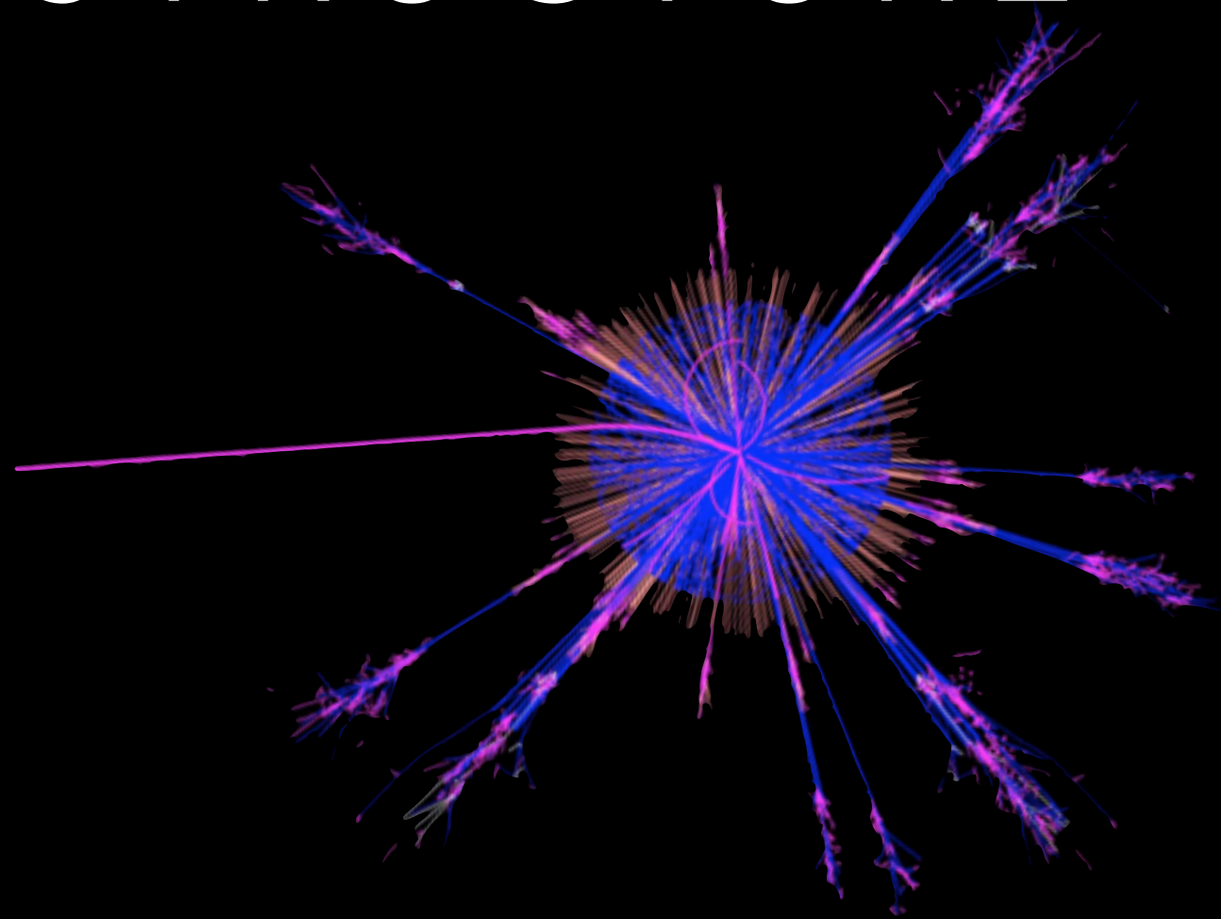
NYU Center for Data Science

Center for Cosmology and particle physics

HEPDATA AND IT'S ROLE IN HEP'S

CYBER INFRASTRUCTURE

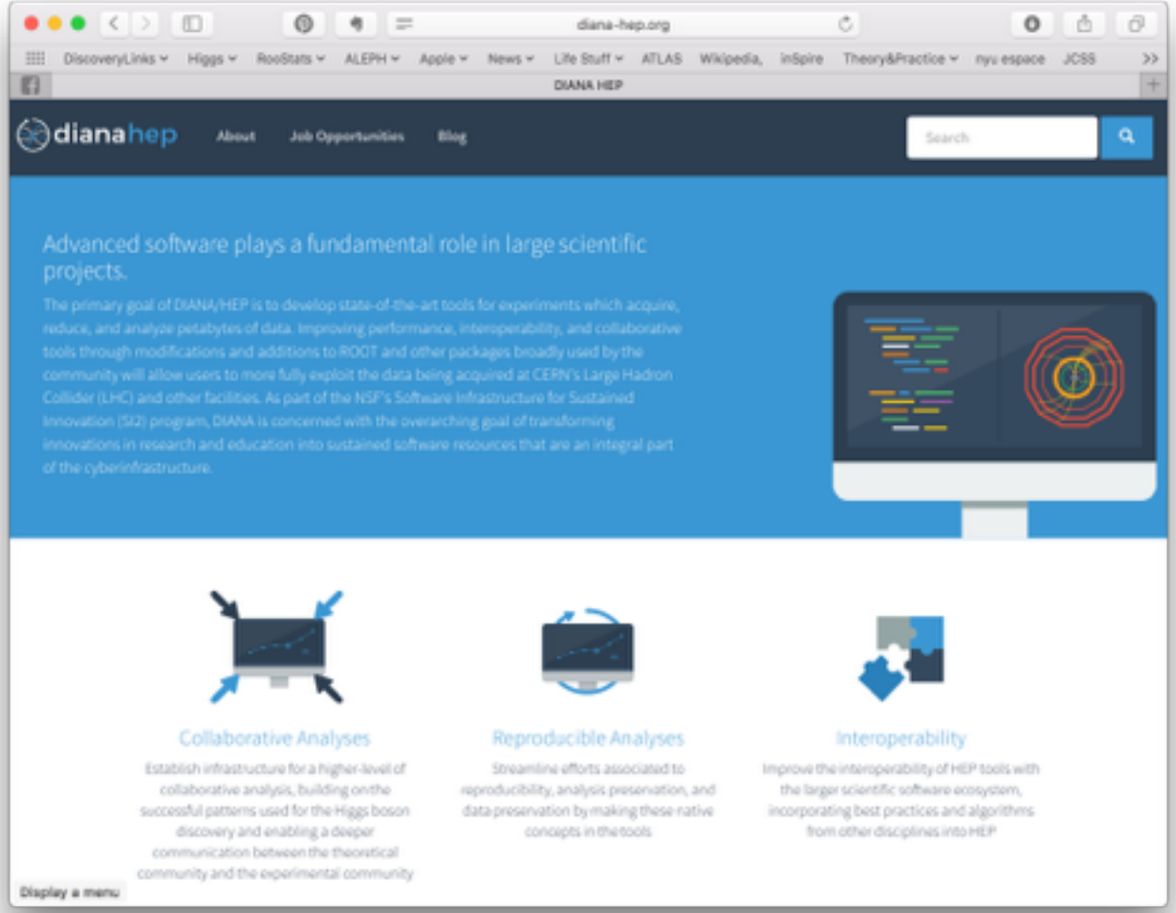


@KyleCranmer
New York University
Department of Physics
Center for Data Science

DASPOS AND DIANA

DASPOS and DIANA are two large projects funded by the National Science Foundation related focusing on issues around software and data for high energy physics.

Working closely with CERN Analysis Preservation (CAP) portal, INSPIRE, and HEPData



ECOSYSTEM OF TOOLS

Goal is to have integrations between these tools.
What do we want to do? What is missing?

HEPData

INSPIRE

CAP

RIVET

RECAST

CDS

Wiki pages

ROOT

GitHub/GitLab

Mathematica

Jupyter/IPython

Zenodo

MadAnalysis
Checkmate
Atom, FastLim, …

MadGraph
Sherpa
Pythia
…

ANALYSIS CATALOGUE

Goal is to have integrations between these tools.
What do we want to do? What is missing?

HEPData INSPIRE CAP

RIVET RECAST CDS Wiki pages

One analysis from a particular collaboration might have records in each of the components above.

