

# Data & Analysis Preservation in ATLAS

Lukas Heinrich

# Data Preservation Policies:

Focus on preservation of RAW data. Derived data (e.g. small analysis object formats) subject to lifetime policies.

Rely on software preservation for ability to reproduce downstream formats. Detailed database on data provenance.

Trade-off: Compute vs Storage

Calibration data, etc available

Future software will always be able to reprocess full ATLAS dataset (all runs)

e.g. reading old data w/ new software

Approved CB 20<sup>th</sup> February 2015

## ATLAS Data Preservation Policy

February, 2015

**Purpose of this Policy Document**  
The principal intent of this document is to describe the ATLAS policy ensuring that its data are maintained reliably in a form accessible to ATLAS members. A separate document describes the ATLAS policy for making its data available, and potentially useful, to scientists who are not members of ATLAS.

**ATLAS Data**  
In this context, ATLAS data comprises the acquired raw data, simulated data, the derived data products stored and catalogued in the ATLAS Distributed Data Management system, the calibration data, metadata, transformations (code, including that for simulation) and the documentation required to create the derived products and use them to obtain physics results (\*).

ATLAS is committed to preserving all raw data from collisions in such a way that they can be reprocessed and reanalysed for the active lifetime of the collaboration.

**Preservation of the Data in Common Formats**

**Non-Reproducible Data**  
The preservation policy for raw data aims at reducing the risk of loss due to either technical failure or disasters, such as fire or earthquake, to a very low level. At least two copies of ATLAS raw data are stored on accessible archival media at a subset of WLCG (Worldwide LHC Computing Grid) sites. No site holds all copies of particular part of the data.

Other non-reproducible data including calibration data, metadata, documentation, and transformations are stored within the WLCG in professionally engineered and backed up databases or file systems. Old versions of these data are archived.

**Derived Data**  
Derived data and simulated data are, in principle, fully reproducible at any time provided the appropriate calibration data, metadata, transformations, CPU architecture or emulations thereof, and documentation are available.

(\*) ATLAS physics data do not contain any personally identifiable information.

**Commitment of the Institutions Hosting ATLAS Data**  
The WLCG MOU [1] describes the commitment of the institutions hosting ATLAS data.

**Preservation of Physics Results & the Ability to Re-derive Them**  
To produce physics results from ATLAS data, human resources are required in addition to the preservation of data, metadata and documentation. Some of the processes, such as the internal peer review of intermediate and final results, cannot be captured as fully documented, reproducible procedures. The ATLAS Collaboration intends to maintain the knowledge necessary to perform and review physics analysis for as long as possible after data taking ceases.

ATLAS internal documentation relating to physics results, for example the detailed analysis notes that are input to internal review processes, are maintained in a professionally operated document management system.

Scientific outputs published in journals, or submitted to repositories such as arXiv and HEPDATA are assumed to be preserved by the journals or repository operators. In addition they are also archived by the ATLAS Collaboration.

**Outreach and Educational Formats**  
ATLAS also produces outreach and educational datasets and formats, both for use by ATLAS members and for third parties. While it commits to supporting these activities, it makes no long-term commitment to preserving these datasets and specific formats.

**Data Preservation Beyond the Lifetime of the Collaboration**  
At the point at which the collaboration becomes inactive, the intention is that the raw collision data and a selection of derived formats will be preserved and made available, along with the appropriate version of the processing software, metadata and associated simulation software.

[1] <http://wlcg.web.cern.ch/collaboration/mou>

## Data **Access** Policies:

So far focus on:

1. providing high-quality distilled data products for research use
2. event-level Open Data for Outreach & Education purposes
3. Association Programs for collaboration

Level 1: data products based on publications.

(See later slides).

Primary target for open access data repository: HepData

- help theorists construct approximate implementations of analyses (e.g. Rivet, ...)

Link to internally archived analyses for e.g. reinterpretation (e.g. RECAST, Likelihoods)



## Data **Access** Policies:

Level 2: special purpose datasets.

Open Data ([opendata.cern.ch](https://opendata.cern.ch)) currently focused on Outreach & Education use-cases.

For researchers interested in collaborating on research projects, ATLAS has three association mechanisms:

- Short Term Association (STA)
- Analysis Consultants & Experts (ACE)
- Monte Carlo Generator Interactions (MCI)

