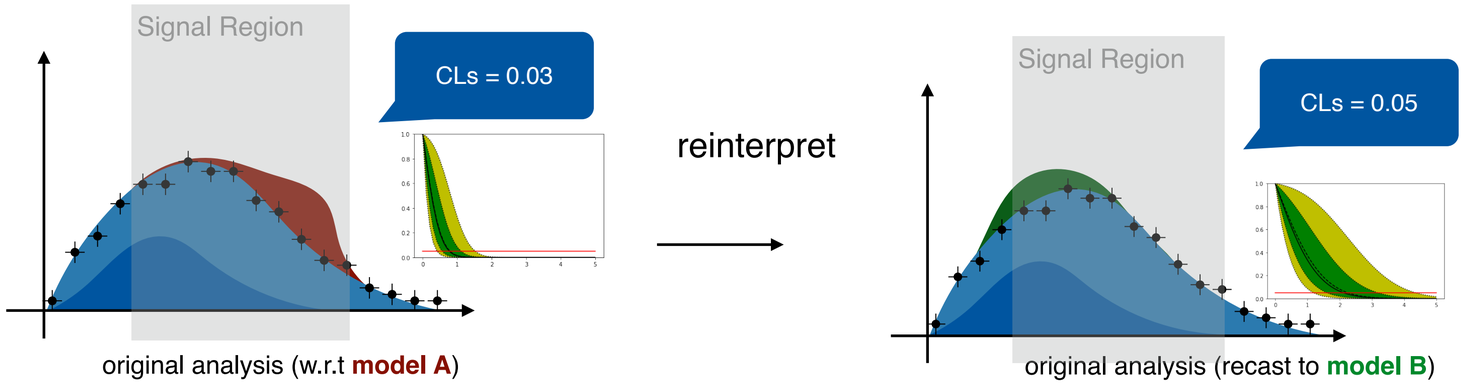
- **historical record of scientific process**
- **useful tool to expand science reach of ATLAS**

ATLAS is investing in this area internally and works closely w/ CERN to develop and use infrastructure (REANA, CAP)

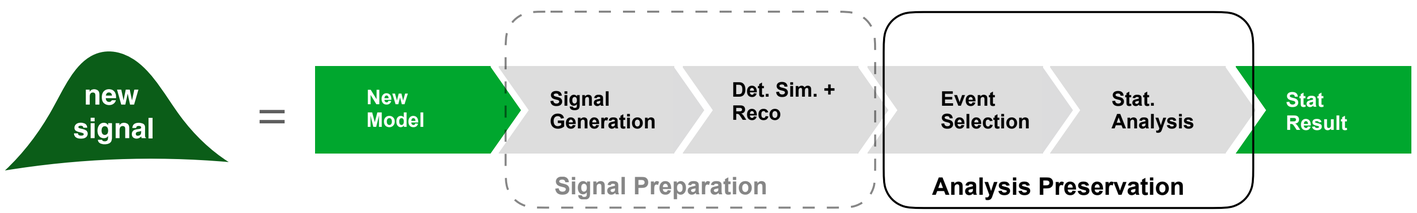
Analysis Implementations

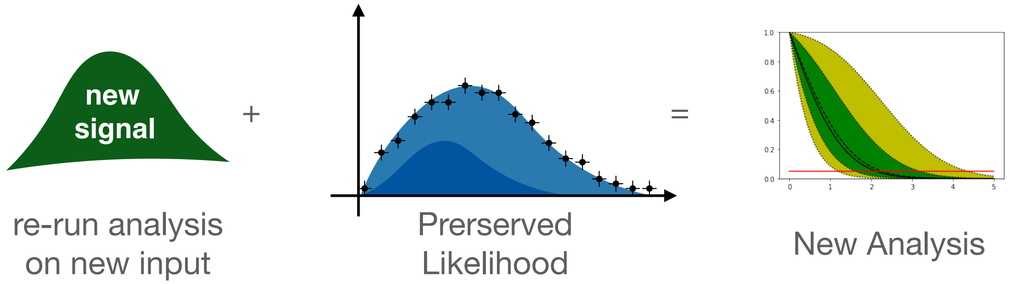
Important Target for Reuse Program: not necessarily the full workflow but special workflows for reinterpretation of existing analyses (RECAST)

