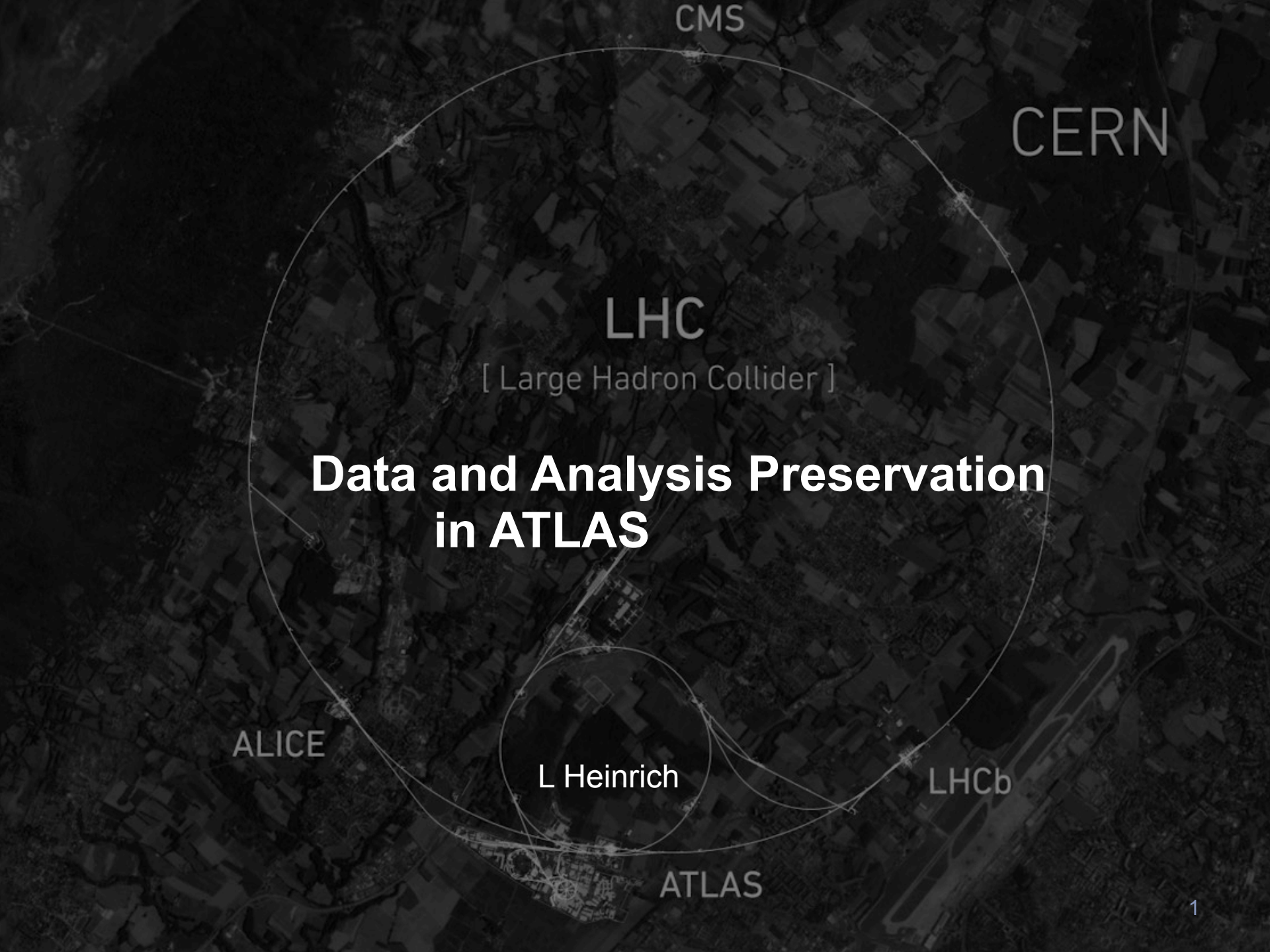


Data and Analysis Preservation in ATLAS

L Heinrich

Data Preservation Policies:

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

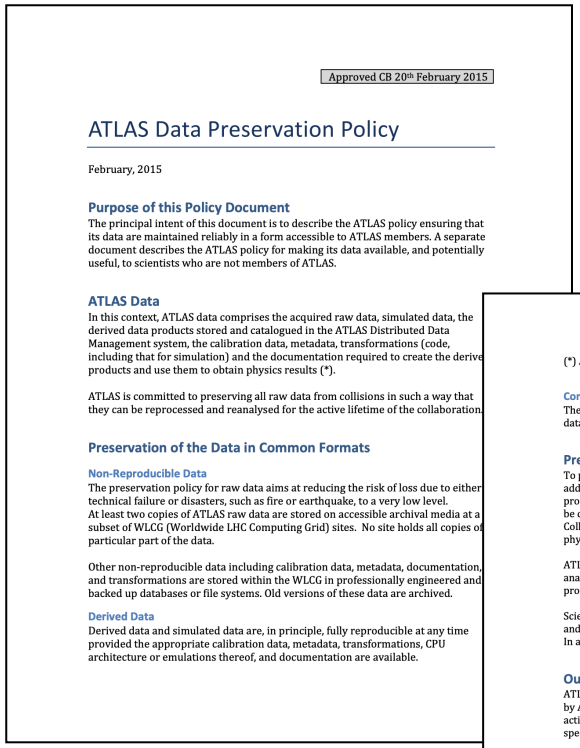
Rely on software preservation for ability to reproduce downstream formats. Detailed database on data provenance.

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 20th 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 derive 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.

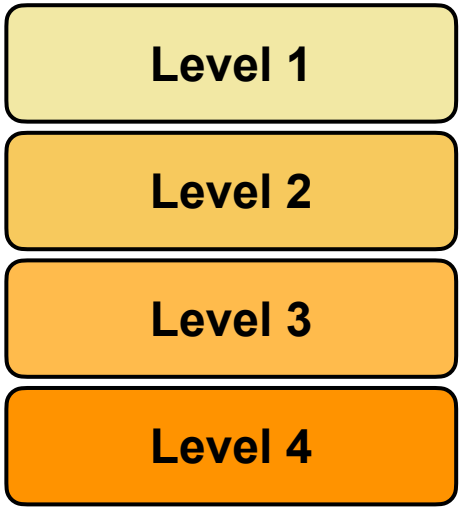
Data Preservation Beyond the Lifetime of the Collaboration

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:

DPHEP Data Nomenclature
Level 1 - Level 4 (document)



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).

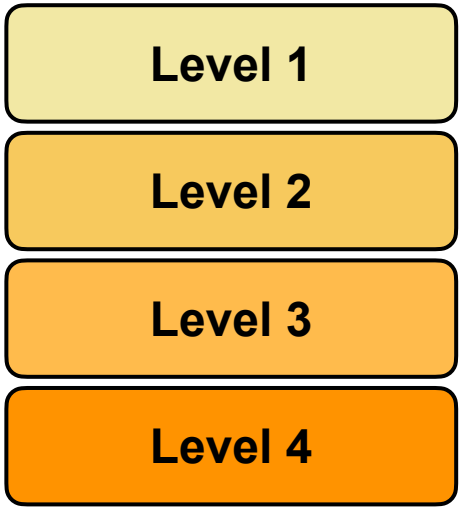
Primary target for open access data repository: HepData

- 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:

DPHEP Data Nomenclature
Level 1 - Level 4 (document)



Level 2: special purpose datasets.