Level 1

Level 2

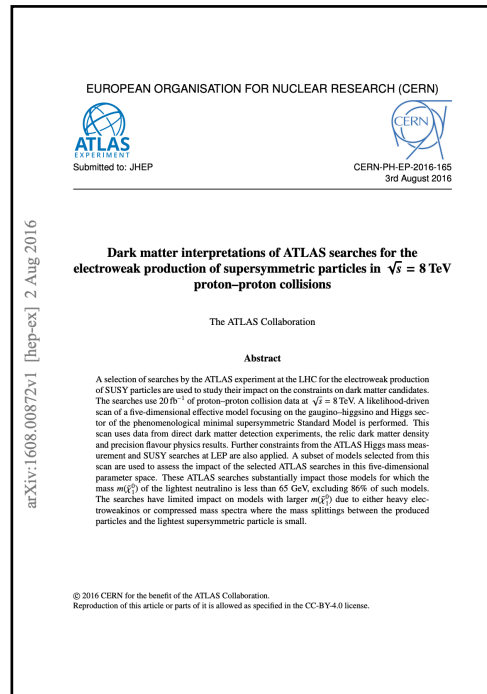
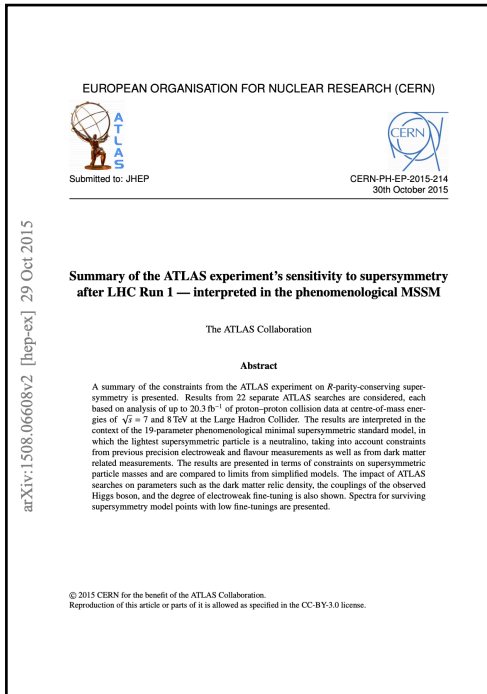
Level 3

Level 4

## Short Term Association (STA)

- 36 STA since 2014
- e.g. advising theorists
- become authors on resulting publications

## Example: Run-1 Summary Publications



## Data Access Policies:

### Analysis Consultants & Experts (ACE):

- 48 ACE since 2016
- access to full ATLAS Monte Carlo for e.g. R&D in fast simulation
- public document signed by ATLAS collaboration
- resulting datasets may become public after publication
  - explicit possibility to publish ML datasets
- credit through acknowledgement reference to method paper (exceptionally: authorship possible)

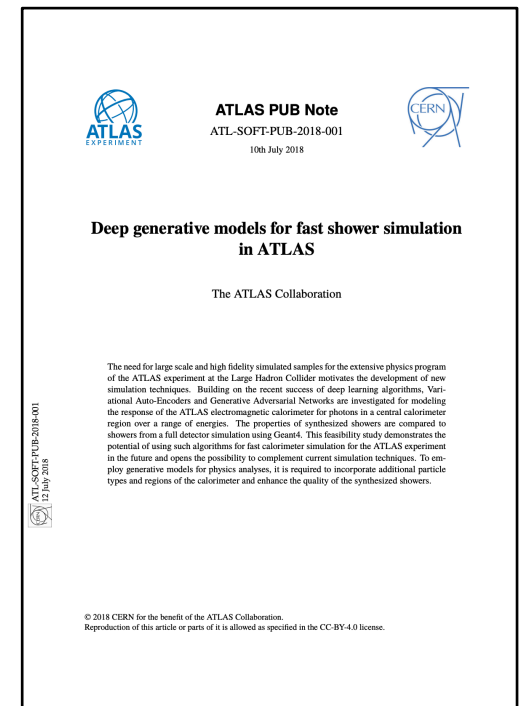
DPHEP Data Nomenclature  
Level 1 - Level 4 ([document](#))

Level 1

Level 2

Level 3

Level 4



## Data **Access** Policies:

### Level 3: Reconstructed Open Data

#### Current Policy:

embargoes reconstructed data for physics exploitation.

- Release in the future possible.
- No technical obstacle, but policy decisions.

Main concern: evidence of sufficient tooling and resources to adequately analyze reconstructed data at scale.

Level 4: Raw Data. Not considered useful for release. But preserved and possible to release beyond the collaboration lifetime

Level 1

Level 2

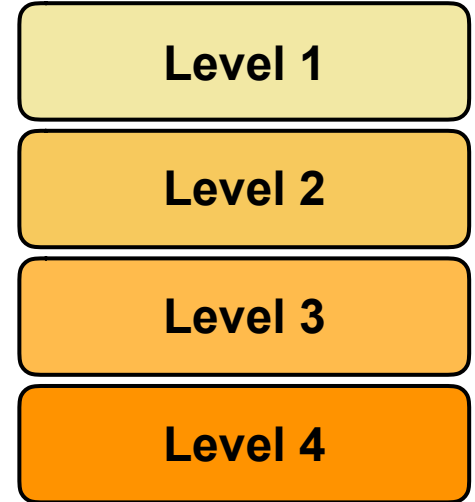
Level 3

Level 4

# Data Access Review

Since original drafting of Data Access Policy significant developments:

- prevalence of Open Data
- Funding Agency views on Open Data / FAIR
- technical capabilities available to non-members



ATLAS is reviewing its Open Data / Data Access policy within the collaboration.