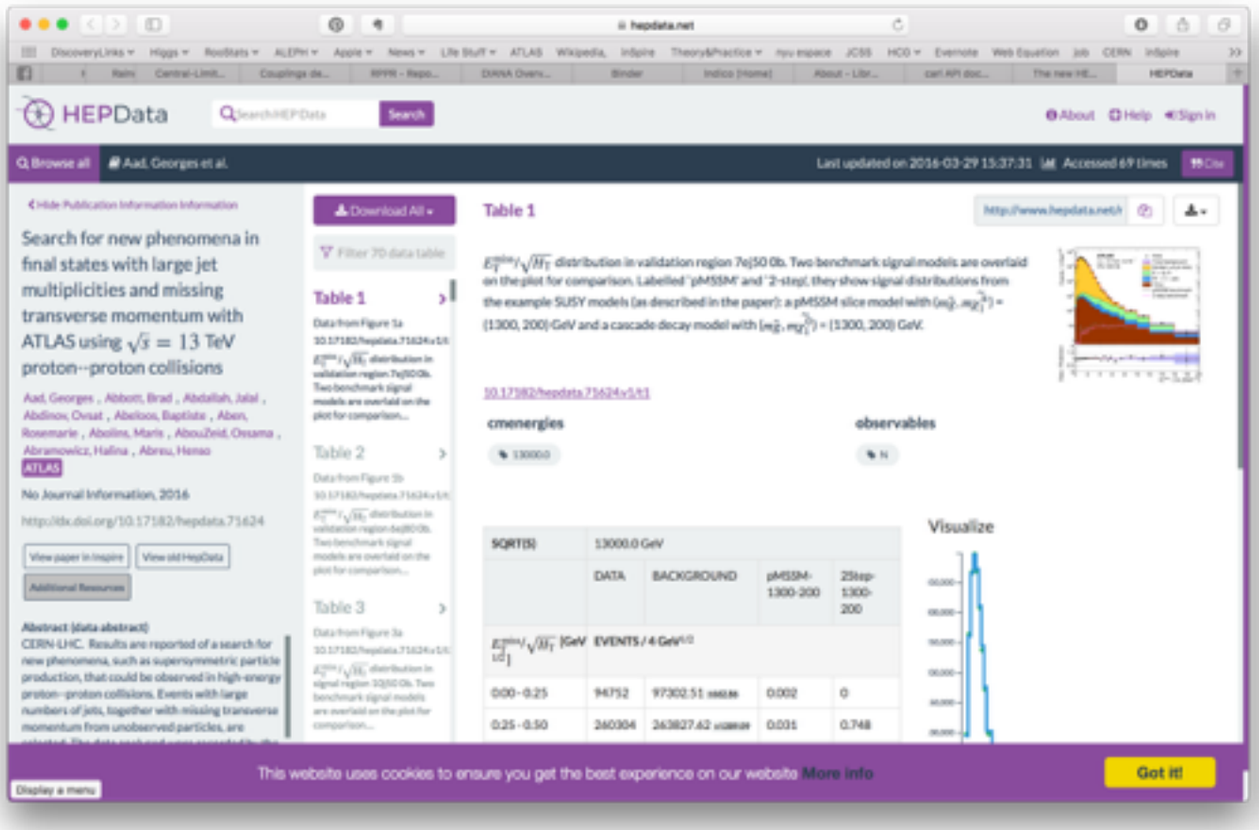
Would be nice to have a unique “key” to identify an analysis.

(probably needs to be more fine grained than publication level)

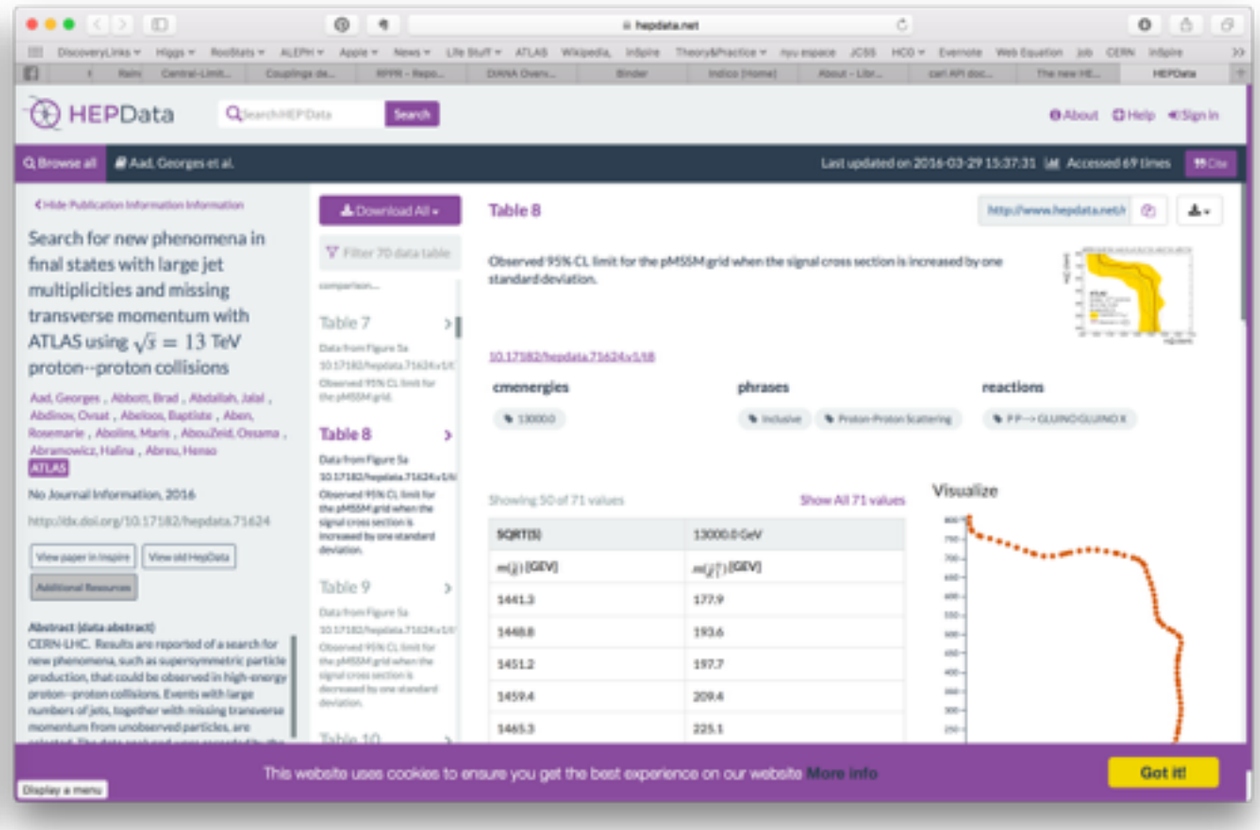
FEEDBACK

There are two main types of data in HEPData

Observations



Interpretation



Would be nice to have meta-data to search based this

Use case: Find all 95% contours for a given simplified model

IMPROVING REPRESENTATION OF PLOTS

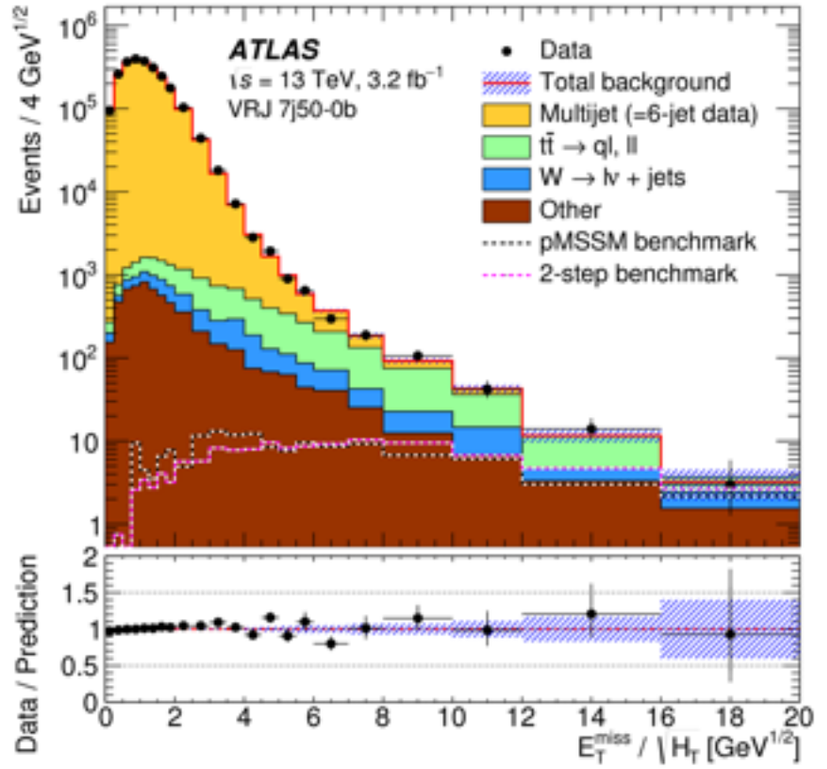
Tools to convert from common ROOT types (histograms, graphs, etc.) to HEPData table greatly improved