Open Data (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)**

Data Access Policies:

DPHEP Data Nomenclature
Level 1 - Level 4 (document)

Short Term Association (STA)

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

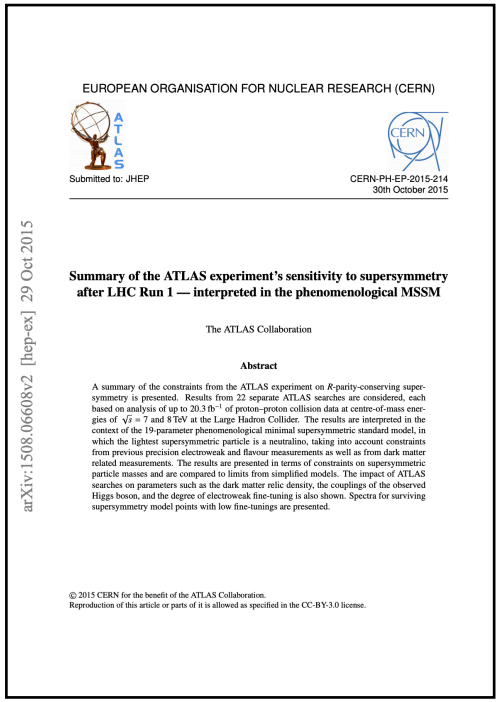
Level 1

Level 2

Level 3

Level 4

Example: Run-1 Summary Publications



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

Submitted to: JHEP

CERN-PH-EP-2015-214
30th October 2015

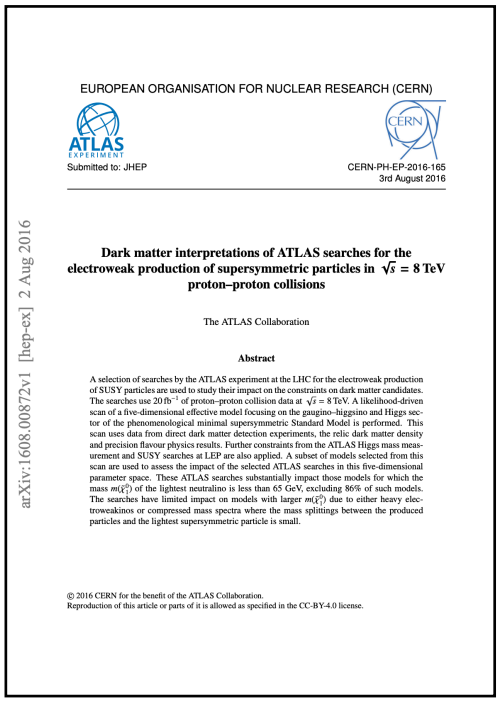
arXiv:1508.06608v2 [hep-ex] 29 Oct 2015

Summary of the ATLAS experiment's sensitivity to supersymmetry after LHC Run 1 — interpreted in the phenomenological MSSM

The ATLAS Collaboration

Abstract

A summary of the constraints from the ATLAS experiment on R-parity-conserving supersymmetry is presented. Results from 22 separate ATLAS searches are considered, each based on analysis of up to 20.3 fb^{-1} of proton–proton collision data at centre-of-mass energies of $\sqrt{s} = 7$ and 8 TeV at the Large Hadron Collider. The results are interpreted in the context of the 19-parameter phenomenological minimal supersymmetric standard model, in which the lightest supersymmetric particle is a neutralino, taking into account constraints from previous precision electroweak and flavour measurements as well as from dark matter related measurements. The results are presented in terms of constraints on supersymmetric particle masses and are compared to limits from simplified models. The impact of ATLAS searches on parameters such as the dark matter relic density, the couplings of the observed Higgs boson, and the degree of electroweak fine-tuning is also shown. Spectra for surviving supersymmetry model points with low fine-tunings are presented.



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

Submitted to: JHEP

CERN-PH-EP-2016-165
3rd August 2016

arXiv:1608.00872v1 [hep-ex] 2 Aug 2016

Dark matter interpretations of ATLAS searches for the electroweak production of supersymmetric particles in $\sqrt{s} = 8$ TeV proton–proton collisions

The ATLAS Collaboration

Abstract

A selection of searches by the ATLAS experiment at the LHC for the electroweak production of SUSY particles are used to study their impact on the constraints on dark matter candidates. The searches use 20 fb^{-1} of proton–proton collision data at $\sqrt{s} = 8$ TeV. A likelihood-driven scan of a five-dimensional effective model focusing on the gaugino–higgsino and Higgs sector of the phenomenological minimal supersymmetric Standard Model is performed. This scan uses data from direct dark matter detection experiments, the relic dark matter density and precision flavour physics results. Further constraints from the ATLAS Higgs mass measurement and SUSY searches at LEP are also applied. A subset of models selected from this scan are used to assess the impact of the selected ATLAS searches in this five-dimensional parameter space. These ATLAS searches substantially impact those models for which the mass $m(\tilde{\chi}_1^0)$ of the lightest neutralino is less than 65 GeV, excluding 86% of such models. The searches have limited impact on models with larger $m(\tilde{\chi}_1^0)$ due to either heavy electroweakinos or compressed mass spectra where the mass splittings between the produced particles and the lightest supersymmetric particle is small.

Data Access Policies:

DPHEP Data Nomenclature
Level 1 - Level 4 (document)

Analysis Consultants & Experts (ACE):

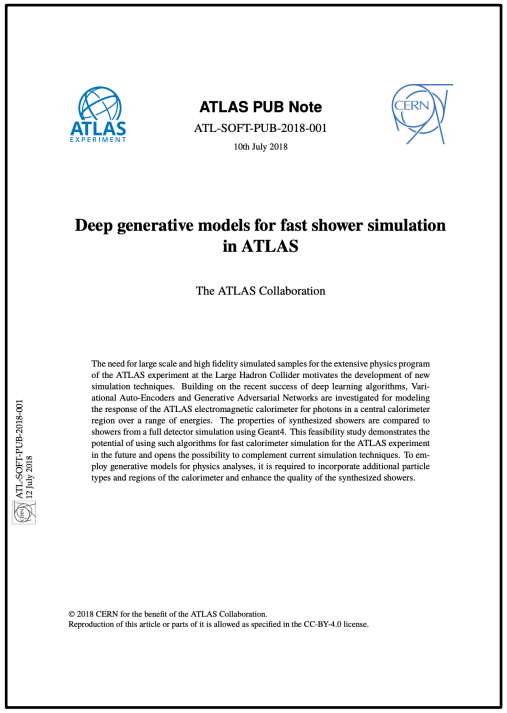
- 48 ACE since 2016
- access to full ATLAS Monte Carlo for e.g. R&D in fast simulation
- public document signed by ATLAS collaboration
- resulting datasets may become public after publication
  - explicit possibility to publish ML datasets
- credit through acknowledgement reference to method paper (exceptionally: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

Level 1

Level 2

Level 3

Level 4



ATLAS EXPERIMENT

ATLAS PUB Note
ATL-SOFT-PUB-2018-001
10th July 2018

CERN

Deep generative models for fast shower simulation in ATLAS

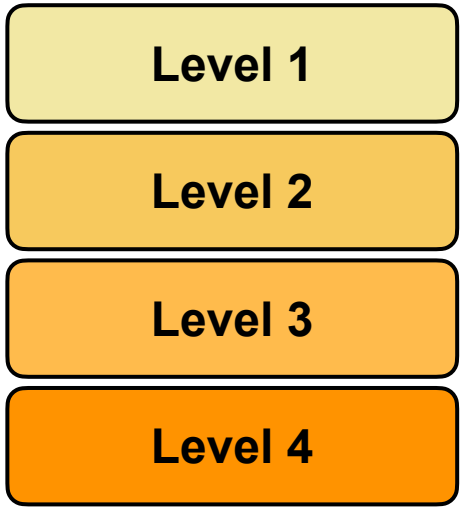
The ATLAS Collaboration

The need for large scale and high fidelity simulated samples for the extensive physics program of the ATLAS experiment at the Large Hadron Collider motivates the development of new simulation techniques. Building on the recent success of deep learning algorithms, Variational Auto-Encoders and Generative Adversarial Networks are investigated for modeling the response of the ATLAS electromagnetic calorimeter for photons in a central calorimeter region over a range of energies. The properties of synthesized showers are compared to showers from a full detector simulation using Geant4. This feasibility study demonstrates the potential of using such algorithms for fast calorimeter simulation for the ATLAS experiment in the future and opens the possibility to complement current simulation techniques. To employ generative models for physics analyses, it is required to incorporate additional particle types and regions of the calorimeter and enhance the quality of the synthesized showers.