## Technical Advancements:

- Open Sourcing of full reconstruction & analysis framework
- R&D towards feasibility of Level-3 Analysis of HL-LHC scale data using e.g. cloud resources
- development of fully calibrated common data format PHYSLITE likeliest candidate for L3 release

Open Source





# Analysis Preservation Efforts

Broadly there are two themes in Analysis Preservation.

"The Museum"



long-term, descriptive,  
archival, historical record  
of scientific activity

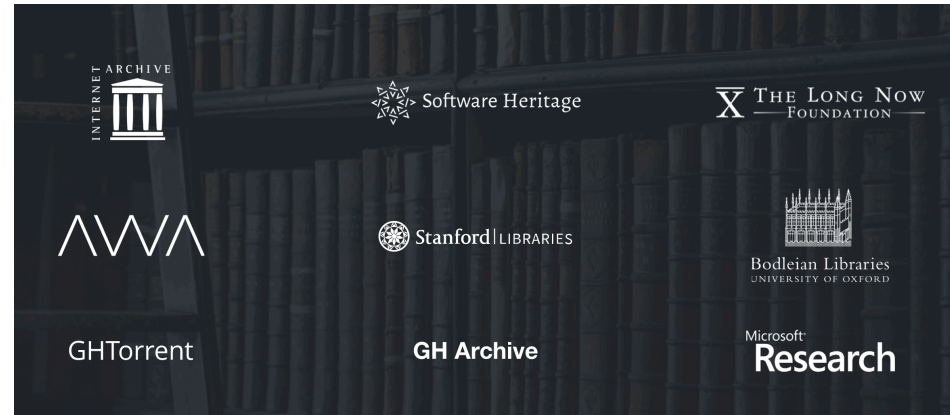
"The Hangar"



short-term, actionable,  
re-usable, deployable  
analysis implementation

# Software as cultural artifact

- partnering w/ e.g. Software Heritage project
- opportunity for LHC expts to contribute



```
chrisgarry / Apollo-11
Watch 1.3k
Code Issues 15 Pull requests 13 Actions Projects 0 Wiki Security Ins Archive Program
Branch: master Apollo-11 / Luminary099 / BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc
wopian Trim whitespace
3 contributors
1059 lines (866 sloc) | 21.8 KB
Raw Bl

1 # Copyright: Public domain.
2 # Filename: BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc
3 # Purpose: Part of the source code for Luminary 1A build 099.
4 # It is part of the source code for the Lunar Module's (LM)
5 # Apollo Guidance Computer (AGC), for Apollo 11.
6 # Assembler: yaYUL
7 # Contact: Ron Burkey <info@sandroid.org>.
8 # Website: www.ibiblio.org/apollo.
9 # Pages: 731-751
10 # Mod history: 2009-05-19 RSB Adapted from the corresponding
11 # Luminary131 file, using page
12 # images from Luminary 1A.
13 # 2009-06-07 RSB Corrected 3 typos.
14 # 2009-07-23 RSB Added Onno's notes on the naming
15 # of this function, which he got from
16 # Don Eyles.
17 #
18 # This source code has been transcribed or otherwise adapted from
19 # digitized images of a hardcopy from the MIT Museum. The digitization
20 # was performed by Paul Fjeld, and arranged for by Deborah Douglas of
21 # the Museum. Many thanks to both. The images (with suitable reduction
22 # in storage size and consequent reduction in image quality as well) are
23 # available online at www.ibiblio.org/apollo. If for some reason you
24 # find that the images are illegible, contact me at info@sandroid.org
25 # about getting access to the (much) higher-quality images which Paul
26 # actually created.
27 #
28 # Notations on the hardcopy document read, in part:
29 #
30 #
```

## GitHub Archive Program

# Preserving open source software for future generations

**It is a hidden cornerstone of modern civilization, and the shared heritage of all humanity. The mission of the GitHub Archive Program is to preserve open source software for future generations.**

Get your code into the [GitHub Arctic Code Vault](#)

# 02/02/2020

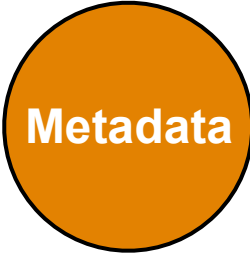
GitHub is partnering with the Long Now Foundation, the Internet Archive, the Software Heritage Foundation, Arctic World Archive, Microsoft Research, the Bodleian Library, and Stanford Libraries to ensure the long-term preservation of the world's open source software. We will protect this priceless knowledge by storing multiple copies, on an ongoing basis, across various data formats and locations, including a very-long-term archive designed to last at least 1,000 years.