- **answer key question: what can existing analysis say about new theories**



- **only need to have workflows to re-run analysis on new signal sample**





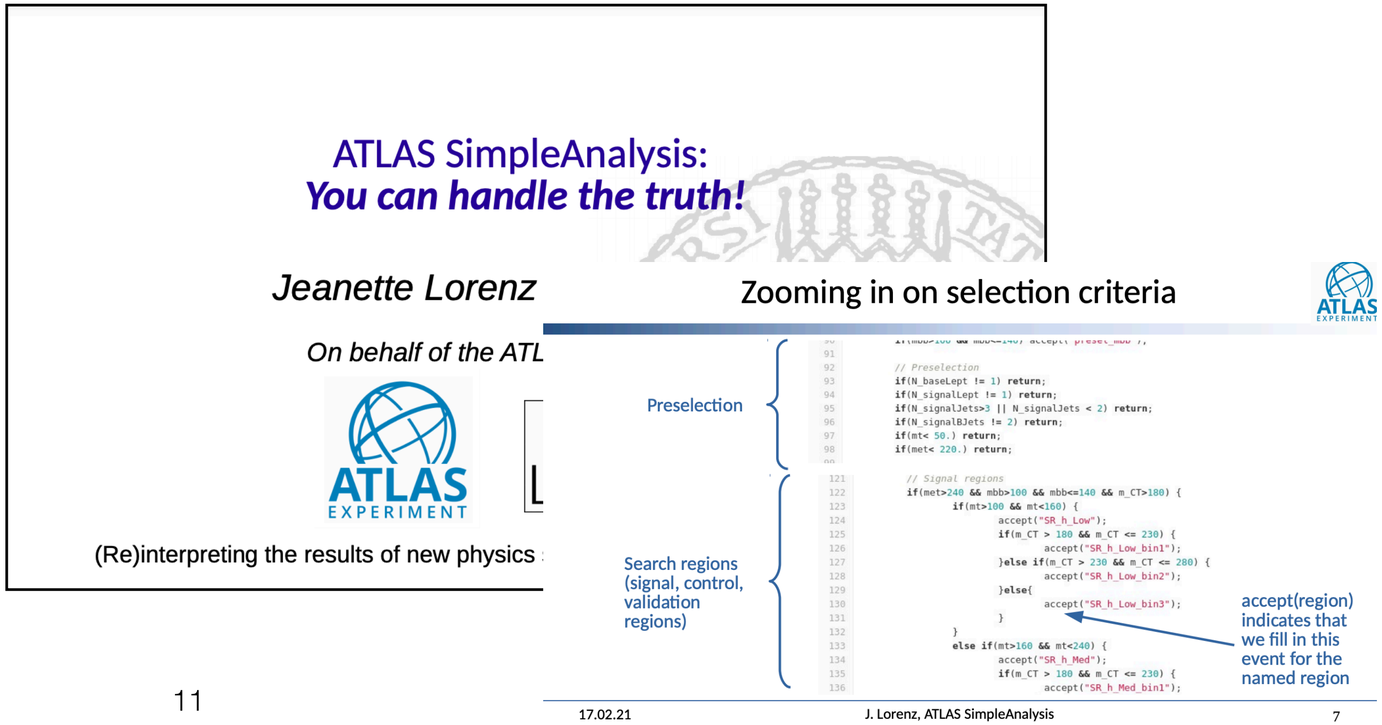
Simplified Analysis implementations

HEP has long history of releasing simplified Analysis Procedures

- **Rivet, HZTool, ... mostly from a Standard Model perspective**

ATLAS has interally been developing a library of simplified analyses (`SimpleAnalysis`) for Beyond-Standard-Model analyses (SUSY, ...)

- **This year ATLAS decided to release this library in-bulk!**



Analysis Implementations

Simplified Analyses have their limits

- **more complex analyses as with decade of experience with LHC**
- **Machine Learning, low-level detector-specific inputs, ...**

Crucial to preserve the acctual analysis and make it accessible

- **caveat: real analysis much more resource intensive than simplified ones**

Three Key components:

capture software

archive analysis code incl. dependencies

capture commands

what do with the captured software

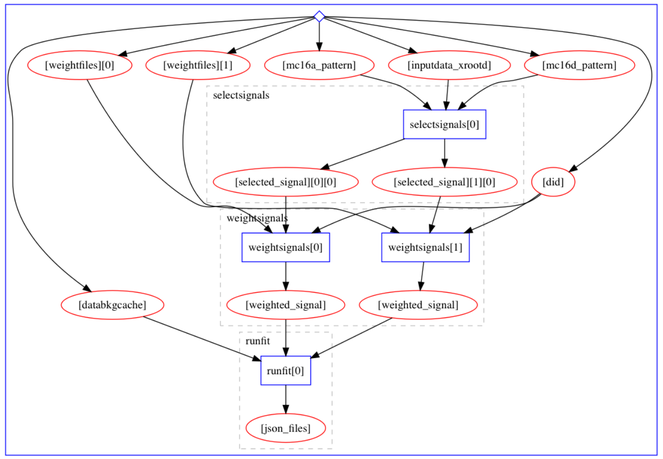
capture workflow

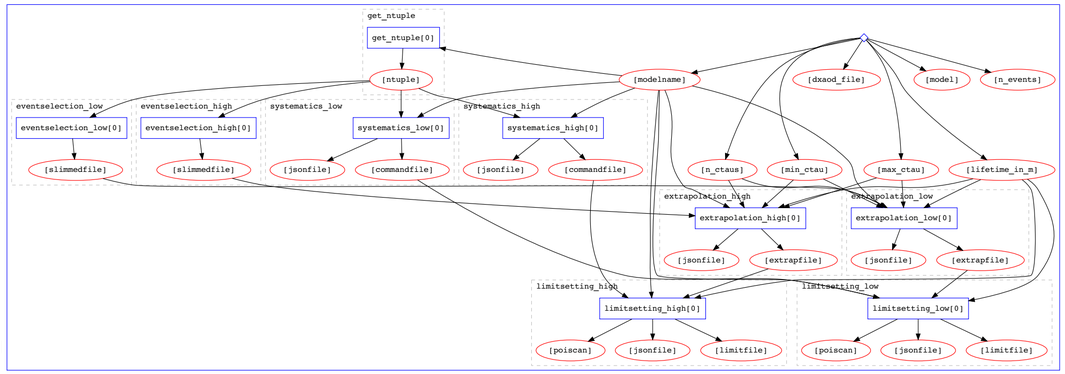
order of individual steps

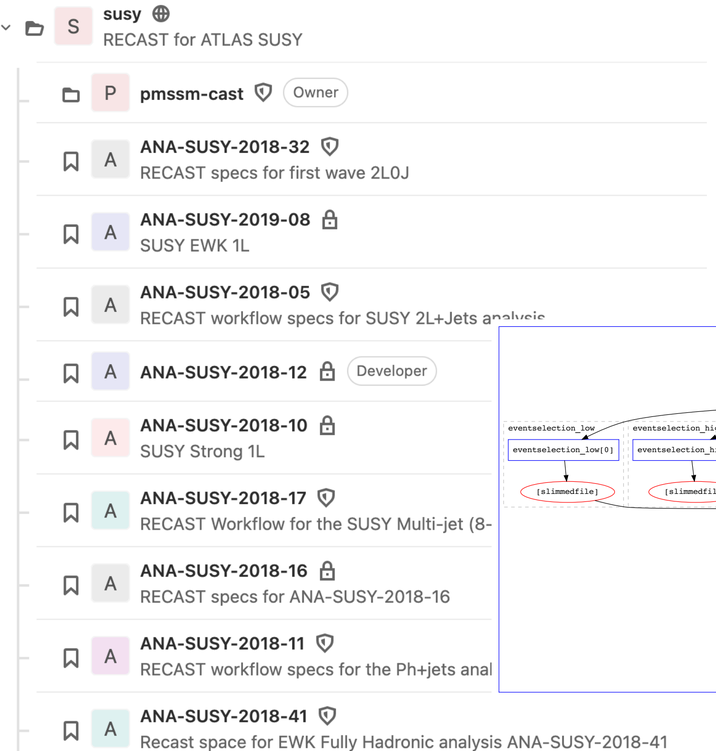
Analysis Implementations

In ATLAS: all major Beyond Standard Model groups require publications to prepare a RECAST Workflow to be run on workflow service (REANA)

- **Basis for new publications**







ATLAS EXPERIMENT

ATLAS PUB Note

ATL-PHYS-PUB-2019-032

11th August 2019

CERN

RECAST framework reinterpretation of an ATLAS Dark Matter Search constraining a model of a dark Higgs boson decaying to two *b*-quarks