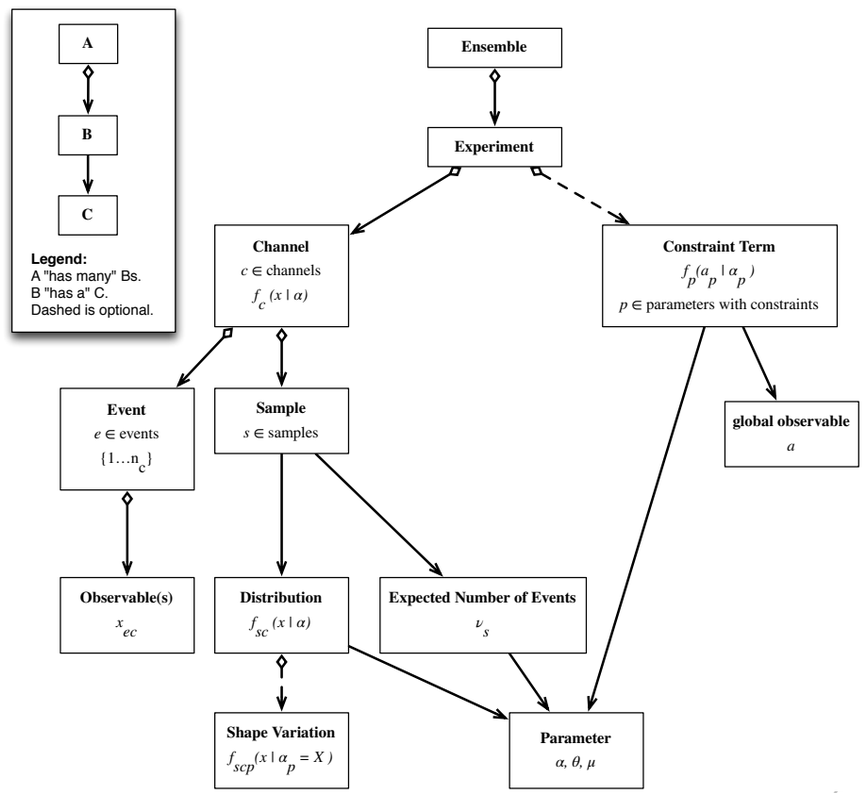
But typical plot has many background components, complicated correlated uncertainties, variational histograms, etc.. Does not fit HEPData tables

- one plot may have hundreds of histograms
- HEPData can store individual histograms, but needs meta-data layer to collect them for one plot and describe what they mean

HistFactory/HistFitter used for large fraction of searches within ATLAS.

- Defines a schema for describing all the components, correlated uncertainties, etc.
- We wrote a converter for the HistFactory XML to yaml, but HEPData not currently able to store this rich structure.





ATL-CB-PUB-2015-001
17 March 2015

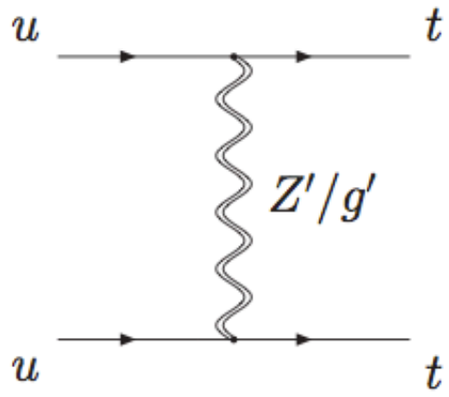
Level-1. Published results

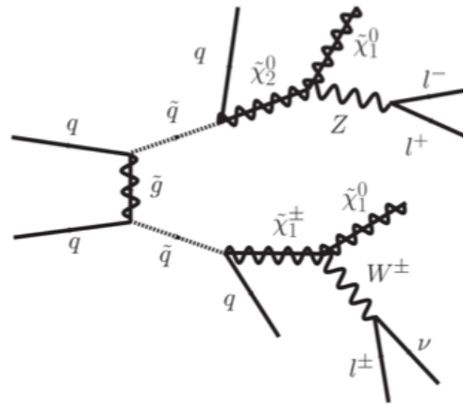
All scientific output is published in journals, and preliminary results are made available in Conference Notes. All are openly available, without restriction on use by external parties beyond copyright law and the standard conditions agreed by CERN.

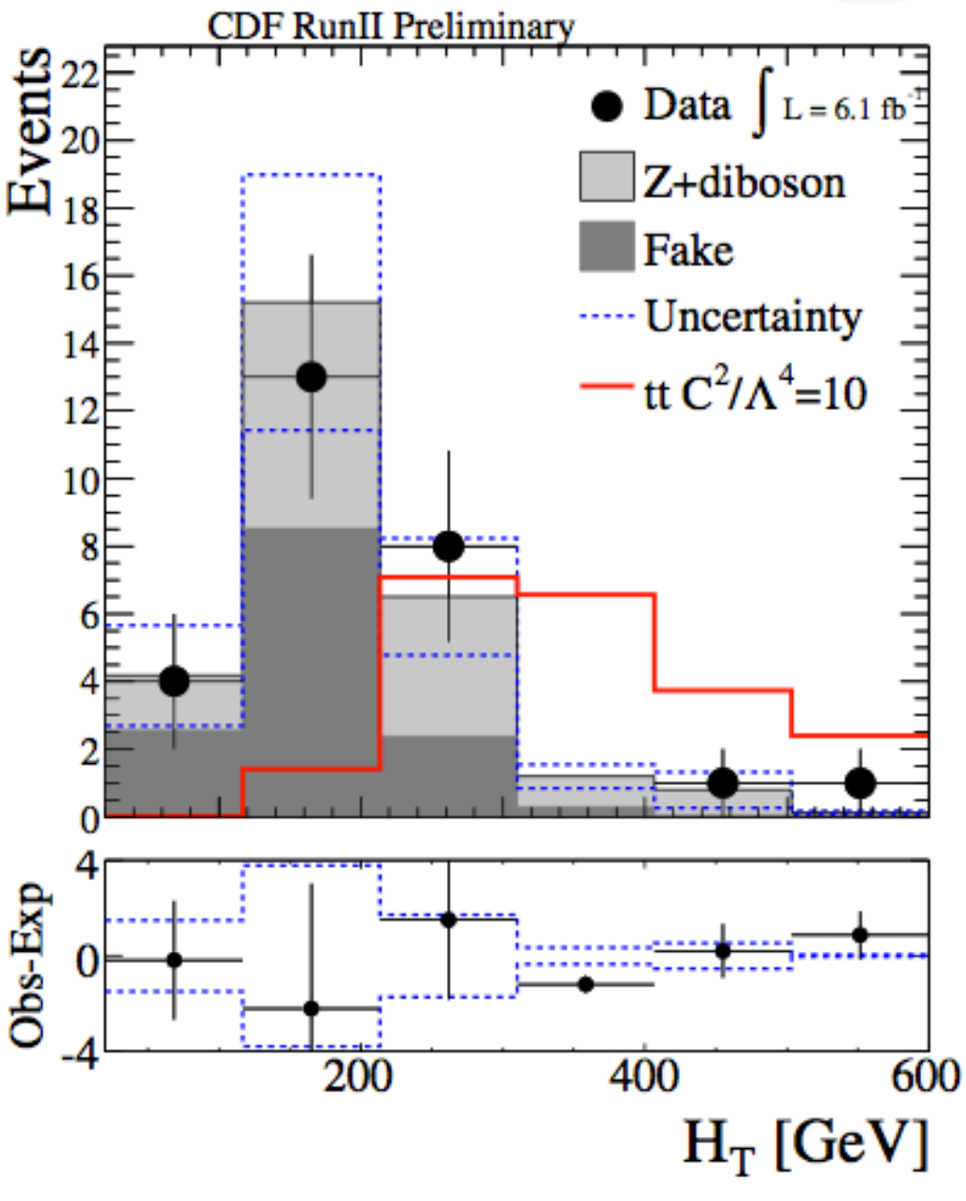
Data associated with journal publications are also made available: tables and data from plots (e.g. cross section values, likelihood profiles, selection efficiencies, cross section limits, …) are stored in appropriate repositories such as HEPDATA[2]. ATLAS also strives to make additional material related to the paper available that allows a reinterpretation of the data in the context of new theoretical models. For example, an extended encapsulation of the analysis is often provided for measurements in the framework of RIVET [3]. For searches 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 [4]. RECAST allows theorists to evaluate the sensitivity of a published analysis to a new model they have developed by submitting their model to ATLAS.