**Both are important and being worked on within ATLAS & CERN**



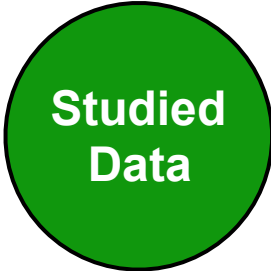
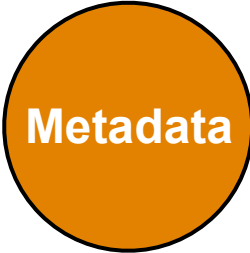
# Both are important and being worked on within ATLAS & CERN

Analysis Team  
Internal Notes  
Bibliographic Info  
...



# Both are important and being worked on within ATLAS & CERN

Analysis Team  
Internal Notes  
Bibliographic Info  
...

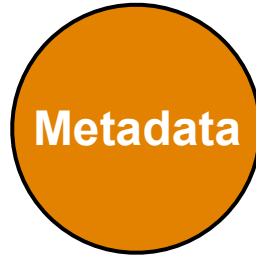


Ntuples / Trees  
for Data & MC  
...



# Both are important and being worked on within ATLAS & CERN

Analysis Team  
Internal Notes  
Bibliographic Info  
...



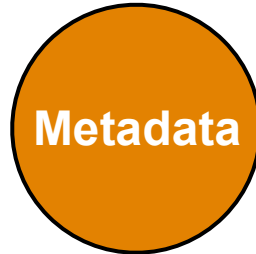
Ntuples / Trees  
for Data & MC  
...



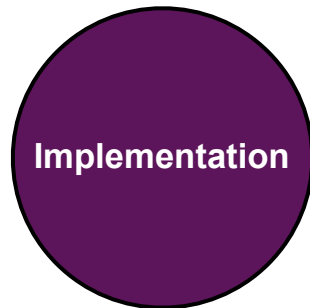
UFO models  
Likelihoods  
Limits  
Measurements  
Eff. Tables  
HepData  
...

# Both are important and being worked on within ATLAS

Analysis Team  
Internal Notes  
Bibliographic Info  
...



Ntuples / Trees  
for Data & MC  
...

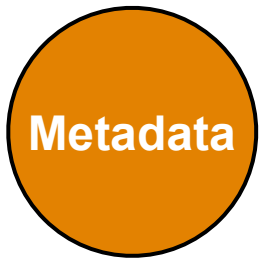


Code  
Runtime Environment  
Workflows



UFO models  
Likelihoods  
Limits  
Measurements  
Eff. Tables  
HepData  
...

# ATLAS has detailed tracking of analysis from inception to publication using in-house system **GLANCE**

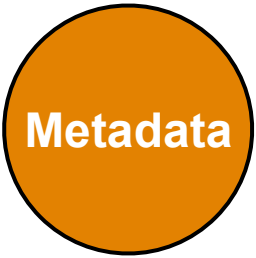


- Links to Code repositories
- Metadata on
  - used data formats
  - triggers
  - physics processes
  - ...
- Analysis Team / Editorial Board
- Links to
  - Internal docs
  - Published docs

The screenshot shows the ATLAS Analysis system interface. The top panel displays the 'Sbottom multi-b ANA-SUSY-2018-31' analysis page. It includes a search bar, a user profile, and a list of team members with links to their profiles. The page is in 'Phase 0' and is 'Finished'. The bottom panel shows the 'Sbottom multi-b SUSY-2018-31' analysis page. It includes a search bar, a user profile, and a list of team members. The page is in 'Phase 1' and is 'Active'. The page content includes a summary, collision details, and a phase 1 data section with start date and analysis code links.



# Automatic Import into CERN Analysis Preservation



Full database access to internal analysis tracking for CAP. Working on automatic ingestion.



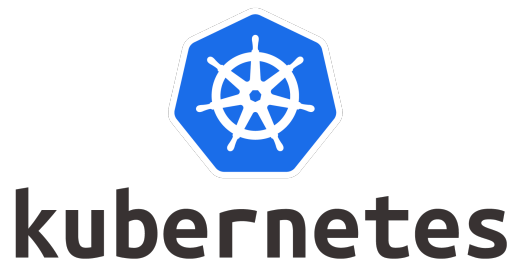
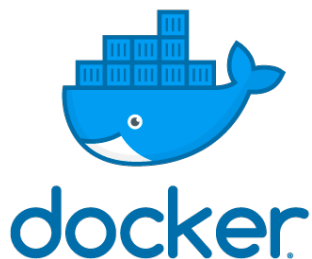
The image displays two overlapping browser screenshots illustrating the automatic import process. The left screenshot shows the ATLAS Analysis website (glance.cern.ch) with a document page for "JDM - Dark Matter Summary" (EXOT-2017-32). The right screenshot shows the CERN Analysis Preservation interface (analysispreservation-dev.cern.ch) with a submission form for the same document. The submission form includes fields for "Dark Matter and Dark Energy Summary 13 TeV", "Glance ID" (123), "Abstract", and "People Involved". Below the form, there is a section for "Information from GLANCE database" which automatically takes data from the GLANCE ID, including "GLANCE ID" (123), "Short Title" (JDM - Dark Matter and Energy Summary), "Full Title" (Searches for Dark Matter), "Publication title" (Dark Matter and Dark Energy Summary 13 TeV), and "Ref Code" (ANA-EXOT-2017-12345). An arrow points from the left screenshot to the right one, indicating the flow of data from the ATLAS Analysis website to the CERN Analysis Preservation interface.