ATL-SOFT-PUB-2018-001
12 July 2018

Data Access Policies:

DPHEP Data Nomenclature
Level 1 - Level 4 (document)



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

Data Access Review

DPHEP Data Nomenclature
Level 1 - Level 4 (document)

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

Level 1

Level 2

Level 3

Level 4

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

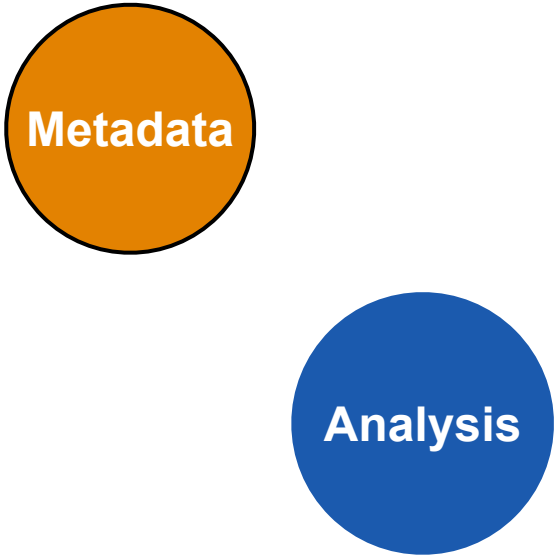
"The Hangar"

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

Both are important and being worked on within ATLAS & CERN

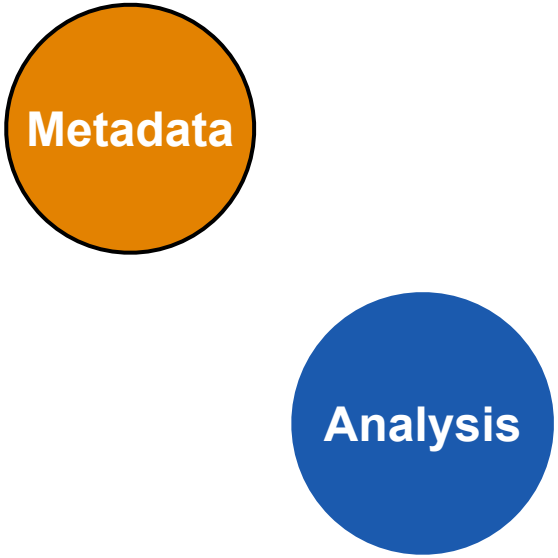
Both are important and being worked on within ATLAS & CERN

Analysis Team
Internal Notes
Bibliographic Info
…



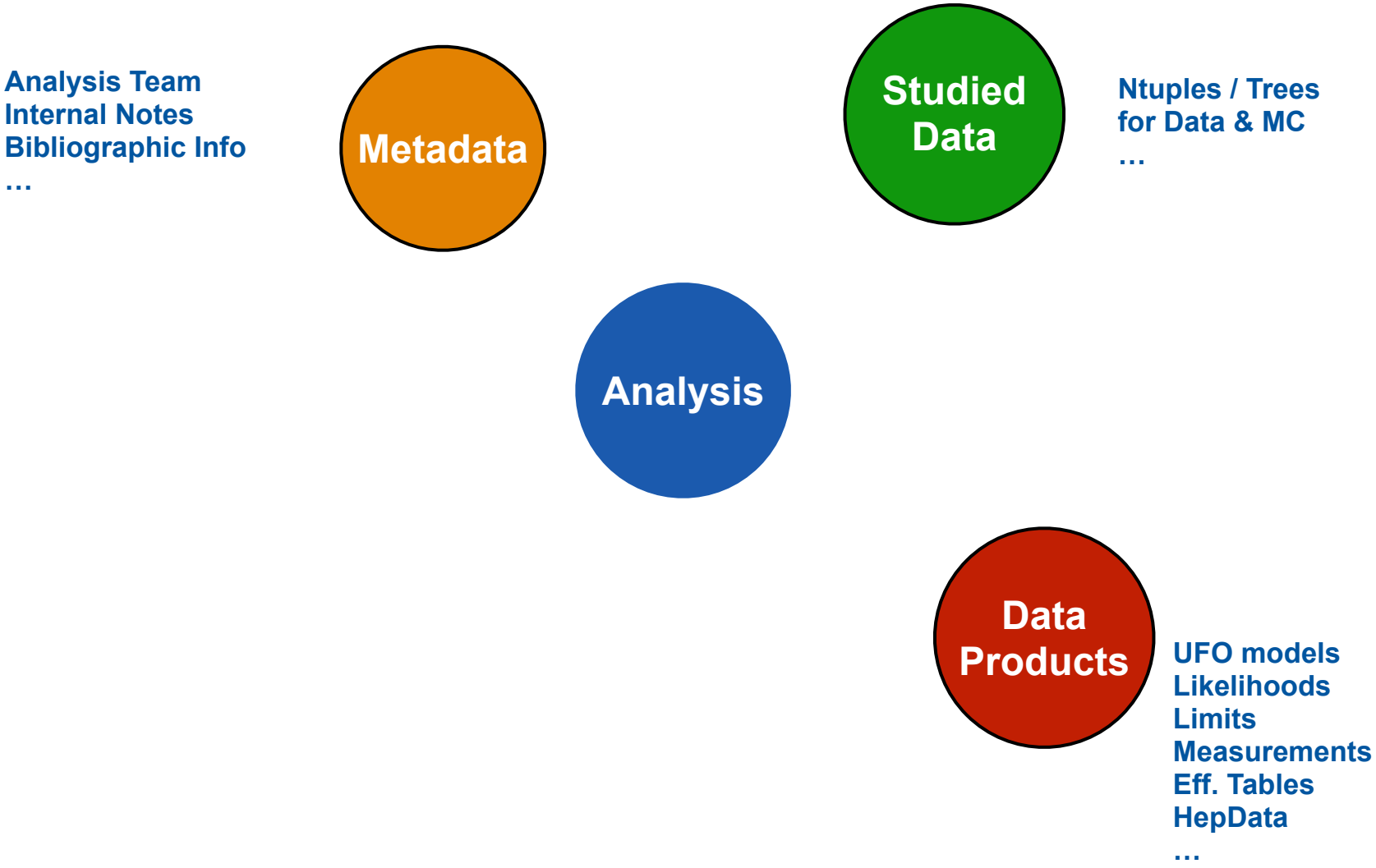
Both are important and being worked on within ATLAS & CERN

Analysis Team
Internal Notes
Bibliographic Info
…

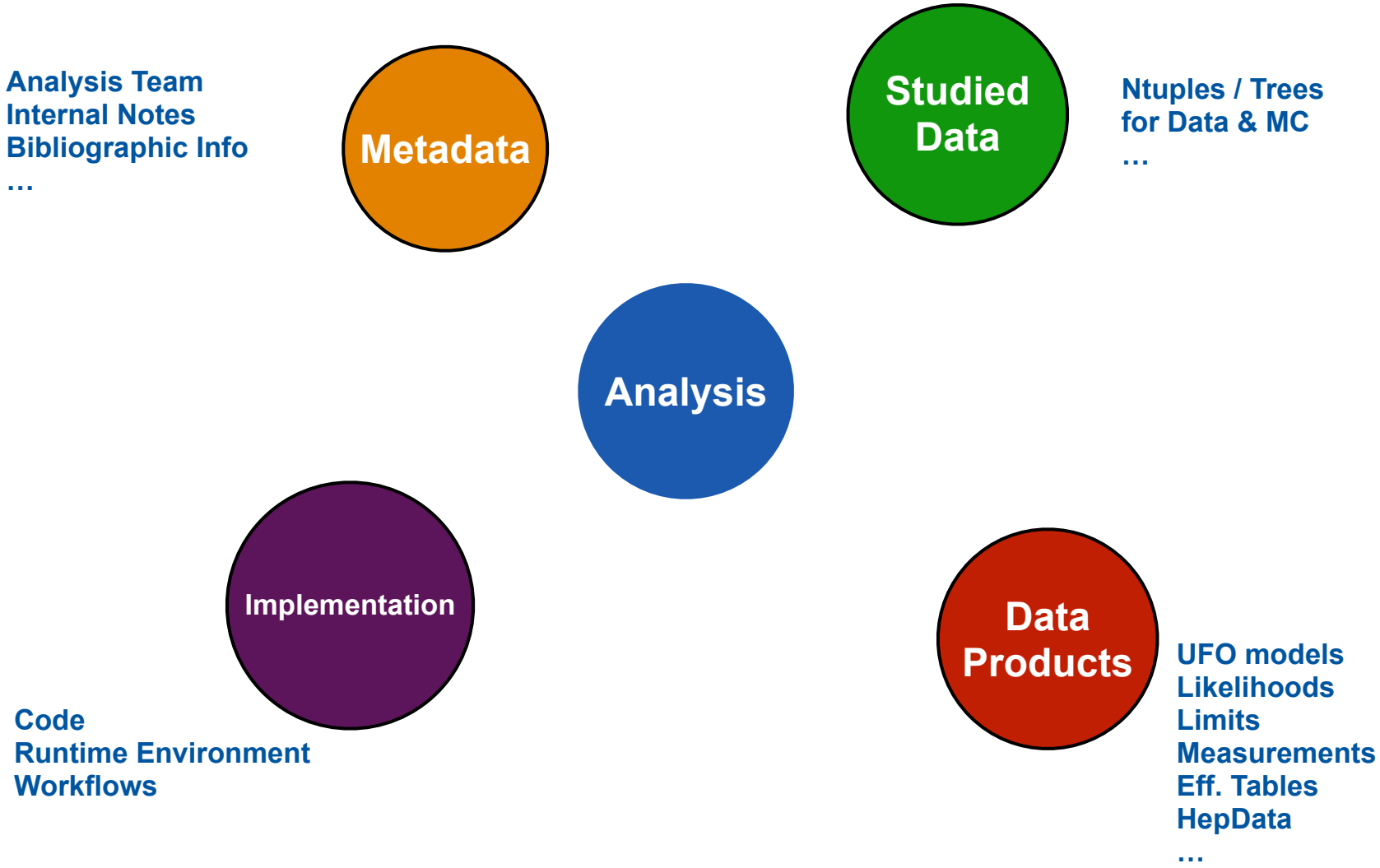


Ntuples / Trees
for Data & MC
…

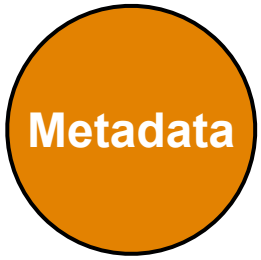
Both are important and being worked on within ATLAS & CERN