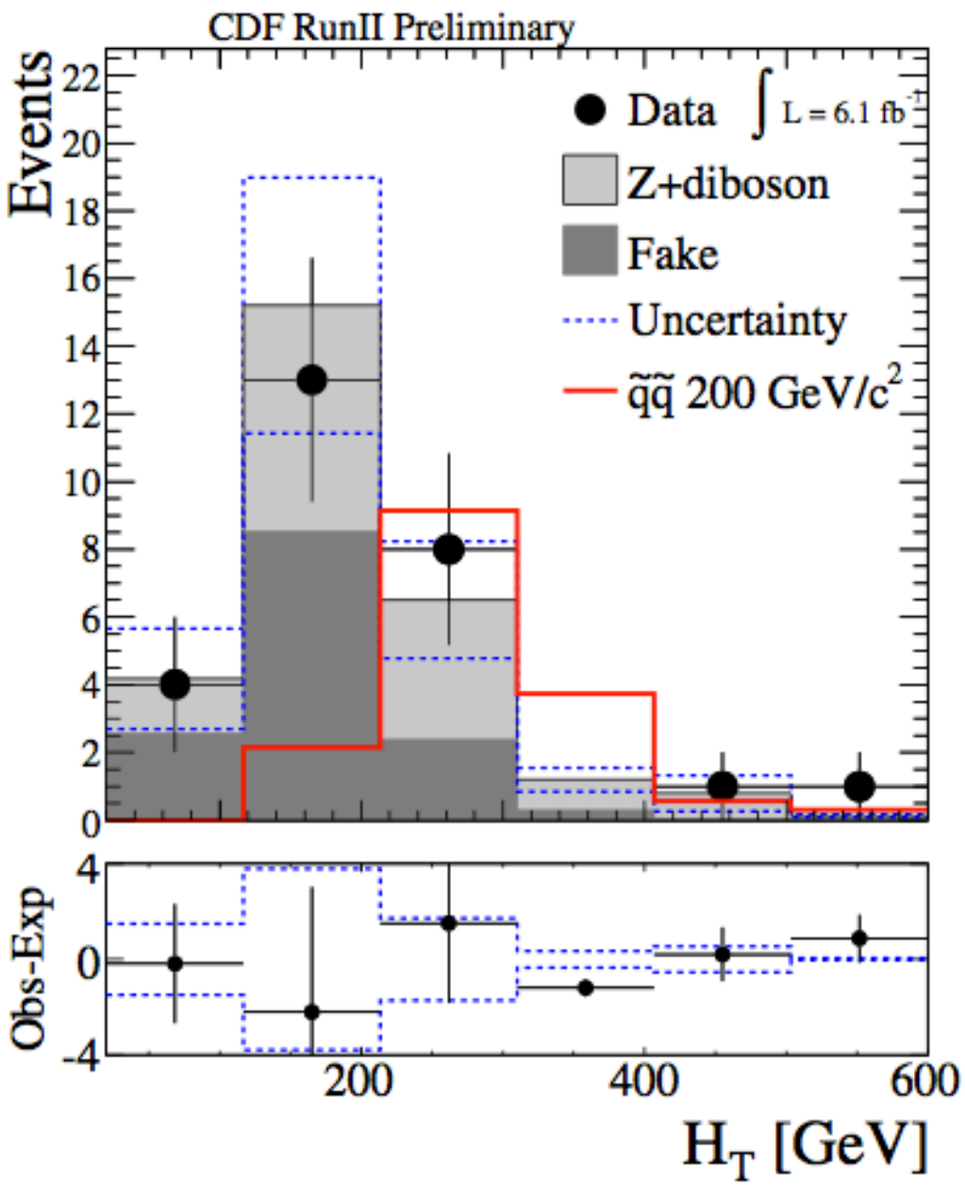
RECASTING

(like Rivet with folding instead of unfolding)







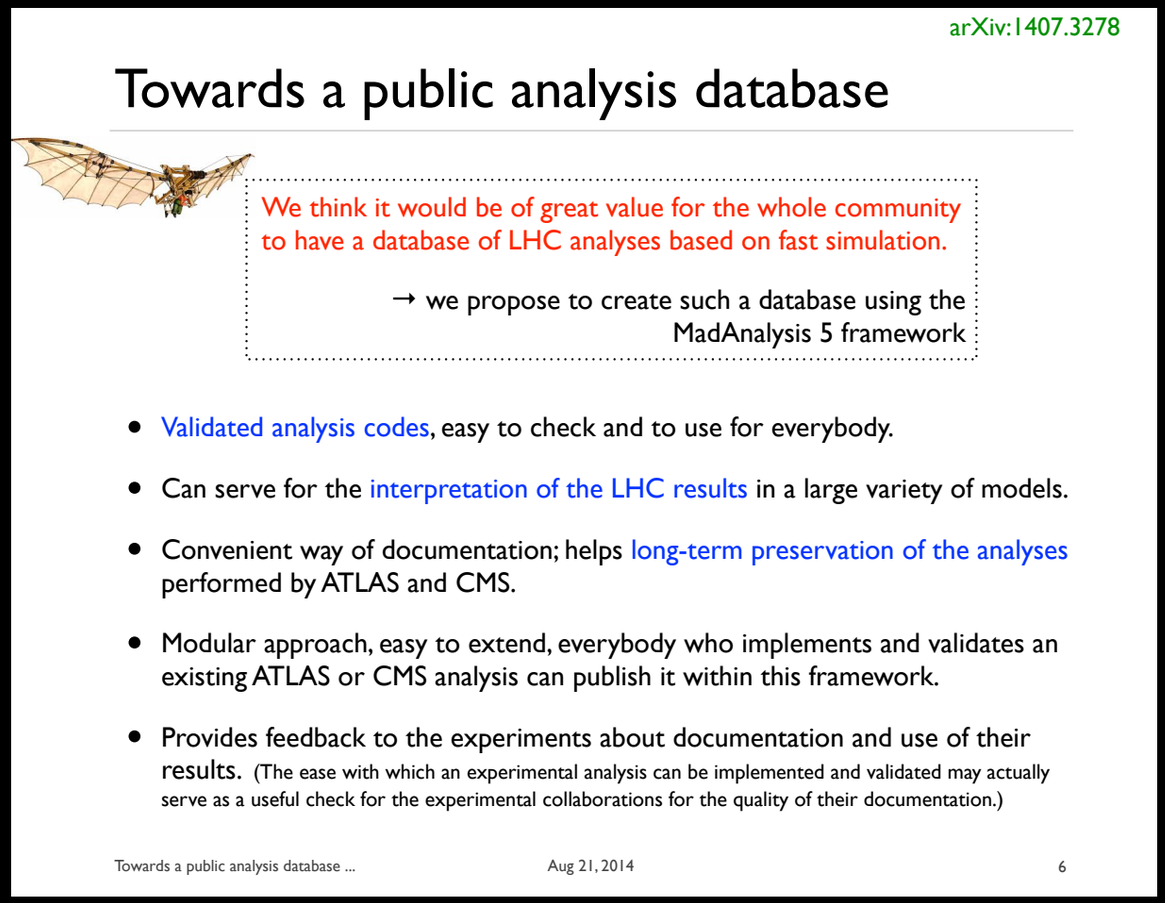


PHENO RECASTING SOFTWARE

- Several tools being developed by phenomenologists to address the need for an organized approach to recasting (but using unofficial and/or approximate methods.