# ATLAS is investing in re-useable / re-producible analysis

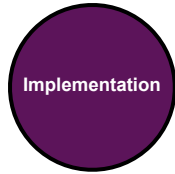
## Technology Choice for software archival:

- Git
- Linux Containers

Currently best-of-breed tools, widely adopted beyond HEP.



# Containers in ATLAS: reproducible software environments



- integrated in Analysis Software Release Schedule
- teach continuous testing / validation / preservation in ATLAS induction / software tutorials
- integrated into distributed computing infrastructure (containers on Grid)

atlas  
Community Organization

Repositories

Showing 7 of 7 repositories

- atlas/athena  
By atlas • Updated 9 hours ago  
ATLAS Athena Release  
Container
- atlas/athanalysis  
By atlas • Updated 5 days ago  
ATLAS Athena Analysis Release  
Container

## Merge branch 'lheweights' into 'master'

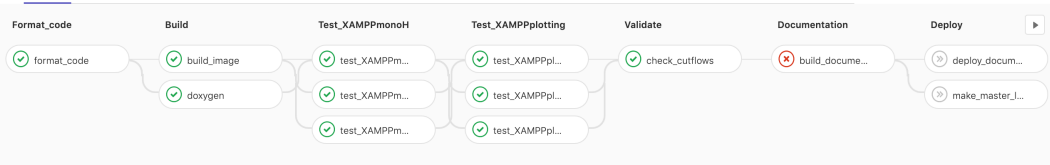
add features to evaluate LHE systematics in background samples  
See merge request 1211

13 jobs for master in 39 minutes and 18 seconds (queued for 1 second)

latest

7c3971d2

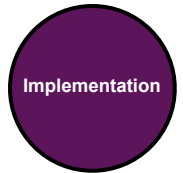
Pipeline Jobs 13 Failed Jobs 1



## Dockerfile 400 Bytes

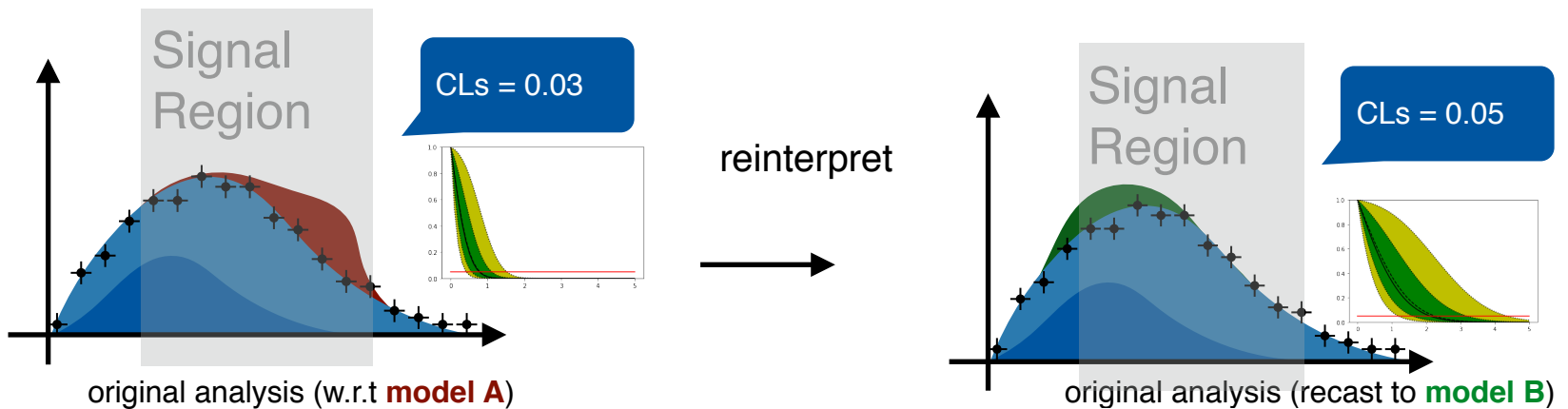
```
1 # The release set in this Dockerfile defines the release
2 # and is parsed by every setup script and by the installation
3 # Be aware of this effect if you edit the release here!
4 FROM atlas/athanalysis:21.2.85
5 ADD . /xampp/XAMPPmonoH
6 WORKDIR /xampp/build
7 RUN source ~/release_setup.sh && \
8     sudo chown -R atlas /xampp && \
9     cmake ../XAMPPmonoH && \
10    make -j4
```