The ATLAS Collaboration

ATLAS EXPERIMENT

ATLAS PUB Note

ATL-PHYS-PUB-2020-007

27th March 2020

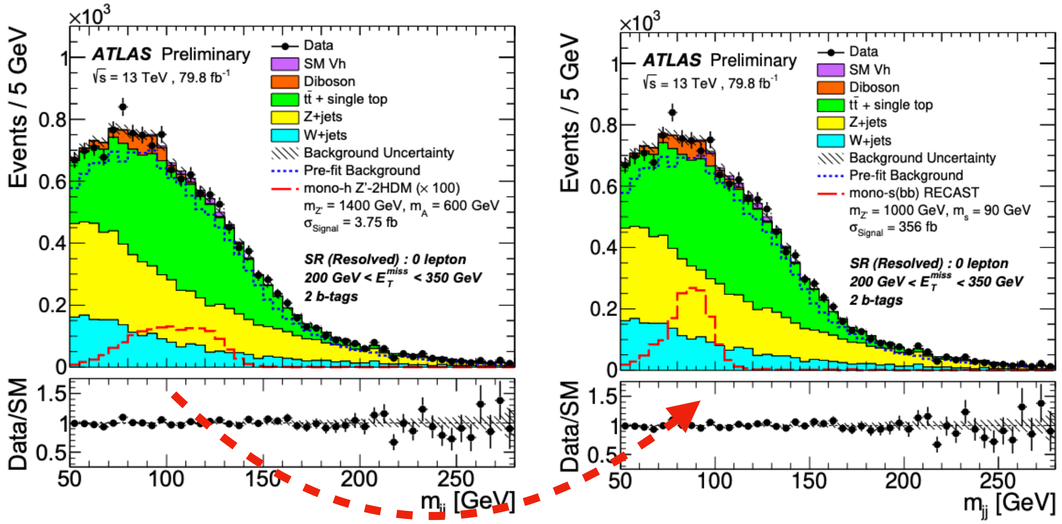
CERN

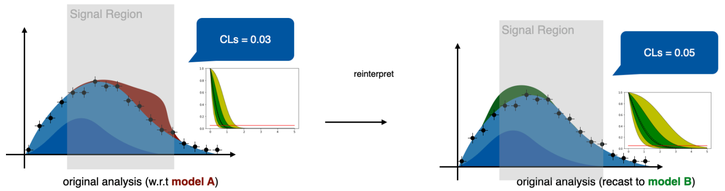
Reinterpretation of the ATLAS Search for Displaced Hadronic Jets with the RECAST Framework

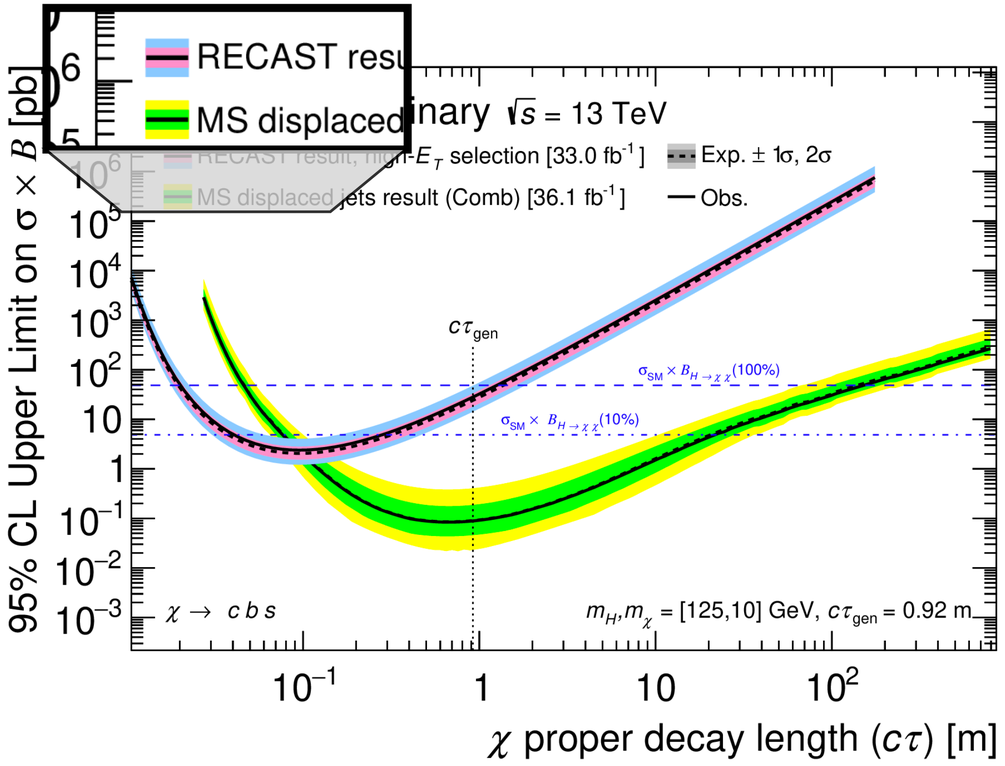
The ATLAS Collaboration

Analysis Implementations

Systematic Approach for preserved analyses allow us to extend physics reach using our existing analysis!