  - ATOM
  - FastLim
  - MadAnalysis
  - SModelS
  - XQCAT
  - CheckMate
  - unofficial contributions to Rivet

Sabine Kraml

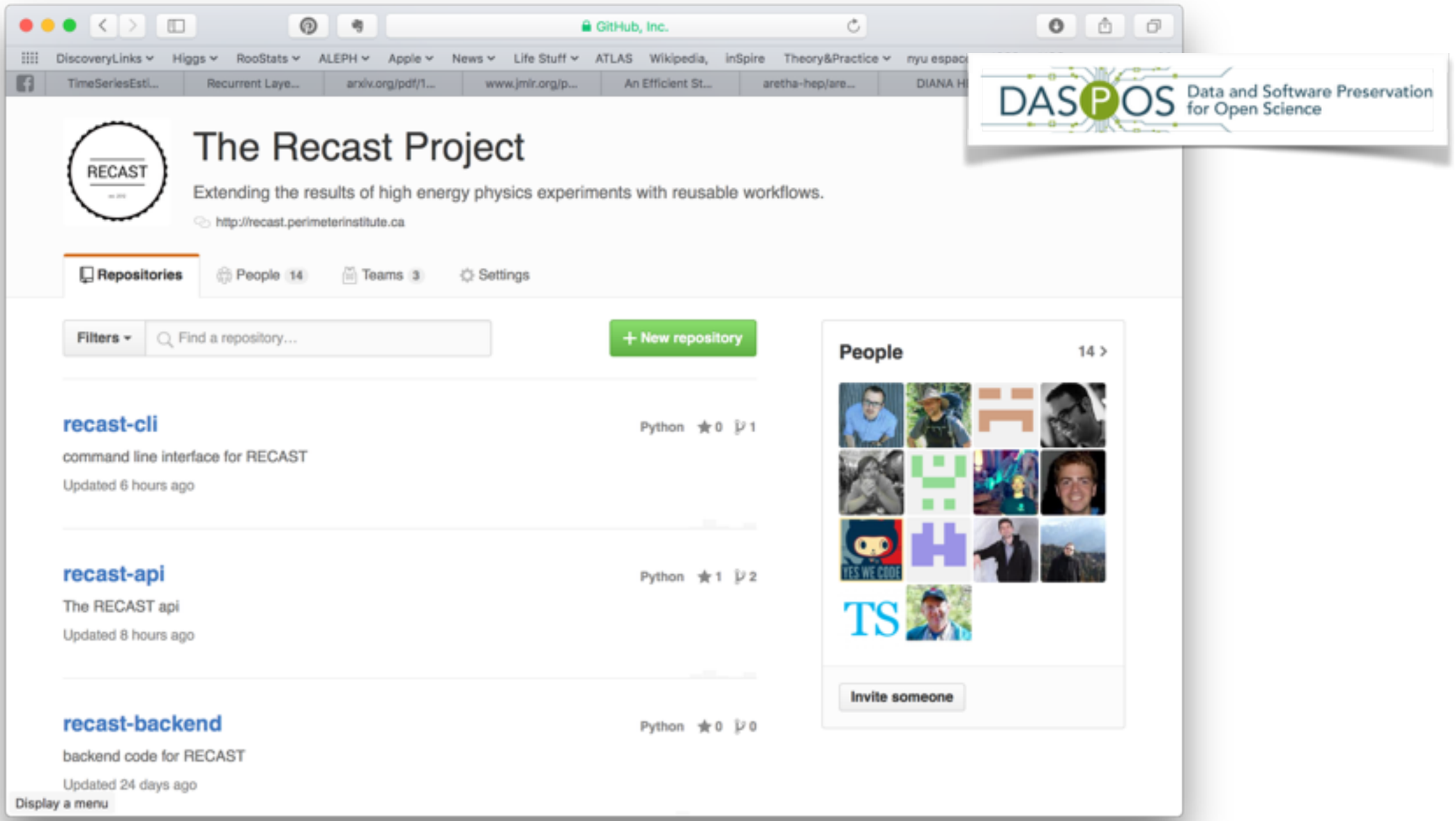


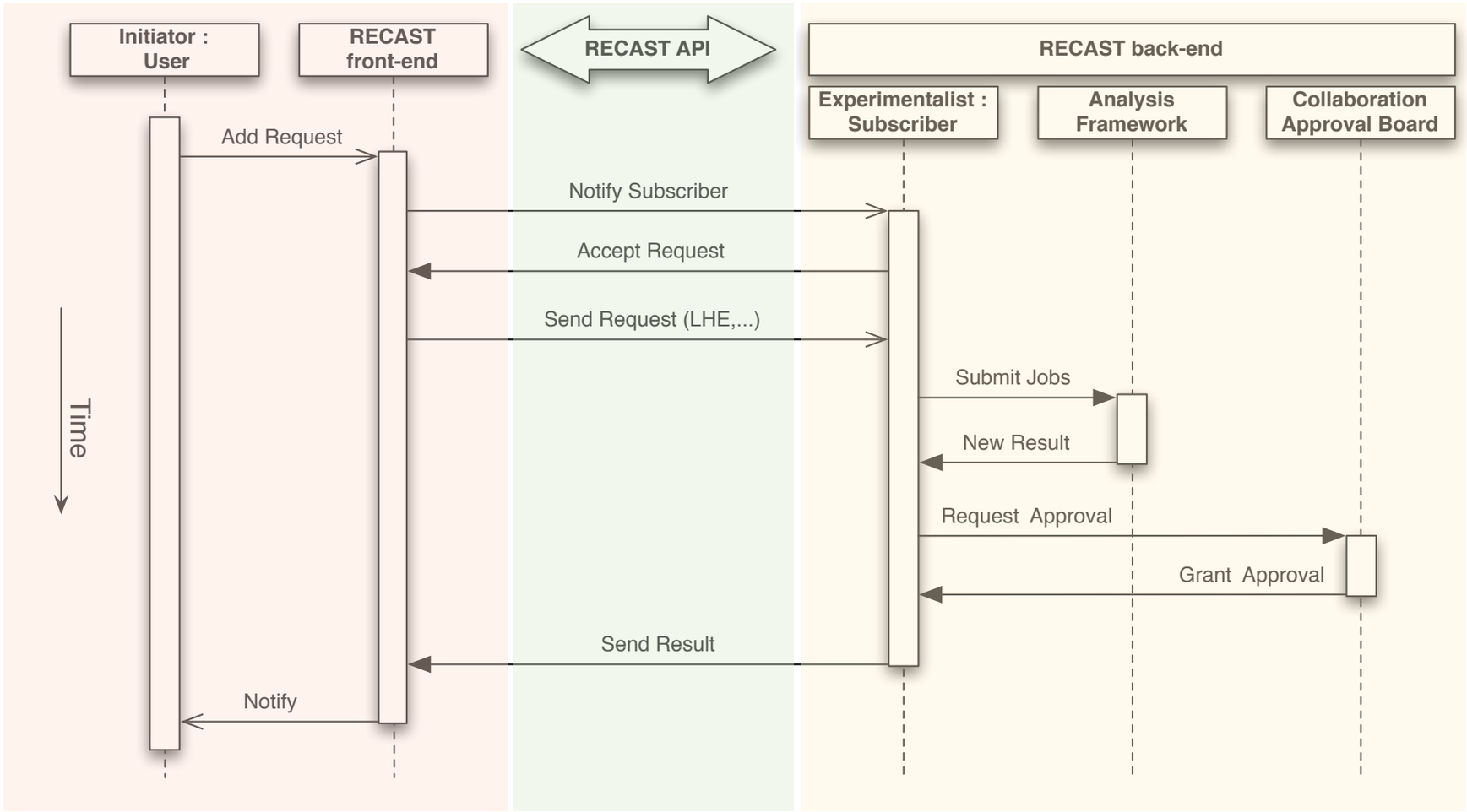
- Also possible to interface RECAST front-end with these unofficial tools

RECAST

Many people contributing now. Using DASPOS's JSON schema developed by Lukas Heinrich for packaging realistic LHC analyses

We have some example analyses stored in CERN Analysis Portal (CAP) in this format that we can run

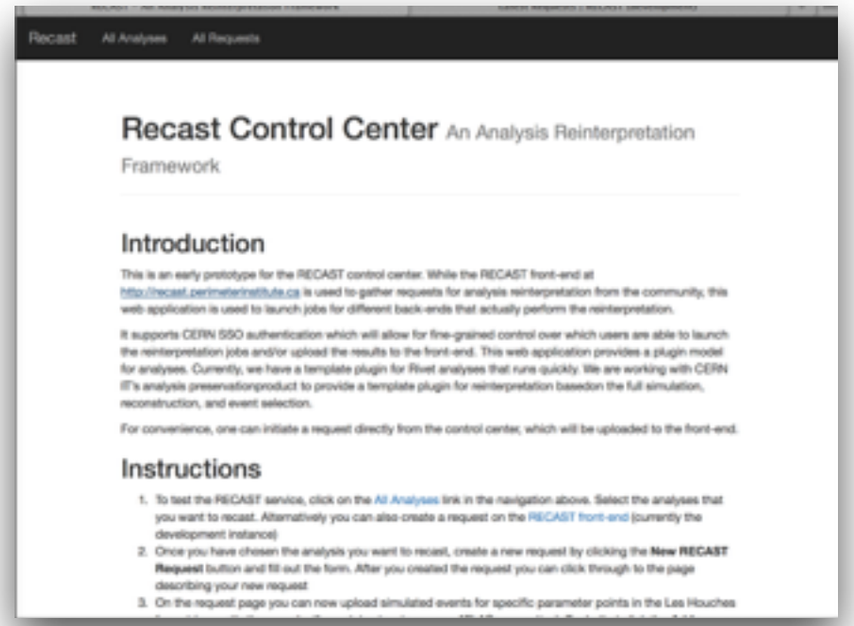




recast.perimeterinstitute.ca

Front-end prototype designed by K.C. and Itay Yavin, live since 2012.