# Major physics groups have adopted Analysis Preservation as part of their approval procedure.



## Currently focused on BSM program (SUSY & Exotics)

## Main use-case: RECAST (reinterpreting searches)



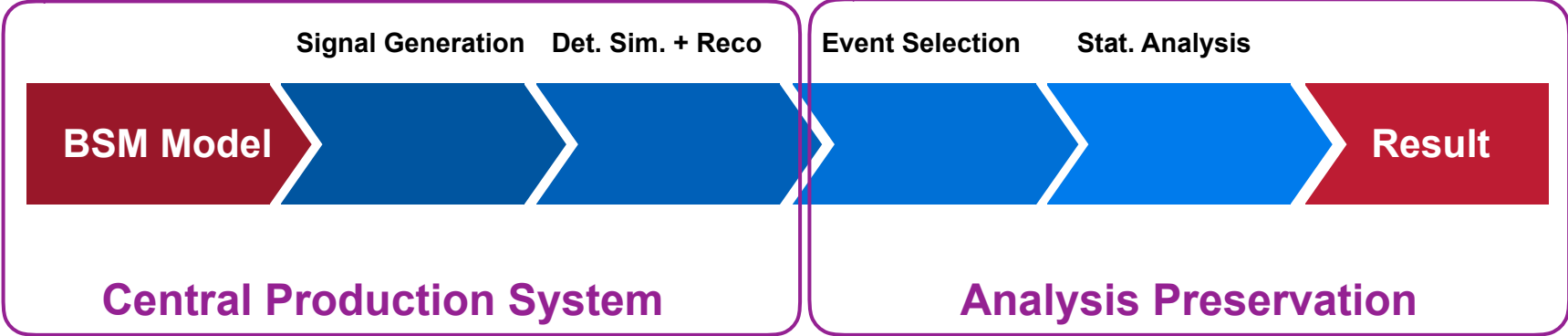
For operational analysis preservation need to preserve full pipeline. Demarcation line: central production.

Analysis is the part of the pipeline that is not handled centrally by the experiment.

Software Preservation of central code is a **separate/easier problem.**

Corollary: if more of analysis is done centrally, the easier they are to preserve.

e.g.data reduction as a service (Derivation System)



# Preservation of Code, Scripts, Workflows:

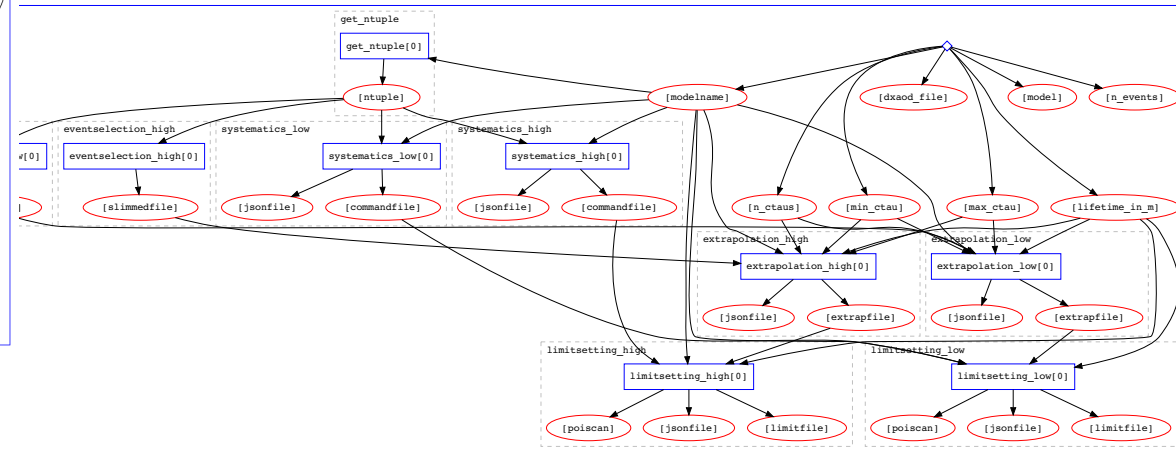
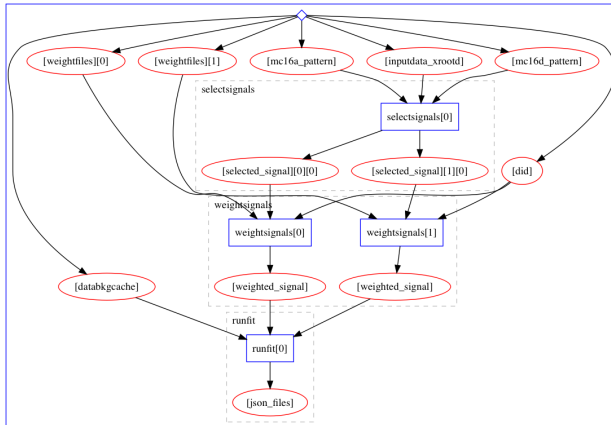
1. capture software  
container images

2. capture commands  
job templates

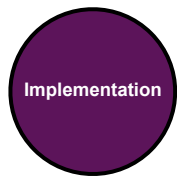
3. capture workflow  
how do I connect the pieces

working with CERN Analysis Preservation  CERN  
Analysis Preservation  
to make ATLAS Analyses

## reana



# Re-execution on independent infrastructure:



Working with  
CERN projects:



# Re-execute analysis from preserved record:


The left screenshot shows the 'Start a workflow' dialog in the CERN Analysis Preservation interface. It includes a 'Select Platform' dropdown set to 'REANA', a 'Select Workflow from the list' dropdown set to 'ATLAS\_RECAST\_REANA', and an 'Auto-start workflow' toggle. 'Cancel' and 'Create Workflow' buttons are at the bottom.