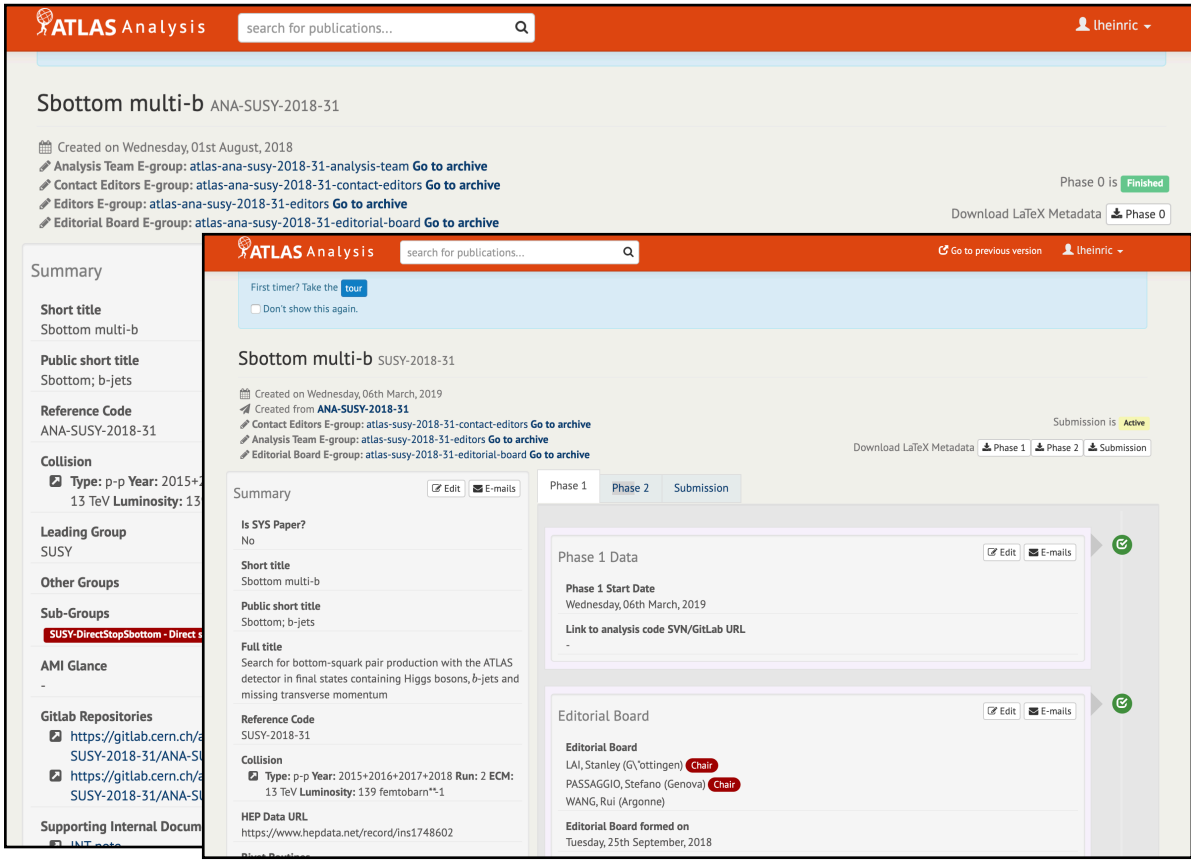
Both are important and being worked on within ATLAS & CERN



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

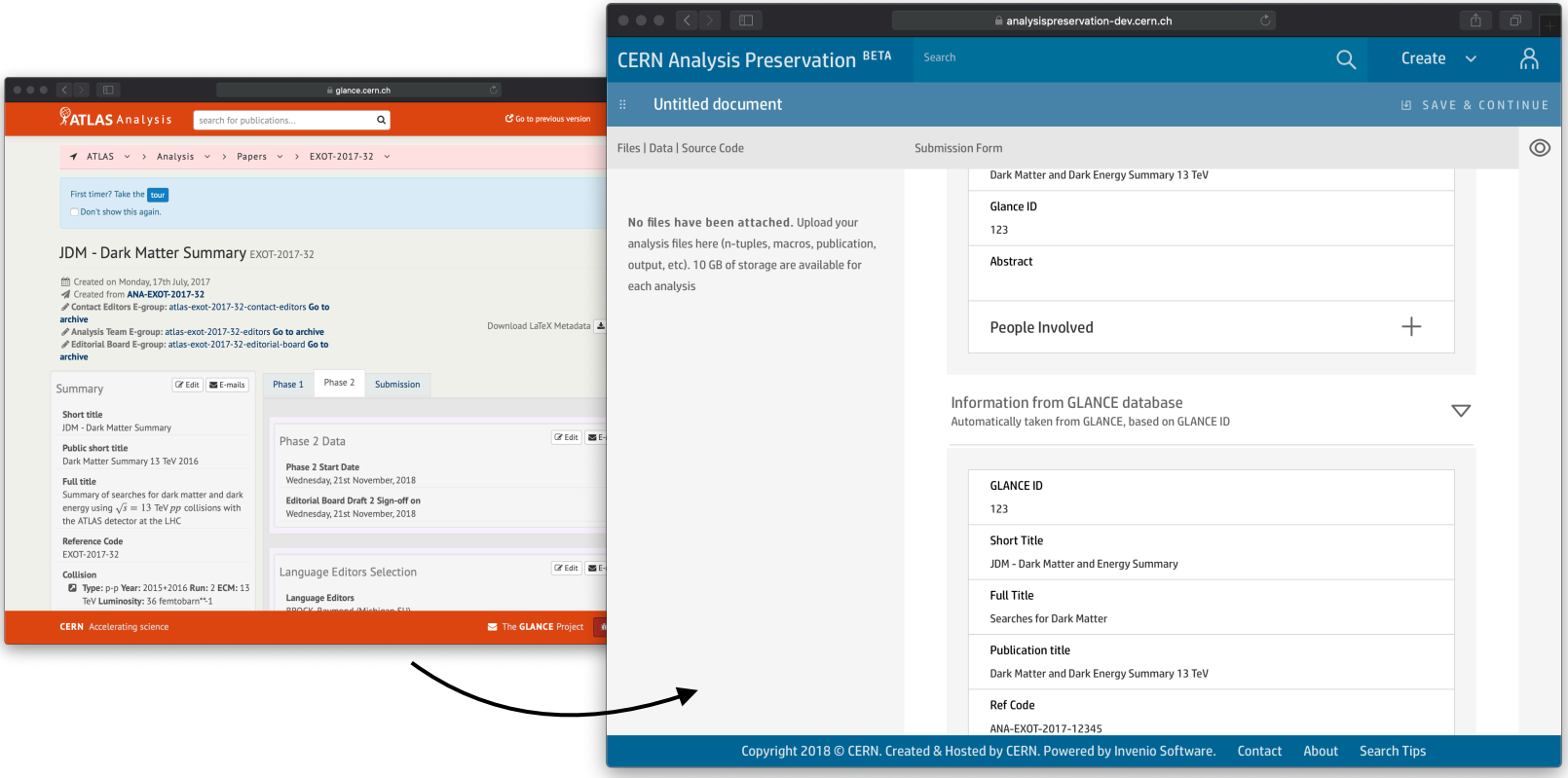


Automatic Import into CERN Analysis Preservation

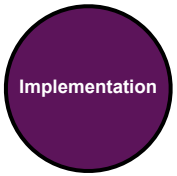
Metadata

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

CERN Analysis Preservation



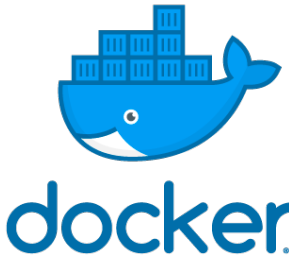
ATLAS is investing in re-useable / re-producible analysis