Analysis Implementations

Continuing to work with REANA and CERN Analysis Preservation develop infrastructure for reuse

reana

Reproducible research data analysis platform

CERN
Analysis Preservation
capture, preserve and reuse physics analyses

reana

Help

recast-cd7cd75c #1
Finished 24 days ago

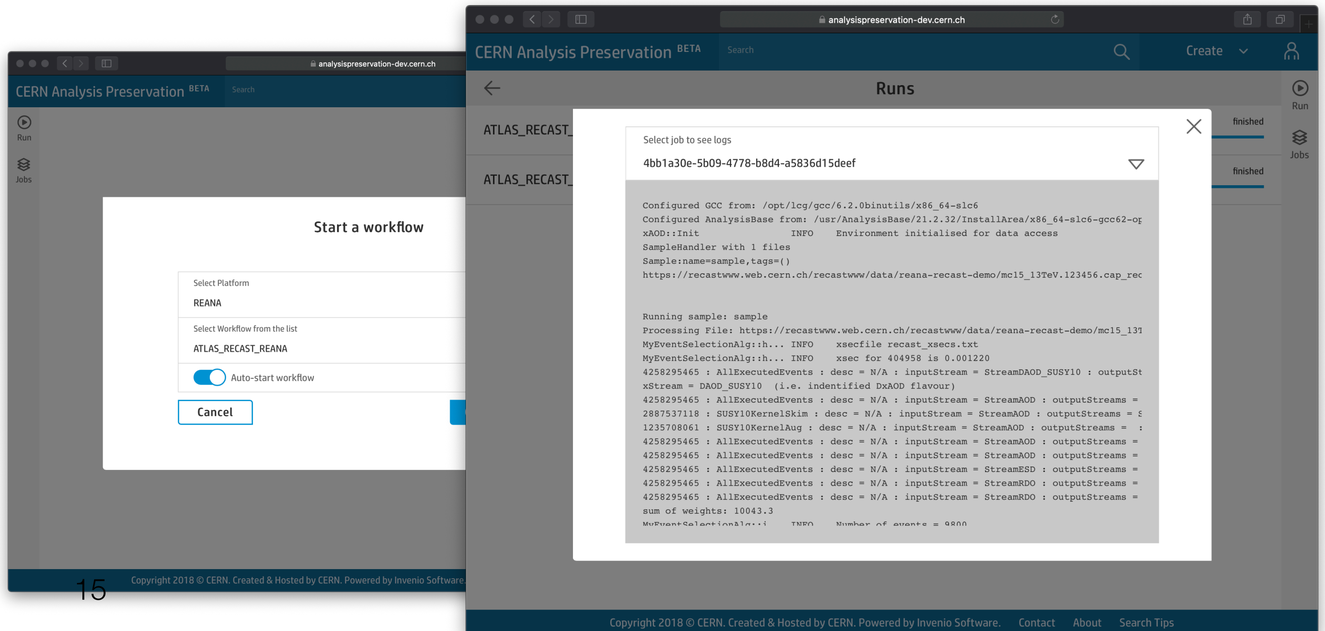
finished in 1h 8m 44s
step 10/6

Logs | Workspace | Specification

Step: selection_stage_mc16a_0

finished | Kubernetes | gitlab-registry.cern.ch/atlas-mpp-xampp... | $ source /r

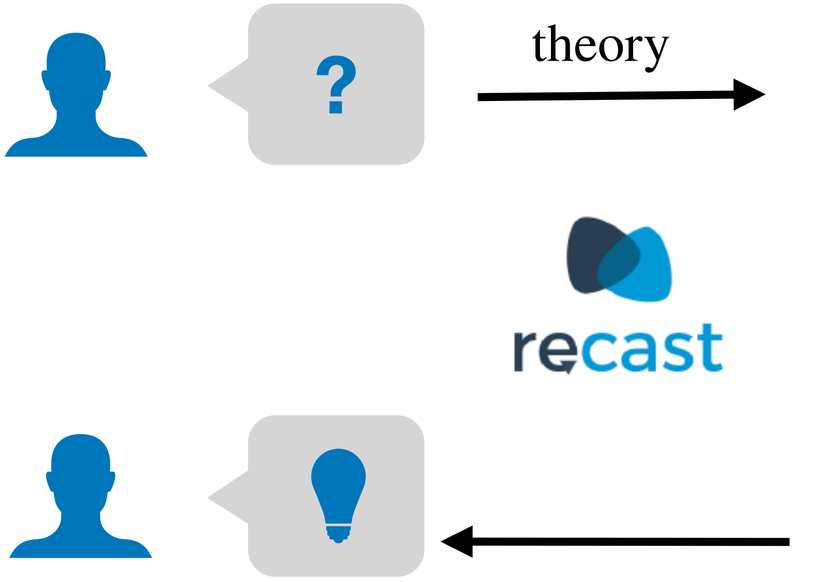
```
job: :
 bash: line 1: /recast_auth/getkrb.sh: No such file or directory
Configured GCC from: /opt/lcg/gcc/8.3.0-cebb0/x86_64-centos7/bin/gcc
Taking LCG releases from: /opt/lcg
Taking Gaudi from: /usr/GAUDI/21.2.108/InstallArea/x86_64-centos7-gcc8-opt
Configured AthAnalysis from: /usr/AthAnalysis/21.2.108/InstallArea/x86_64-centos7-g