recast-demo.cern.ch



New! Great work by Lukas Heinrich (NYU), contributions from Ken Bloom via DASPOS and Frank & Tibor of CERN-IT !

front-end

control center

Home » Analyses Catalog » Demo with working rivet-based back-end » List Requests » test for UCI

View | Edit | Edit | Contact Requester | Show Results | Devel

1. request initiated

Analysis: Demo with working rivet-based back-end

Status: Completed

Requester: lheinric

Recast Audience: all

Model Name: CMSSM

Selected Subscriber(s): lheinric, cranmer

Mon, 02/02/2015 - 14:26 - Activated
Wed, 02/04/2015 - 03:06 - Completed

Request Description and Potential

Reason for request:
because we can
Additional Information:
No information available

5. response public

Recast Request test for UCI

Request Details

analysis	Demo with working rivet-based back-end
status	1
model-type	None
uuid	4cdc558c-8f4a-eab4-fdbf-cd91a34db4b2
new-model-information	None
title	test for UCI
predefined-model	CMSSM
reason-for-request	because we can
requestor	lheinric
audience	None
subscribers	lheinric
additional-information	None

4. upload response

2. process request

+Add Parameter Point | ⓘUpload to RECAST

Parameter	Desciption	Number of Events	Cross-Section	
parameter-0	test for UCI	1000	20	process results

3. review results

Home | Analyses Catalog | Requests | My Subscriptions | About | Developers | News | Help

Home » Analyses Catalog » Demo with working rivet-based back-end » List Requests » test-upload-2 » Show Results » Recast Response for Request #test-upload-2

Recast Response for Request #test-upload-2

View | Edit | Devel

Submitted by lheinric on Sun, 01/18/2015 - 09:54

Request: test-upload-2
ROOT file with TH1: 20150118095414b5872ab0-1a2b-1044-c154-5cead413bc8f.zip
Status: Completed

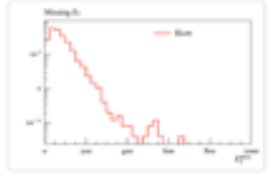
Results for request 4cdc558c-8f4a-eab4-fdbf-cd91a34db4b2 - parameter-0

Efficiency

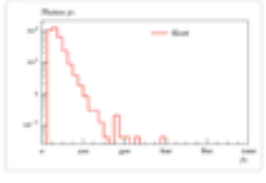
0.18272727272727274

Plots

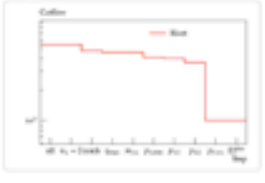
> MET:



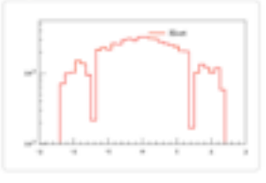
> PhotonPt:



> Cutflow:

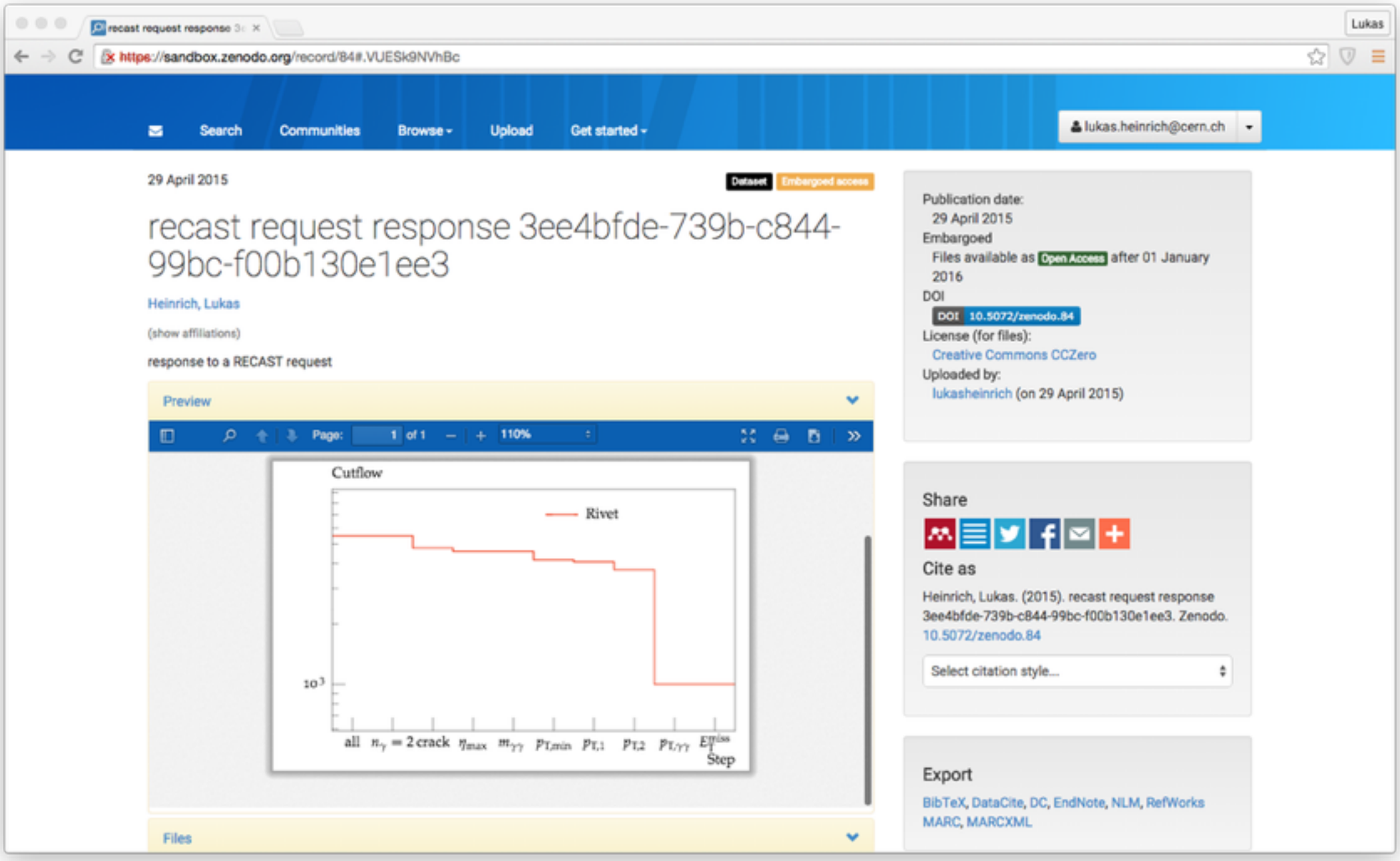


> PhotonEta:



EXAMPLE RECAST → HEPDATA / ZENODO

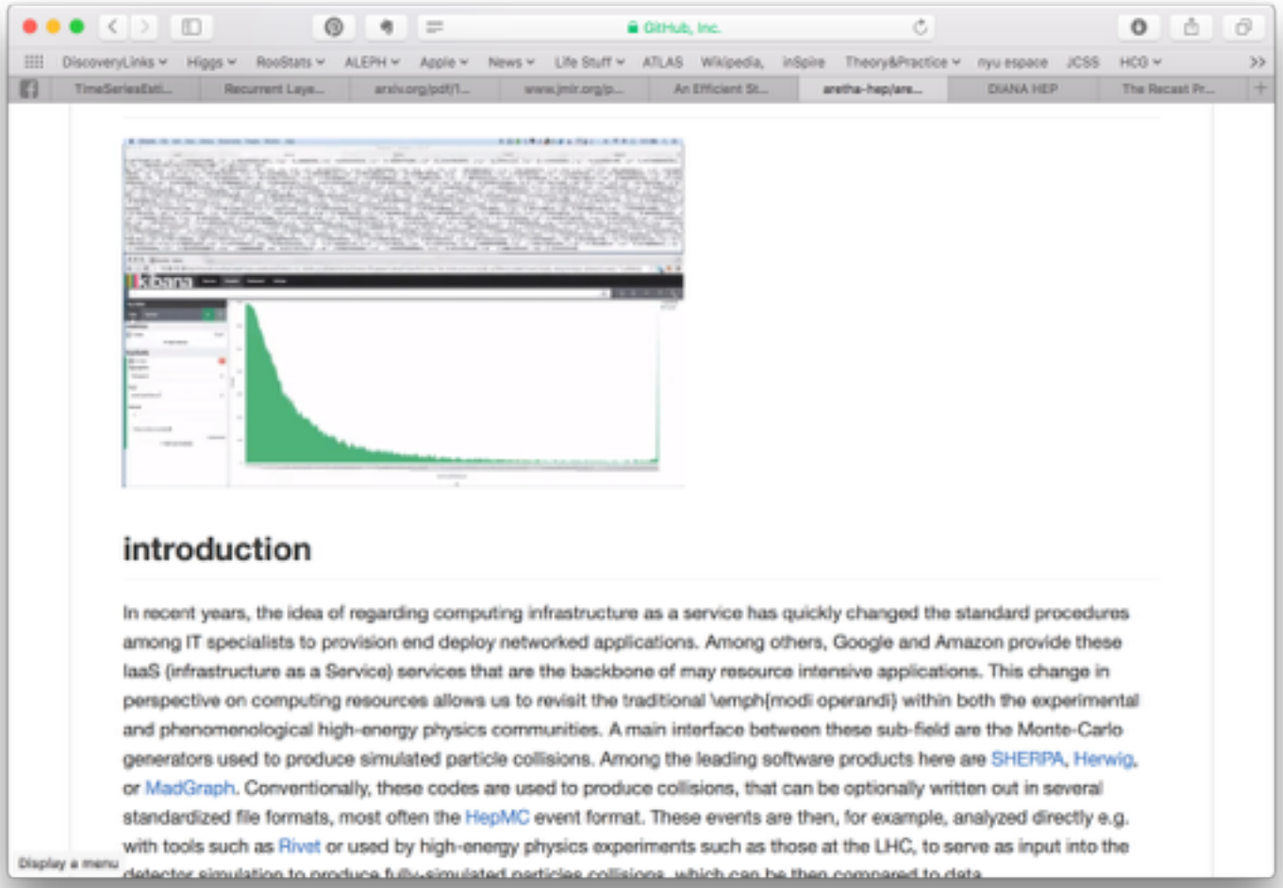
After re-running analysis on new physics model, experiments might want to push result of new interpretation to HEPData. Technically we can do this with Zenodo. Would be nice for HEPData to have API connection to upload result.



MONTE CARLO AS A SERVICE

Lukas has prototyped a web service called Aretha that encapsulates Monte Carlo tools and wraps them as a web service.

- Specific version of "cards" configuring Monte Carlo generator
- specific installation (stored in a docker container) that ensures version of generator and other dependencies (compiler etc.)



https://github.com/aretha-hep/aretha-doc

- give DOIs to the cards and container
- generate more consistent MC on demand

MODEL CATALOGUE

One theoretical model might be referred to in many of these tools.
Would be nice to have a unique identifier for the model