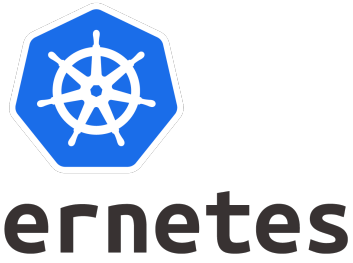


Technology Choice for software archival:

- Git
- Linux Containers

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

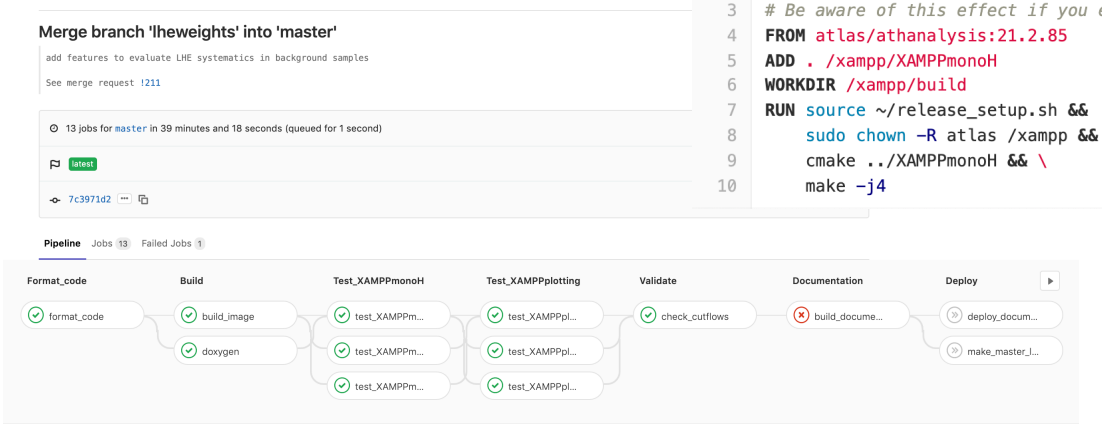




Containers in ATLAS: reproducible software environments

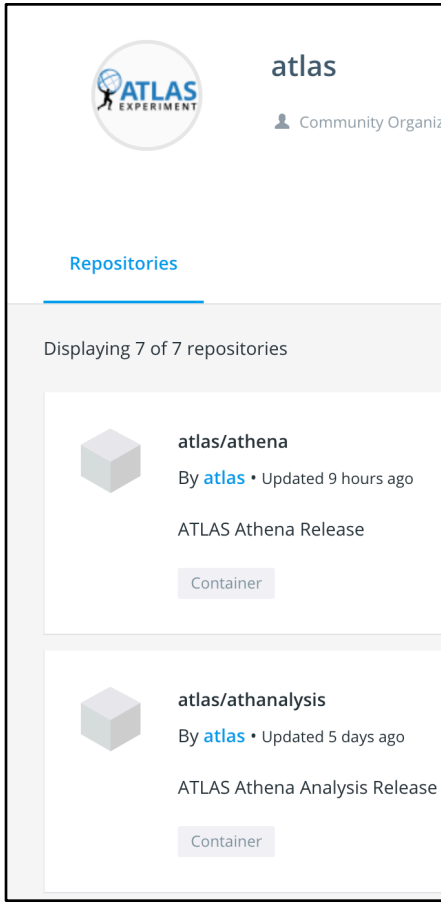
Implementation

- integrated in Analysis Software Release Schedule
- teach continuous testing / validation / preservation in ATLAS induction / software tutorials
- integrated into distributed computing infrastructure (containers on Grid)
- used >1000 times per day for CI etc



Dockerfile 400 Bytes

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

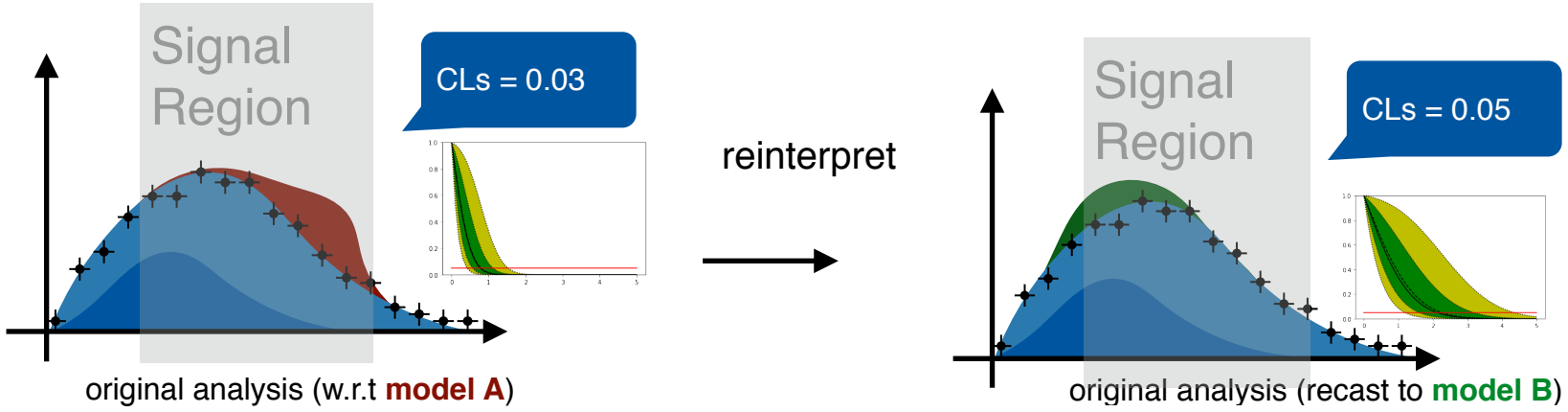


Major physics groups have adopted Analysis Preservation as part of their approval procedure.

Implementation

Currently focused on BSM program (SUSY, Exotics, Higgs/Diboson Searches)

Main use-case: RECAST (reinterpreting searches)



original analysis (w.r.t **model A**)

original analysis (recast to **model B**)

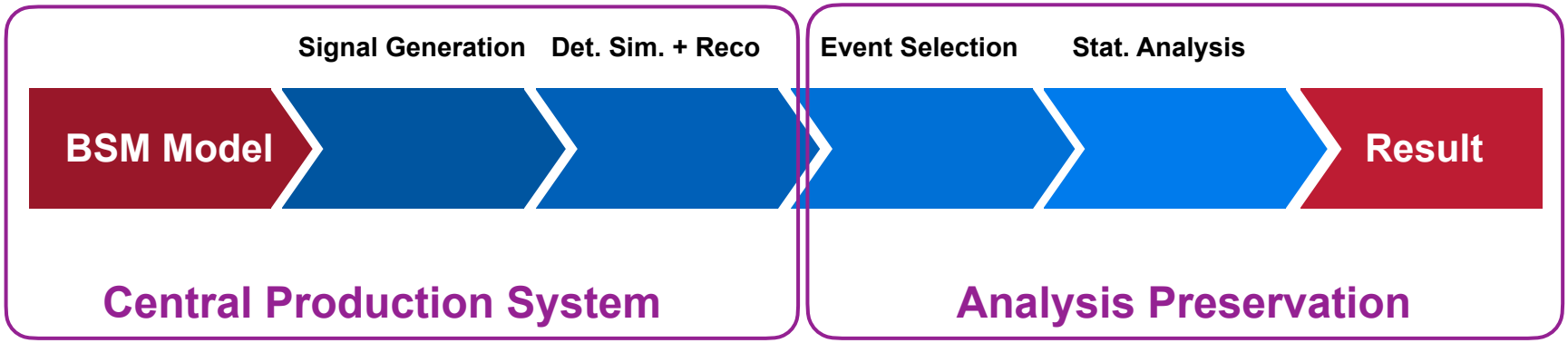
For operational analysis preservation need to preserve full pipeline. Demarcation line: central production.