--------------------
Directory contents:
CMakeLists.txt
Dockerfile
README.md
```



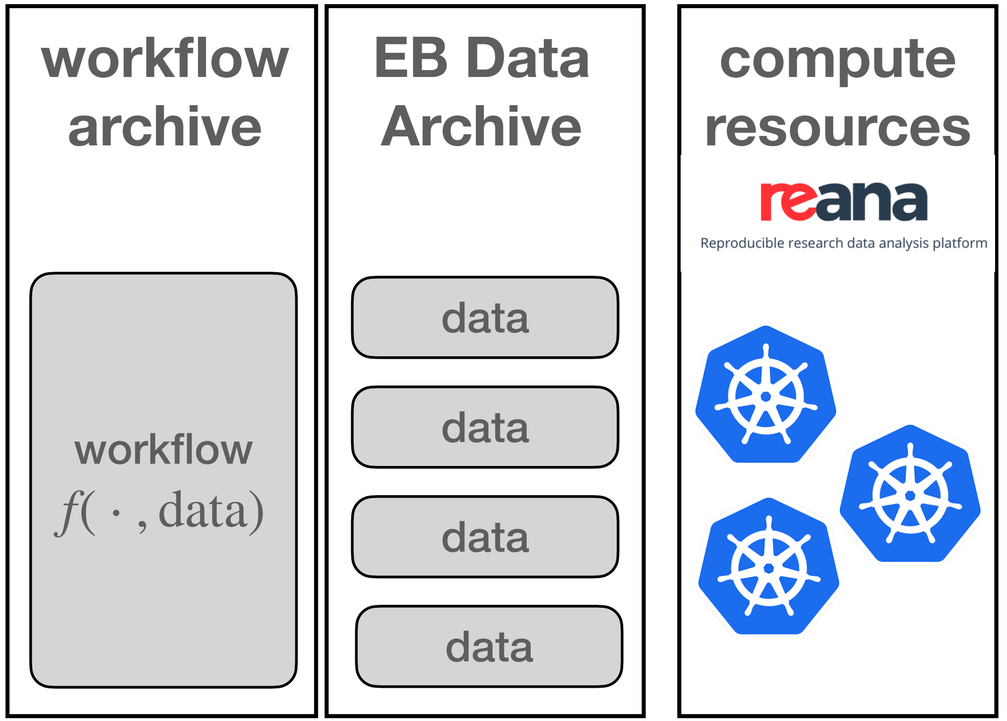
RECAST

Preserved Analysis can provide "semi-open" access at a higher, abstract level

- **Science Gateway that lets theorists "query" the data ("here's my theory, is it excluded by your data?)**



External Scientists



CERN

Outlook

ATLAS has a broad Open Science vision focused on providing best tools for external community

Recently there have been a number of innovations in various areas

- **New L3 Open Data policy + first releases**
- **Open Likelihoods**
- **Large-scale release of simplified Analysis Procedues**
- **RECAST & REANA - full analysis preservation**

Challenges
- **we're new to L3 data, looking forward to learning from you**
- **HL-LHC makes analysis preservation even harder**