The right screenshot shows the 'Runs' page with a modal window displaying the execution logs for a specific job. The job ID is '4bb1a30e-5b09-4778-b8d4-a5836d15deef'. The logs show the configured GCC, AnalysisBase, and SampleHandler, followed by the execution of the sample 'sample' with various event selection and stream processing steps.


# New Results from preserved analyses

- additional reinterpretations in-flight
- planning to use for Run-2 summary papers

new result

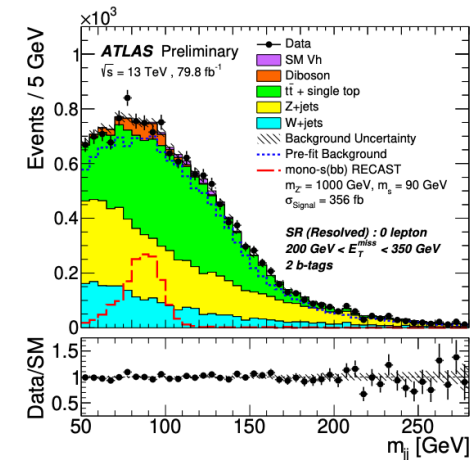
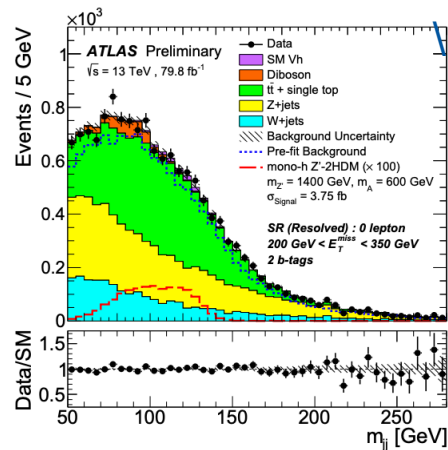



**ATLAS PUB Note**  
ATL-PHYS-PUB-2019-032  
11th August 2019



**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 CONF Note**  
ATLAS-CONF-2018-039  
25th July 2018

Search for Dark Matter Produced in  
with a Higgs Boson decaying to  $b\bar{b}$  at  
with the ATLAS Detector using  $79.8 \text{ fb}^{-1}$  of  
proton-proton collision data

The ATLAS Collaboration

original publication



# ATLAS provides extensive information publicly for their analyses on HepData



HEPData

Traditionally:

Tabulated Data on measured observables.

More Recently:

- pseudo-code for event selection
- efficiency maps
- multivariate discriminants (BDTs, etc)

```
Branch: master use-atlas-bdt-hepdata / Usage_ATLAS_HepData_BDT.ipynb
lukasheinrich Add files via upload
1 contributor

152 Lines (152 sloc) 20.6 KB

In [14]: import ROOT
import array
import matplotlib.pyplot as plt
import numpy as np
import matplotlib inline

In [2]: !wget -O bdt.xml https://www.hepdata.net/record/resource/406719?view=true
--2019-04-01 07:35:38-- https://www.hepdata.net/record/resource/406719?view=true
Resolving www.hepdata.net (www.hepdata.net)... 188.184.64.140
Connecting to www.hepdata.net (www.hepdata.net)|188.184.64.140|:443... conn
HTTP request sent, awaiting response... 200 OK
Length: 953915 (932K) [text/xml]
Saving to: 'bdt.xml'

bdt.xml 100%[=====>] 931.56K 605KB/s in 1.5s
2019-04-01 07:35:40 (605 KB/s) - 'bdt.xml' saved [953915/953915]

In [15]: ROOT.TMVA.Tools.Instance()
reader = ROOT.TMVA.Reader()

var = [array.array('f',[0]) for i in range(7)]
reader.AddVariable("MET",var[0]);
reader.AddVariable("MT",var[1]);
reader.AddVariable("dMT200",var[2]);
reader.AddVariable("m_tophad",var[3]);
reader.AddVariable("m_toplep200",var[4]);
reader.AddVariable("dphi_lep_nu200",var[5]);
reader.AddVariable("dphi_rjet_lep",var[6]);

reader.BookMVA("BDT method", "bdt.xml")

Out[15]: <ROOT.TMVA::MethodBDT object ("BDT") at 0x55c1e6050e0>
: Booking "BDT method" of type "BDT" from bdt.xml
: Reading weight file: bdt.xml
<HEADER> DataSetInfo : [Default]: Added class "Signal"
<HEADER> DataSetInfo : [Default]: Added class "Background"
: Booked classifier "BDT" of type: "BDT"

In [16]: dphis = np.linspace(0,np.pi)
vals = []
for v in dphis:
var[0][0] = 100000
var[1][0] = 30000
var[2][0] = 30000
var[3][0] = 50000
var[4][0] = 10000
var[5][0] = 0.1
var[6][0] = v
vals.append(reader.EvaluateMVA("BDT method"))

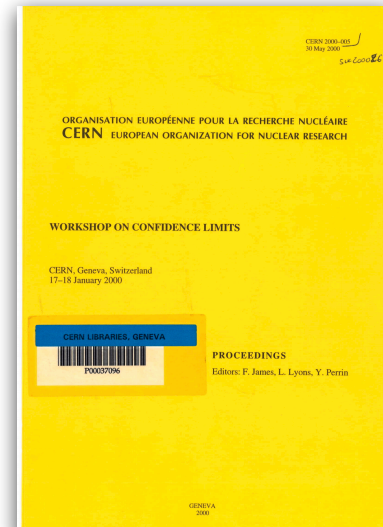
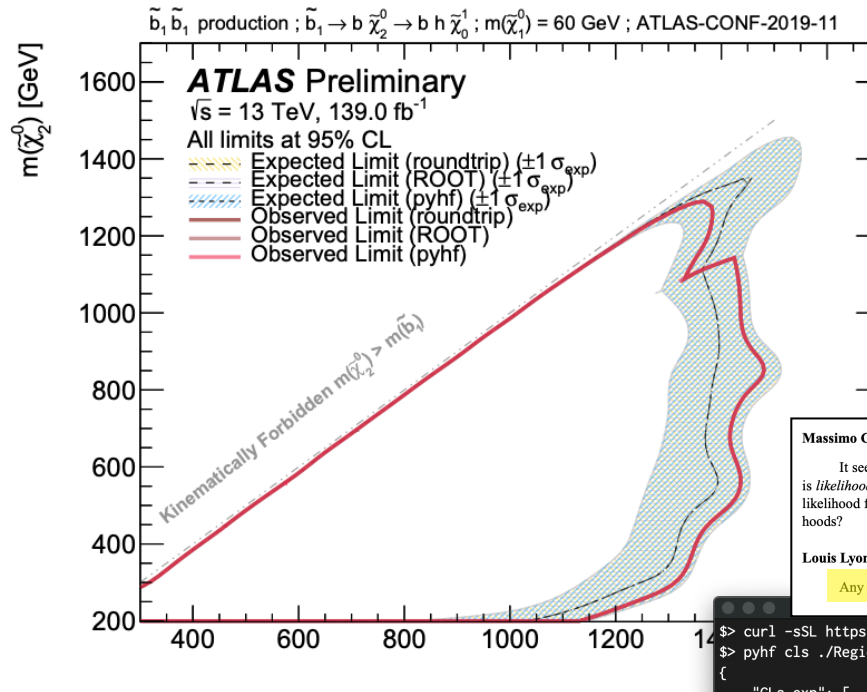
plt.plot(dphis,vals)

Out[16]: [!matplotlib.lines.Line2D at 0x7f9a41e490b8]
```



# New Open Data milestone

- First release of a full likelihood function of a LHC experiment
  - same statistical model (all nuisance parameters) as used in original result.
  - suitable for combination, reinterpretation, etc.



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.

```
$> curl -sSL https://doi.org/10.17182/hepdata.89408.v1/r2|tar -xzf -
$> pyhf cls ./RegionA/BkgOnly.json --patch ./RegionA/patch.sbottom_1400_205_60.json
{
  "Cls_exp": [
    0.144917462643256,
    0.2711393410163219,
    0.47356382348098147,
    0.7268476082357731,
    0.921266748177125
  ],
  "Cls_obs": 0.3439853745556398
}
```

Additional Publication Resources

filter

Common Resources

- Missing Transverse Energy 2
- Effective Mass 2
- Object Based Missing Transverse Energy significance 2
- MaxMin alternative algorithm average  $m_{\text{hcard}}$  2
- Leading jet pT 2
- MaxMin algorithm  $m_{\text{hcard}}$  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
- Acceptance\_SRA\_L\_dm130 2
- Acceptance\_SRC\_incl 2
- Acceptance\_SRA\_L\_m60 2

External Link

Web page with auxiliary material

View Resource

gz File

Archive of full likelihoods in the HistFactory JSON format described in ATL-PHYS-PUB-2019-029. Provided are 3 statistical models labeled RegionA, RegionB and RegionC respectively each in their own sub-directory. For each model the background-only model is found in the file named 'BkgOnly.json'. For each model a set of patches for various signal points is provided.

Download

**ATLAS has a rich data and analysis preservation program**

**Both internal and external preservation to maximize exploitation of ATLAS data.**

**Focus on Outreach & Education for Open Access event-level data**

**For research purposes focus on**

- **high quality data products**
- **joint work with external researchers through ATLAS association mechanisms.**

**Landscape is changing: ATLAS is reviewing its policies.**