Implementation

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:

Implementation

1. **capture software**

   container images

2. **capture commands**

   job templates

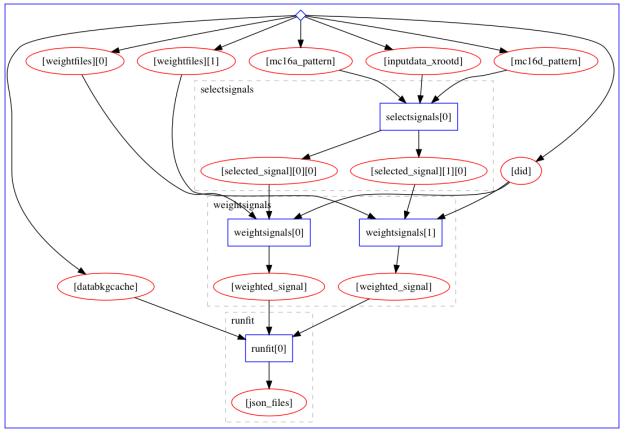
3. **capture workflow**

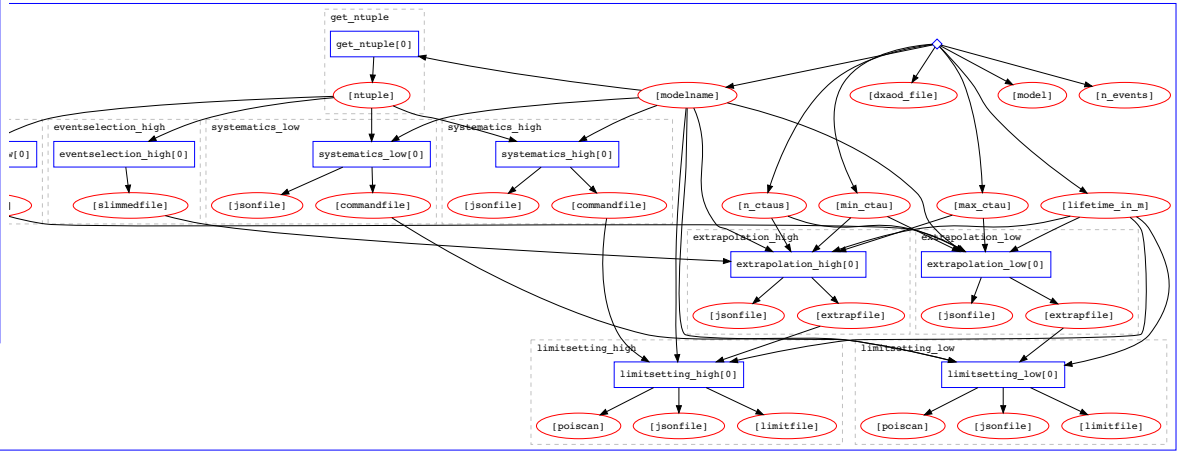
   how do I connect the pieces

working with CERN Analysis Preservation & REANA teams to make ATLAS Analyses

CERN Analysis Preservation

reana





Preservation of Code, Scripts, Workflows:

Implementation

1. **capture software**

   container images

2. **capture commands**

   job templates

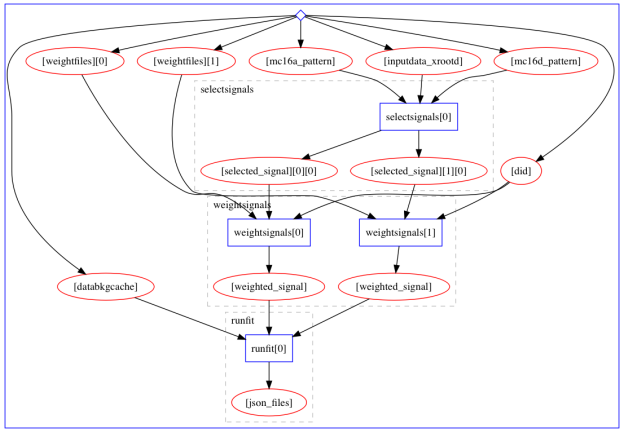
3. **capture workflow**

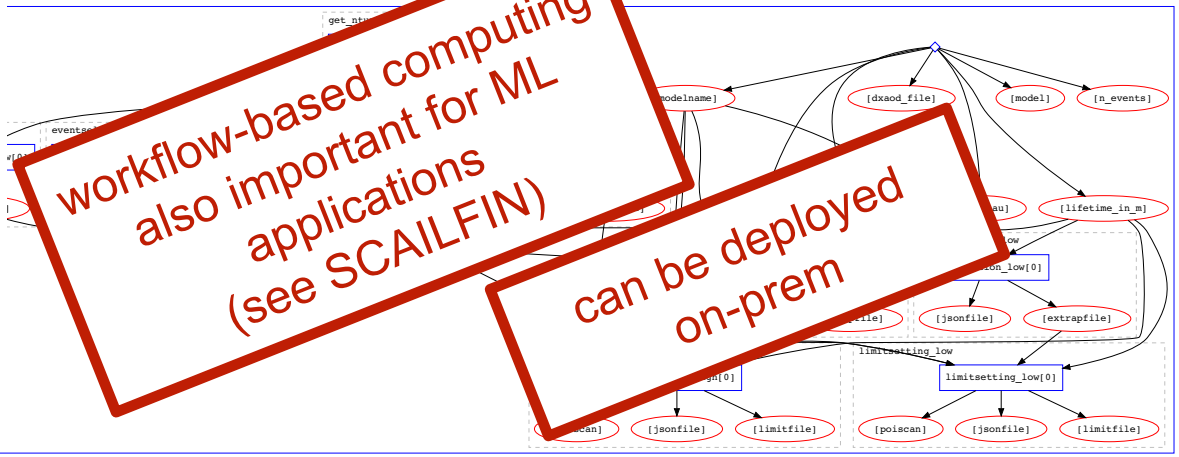
   how do I connect the pieces

working with CERN Analysis Preservation & REANA teams to make ATLAS Analyses

CERN Analysis Preservation

reana





workflow-based computing also important for ML applications (see SCAILFIN)

can be deployed on-prem

Re-execution on independent infrastructure:

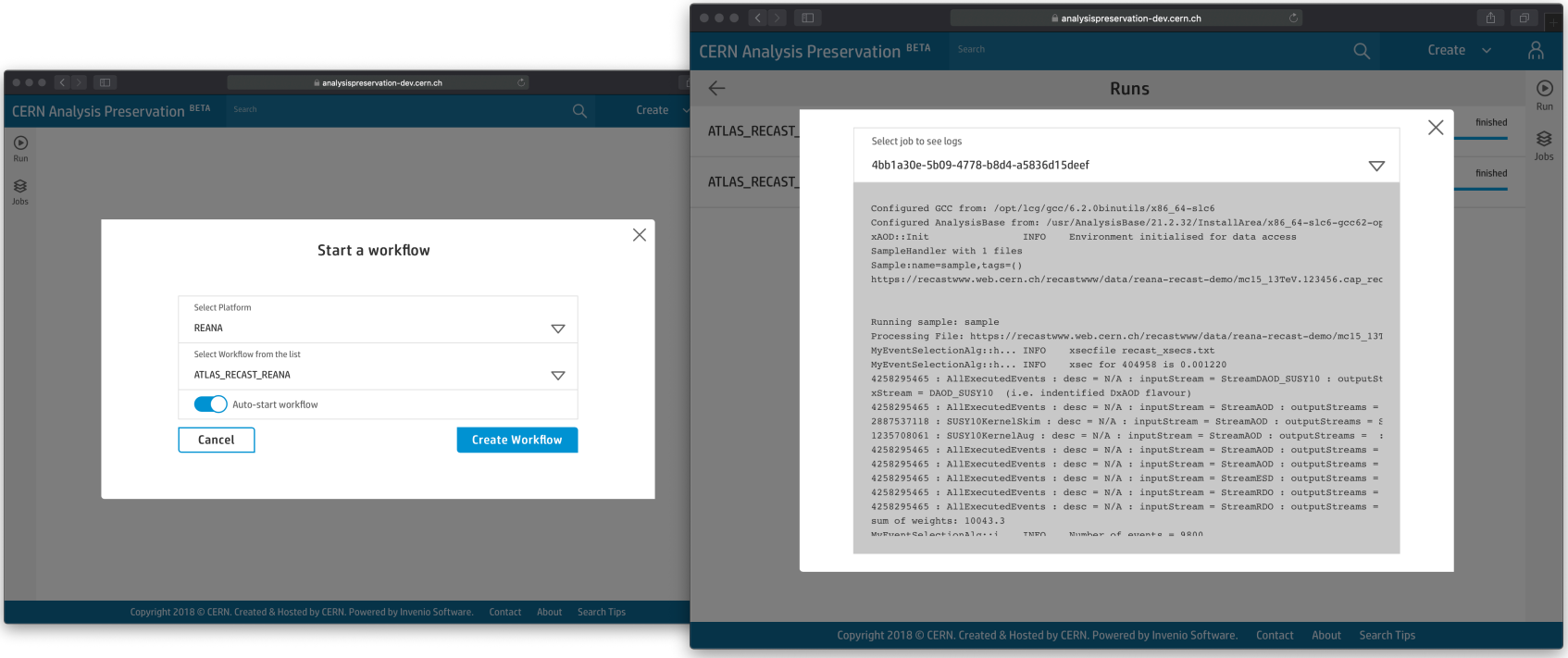
Implementation

Working with CERN projects:

reana

CERN Analysis Preservation

Re-execute analysis from preserved record:



New Results from preserved analyses

Implementation

- currently used to prepare to use for Run-2 summary papers

Example 1: Dark Higgs

new result

ATLAS EXPERIMENT

ATLAS PUB Note

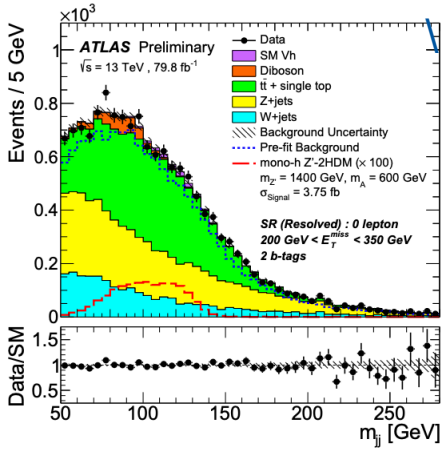
ATL-PHYS-PUB-2019-032

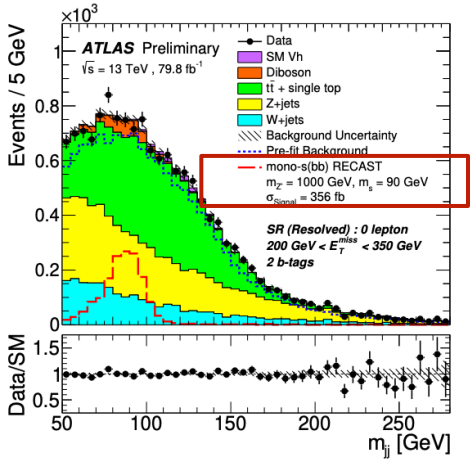
11th August 2019

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 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 fb^{-1} of proton-proton collision data

The ATLAS Collaboration

original publication

New Results from preserved analyses

Implementation

- currently used to prepare to use for Run-2 summary papers

Example 2: Long-lived Particles
- perfect use-case (lots of low-level ML, hard to reproduce externally)

new result

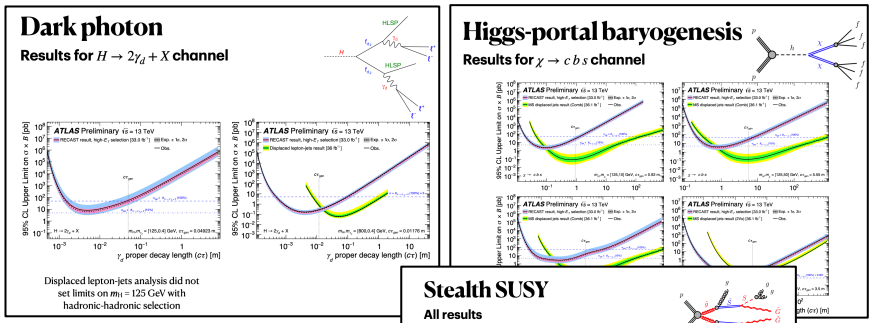
ATLAS EXPERIMENT

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



slides A Morris

Eur. Phys. J. C (2019) 79:481
https://doi.org/10.1140/epjc/s10052-019-6962-6

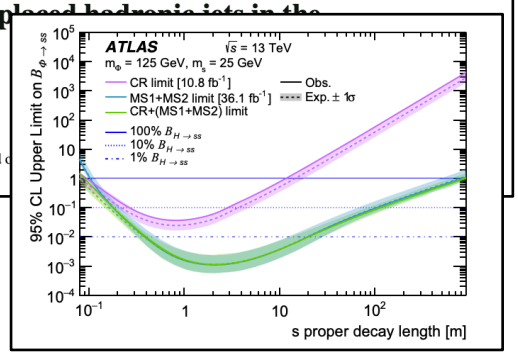
THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Experimental Physics

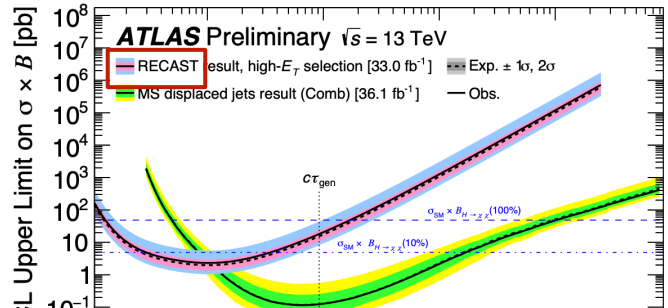
Search for long-lived neutral particles in pp collisions at $\sqrt{s} = 13$ TeV that decay into displaced hadronic jets in the ATLAS calorimeter

ATLAS Collaboration*
CERN, 1211 Geneva 23, Switzerland

Received: 11 February 2019 / Accepted: 18 May 2019 / Published



original publication



New Results from preserved analyses

Implementation

- currently used to prepare to use for Run-2 summary papers

Example 2: Long-lived Particles

- perfect use-case (lots of low-level ML, hard to reproduce externally)

slowly a new paradigm of systematic reuse is emerging.

new result

ATLAS EXPERIMENT

ATLAS PUB Note

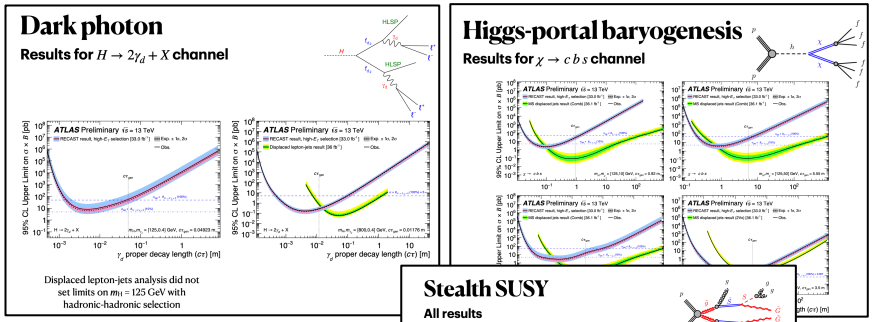
ATL-PHYS-PUB-2020-007

27th March 2020

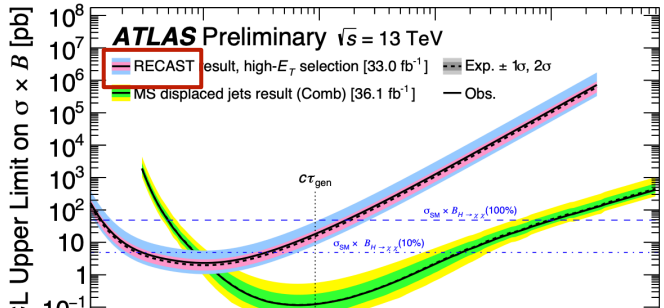
CERN

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

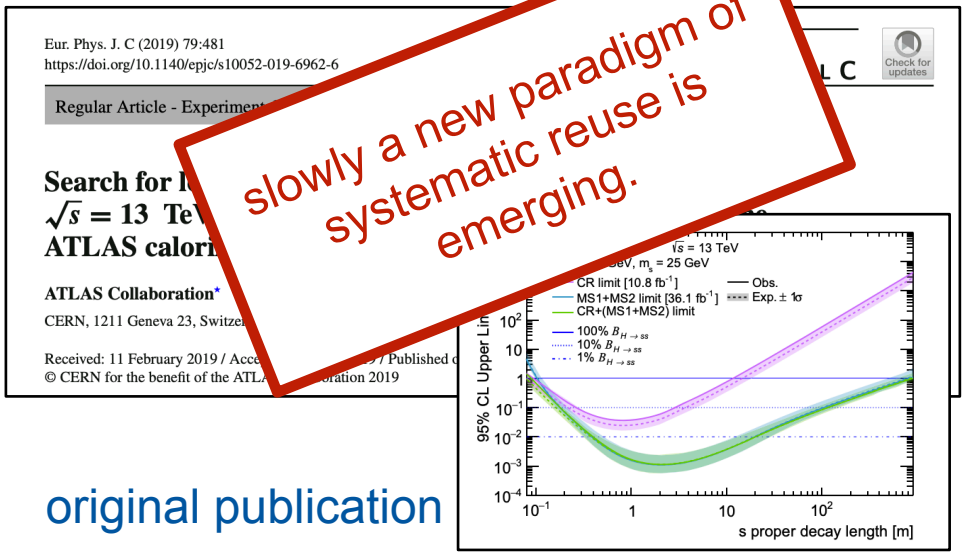
The ATLAS Collaboration



slides A Morris



original publication



New Results from preserved analyses

Implementation

- currently used to prepare to use for Run-2 summary papers

Example 2: Long-lived Particles

- perfect use-case (lots of low-level ML, hard to reproduce externally)

slowly a new paradigm of systematic reuse is emerging.

for new projects (e.g. EIC) this should be done from the start. The tech is there.

new result

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

Collaboration

Eur. Phys. J. C (2019) 79:481
https://doi.org/10.1140/epjc/s10052-019-6962-6

Regular Article - Experiment

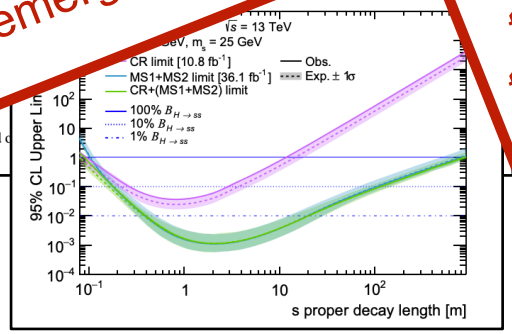
Check for updates

Search for l
$\sqrt{s} = 13$ Te
ATLAS calori

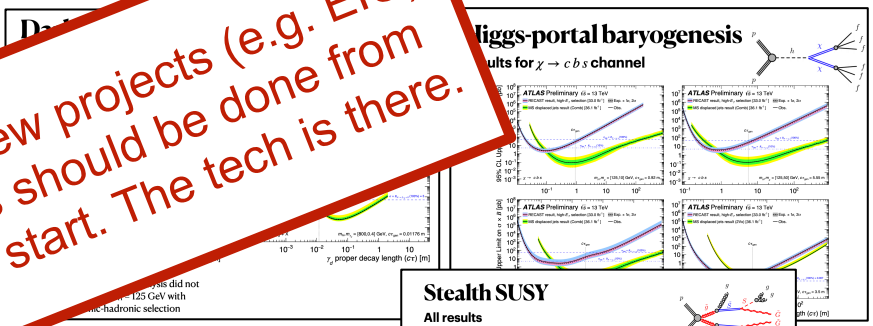
ATLAS Collaboration*

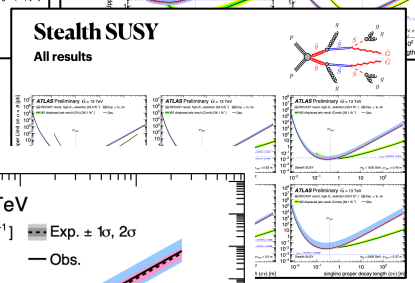
CERN, 1211 Geneva 23, Switze

Received: 11 February 2019 / Acce / Published
© CERN for the benefit of the ATL oration 2019

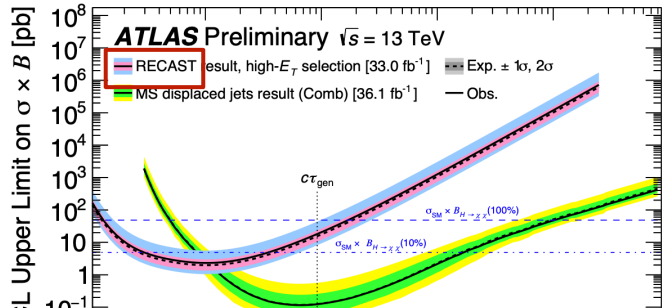


original publication





slides A Morris



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 disciminants (BDTs, etc)

Data Products

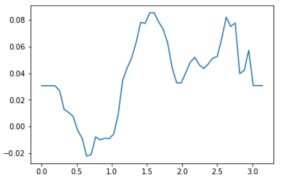
Branch: master ▾ use-atlas-bdt-hepdata / Usage_ATLAS_HepData_BDT.ipynb

lukasheinrich Add files via upload

1 contributor

152 lines (152 sloc) 20.6 KB

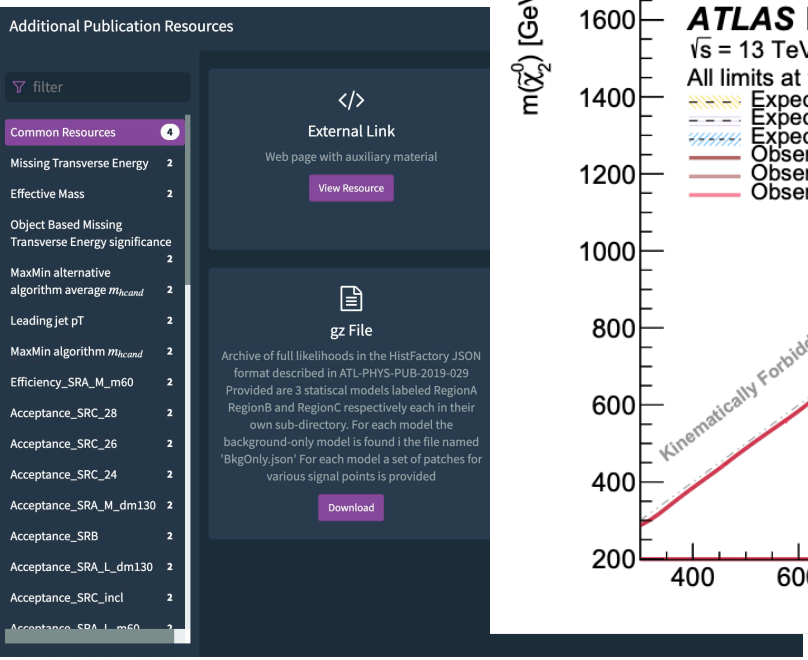
```
In [14]: import ROOT
         import array
         import matplotlib.pyplot as plt
         import numpy as np
         %matplotlib inline

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/4067197vi
Resolving www.hepdata.net (www.hepdata.net)... 188.184.64.140
Connecting to www.hepdata.net (www.hepdata.net)|188.184.64.140|:443... con
HTTP request sent, awaiting response... 200 OK
Length: 953915 (932K) [text/xml]
Saving to: 'bdt.xml'

bdt.xml             100%[===================>] 931.56K   605KB/s    in 1.5

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 0x55cf1e6050e0>

                                  : 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] = 90000
             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>]
```

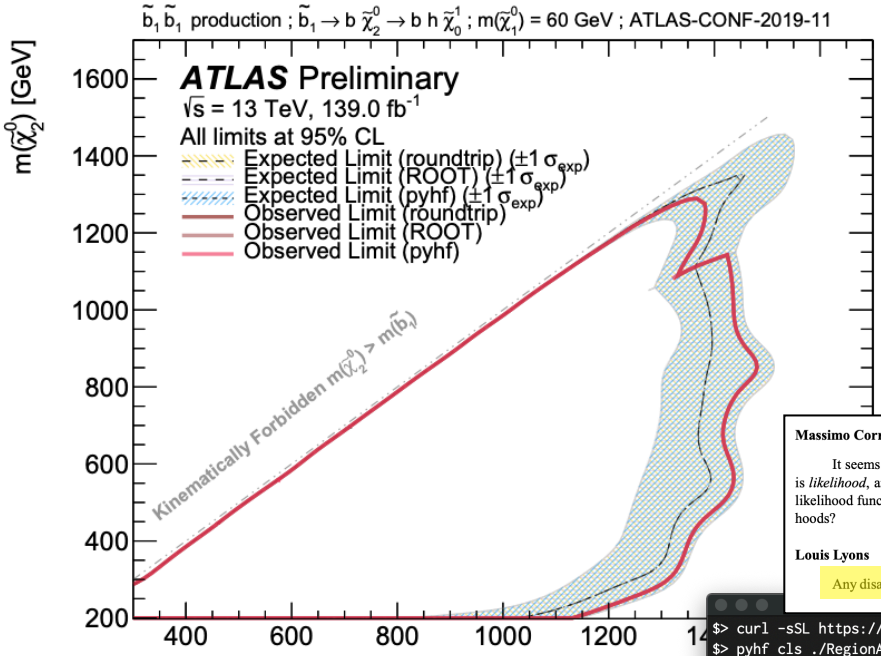


New Open Data milestone

Data Products

- 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
}
$>
```

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.