HEPData

INSPIRE

CAP

RIVET

RECAST

CDS

Wiki pages

ROOT

GitHub/GitLab

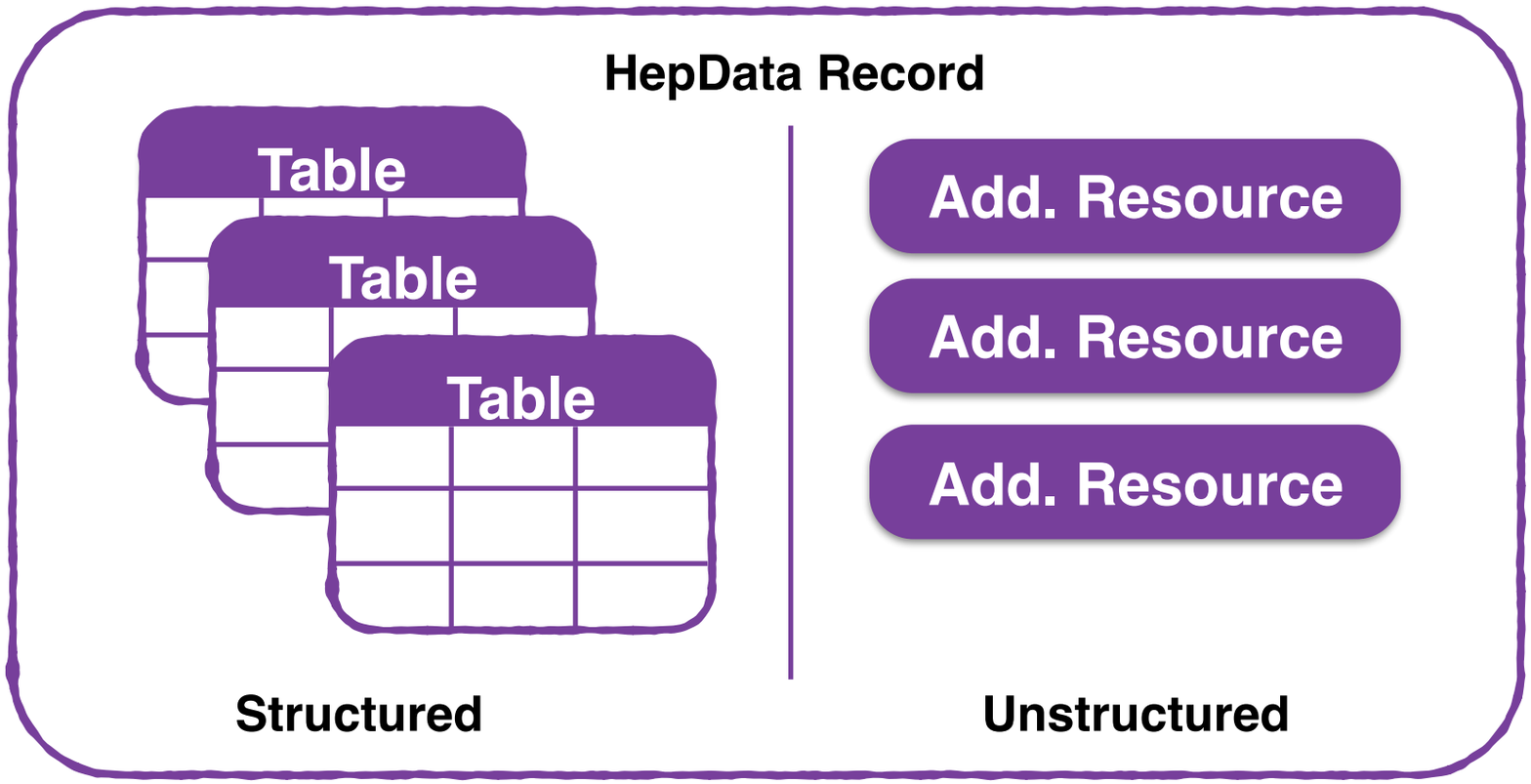
Mathematica

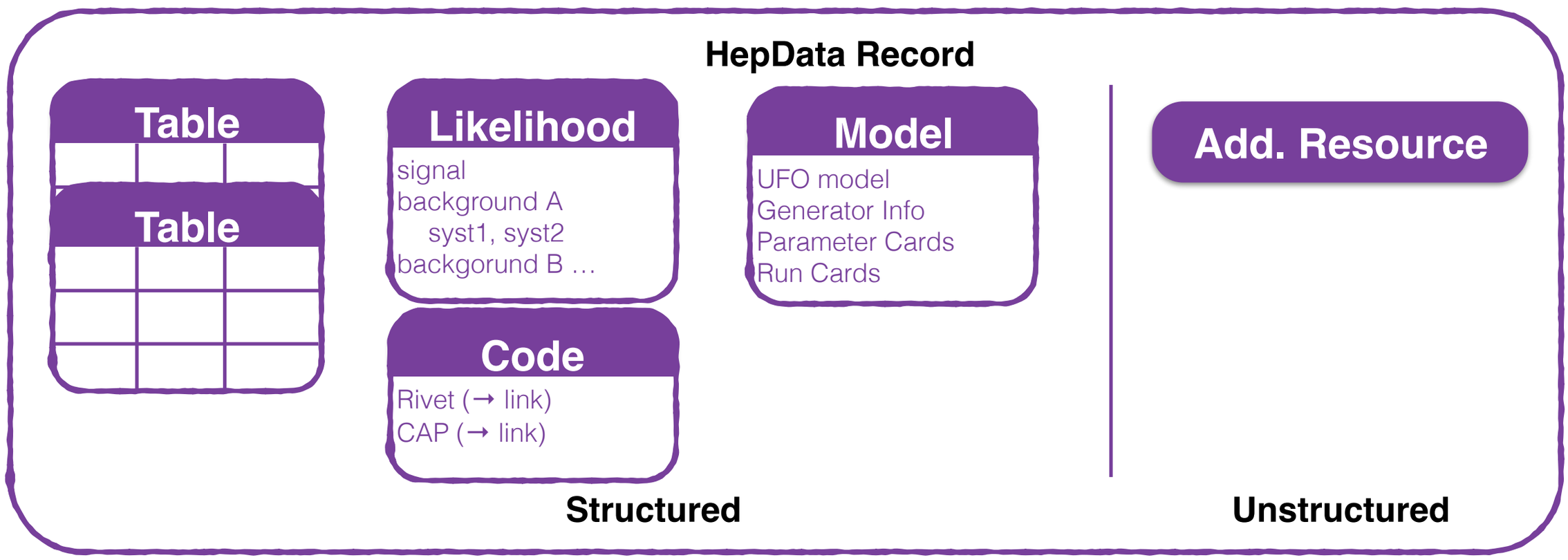
Zenodo

Jupyter/IPython

MadAnalysis
Checkmate
Atom, FastLim, …

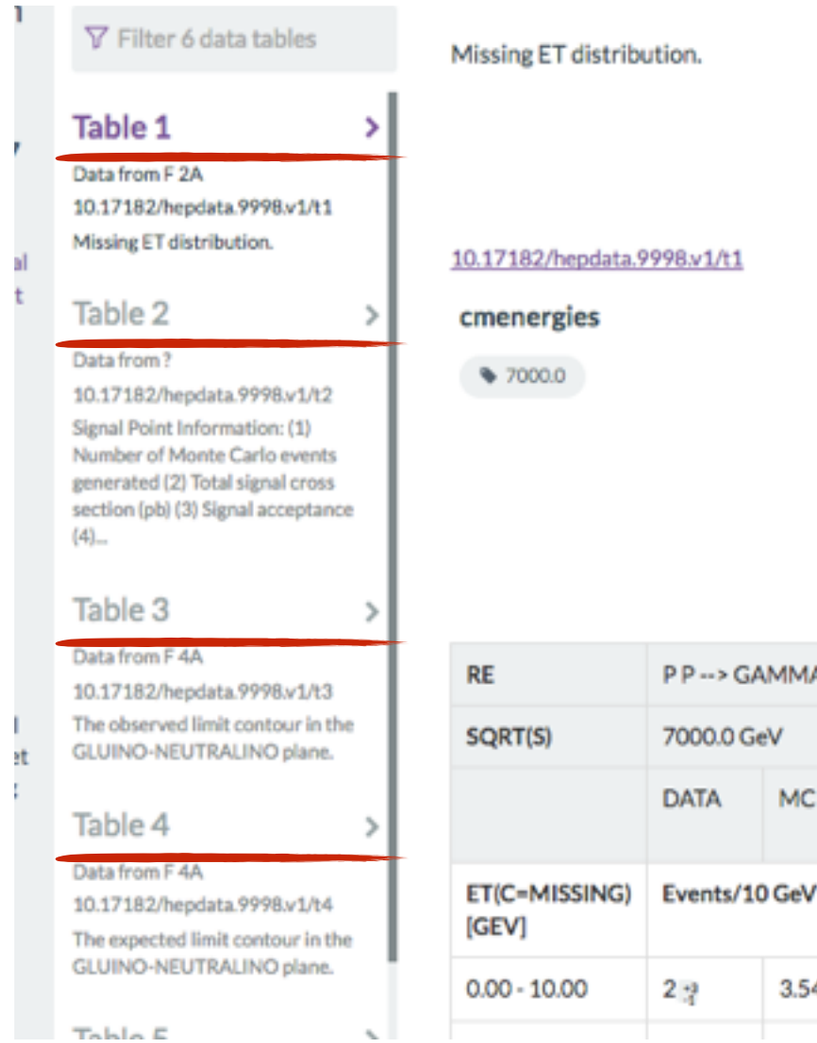
MadGraph
Sherpa
Pythia
…


HepData Record
Table
Table
Table
Add. Resource
Add. Resource
Add. Resource
Structured
Unstructured


HepData Record
Table
Table
Likelihood
signal
background A
syst1, syst2
backgorund B …
Code
Rivet (→ link)
CAP (→ link)
Model
UFO model
Generator Info
Parameter Cards
Run Cards
Add. Resource
Structured
Unstructured

Currently, HepData has only one native data format: **the table**:

- observable measurements as multivariate functions
- simple data model, that works well for many use-cases
- but very limited options to store semantic information in a structured way
  - what kind of data is this? a measured distribution, a limit contour? an efficiency table?
  - mostly need to rely on free-format text table description, hard to read reliably programatically
  - limits usefulness of e.g. HepData Explore



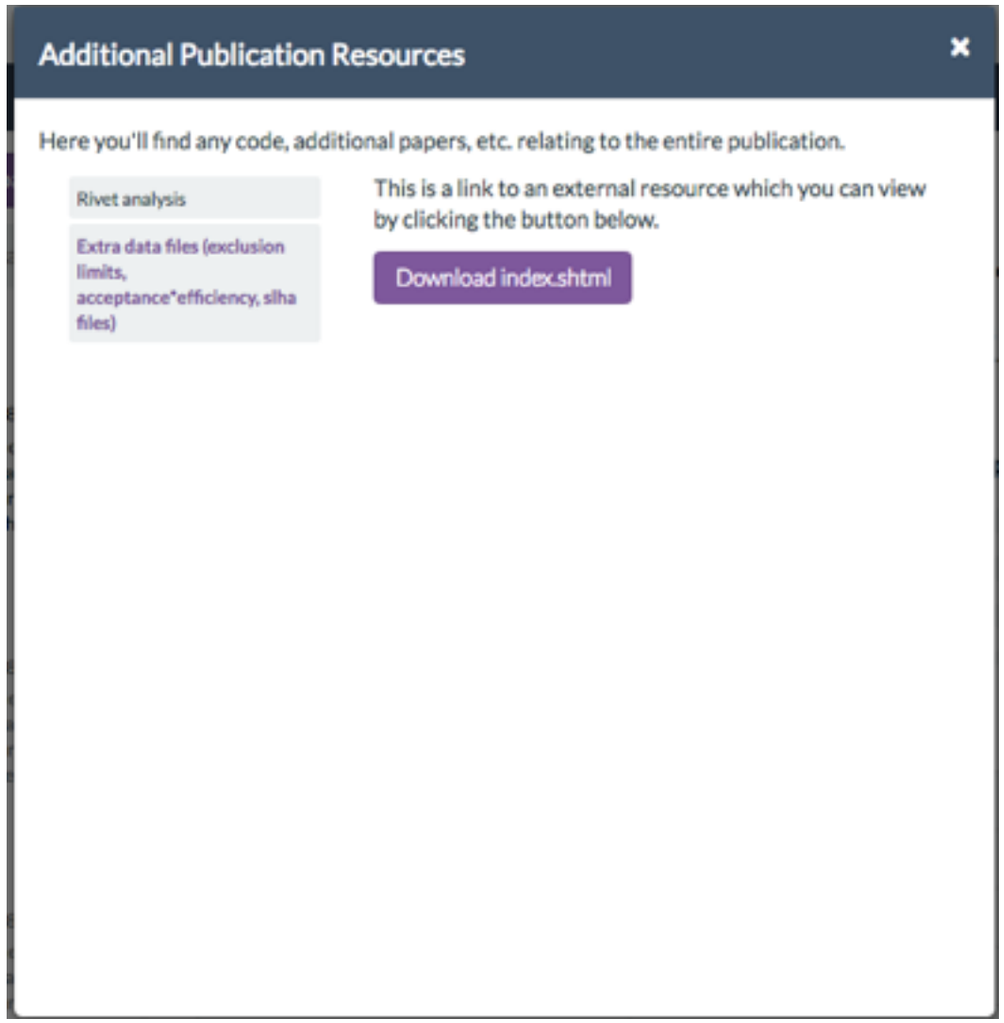
Example:

- Natural HepData query: "Show me all records from ATLAS and CMS that have limit on Gluino mass". Current obstacles:
  - record author (e.g. collaboration) not natively known to HepData (needs query to Inspire)
  - depend on consistency in variable names ("CLs" vs "CL_s" vs "CL", "M_gluino", "M(GLUINO)",…)

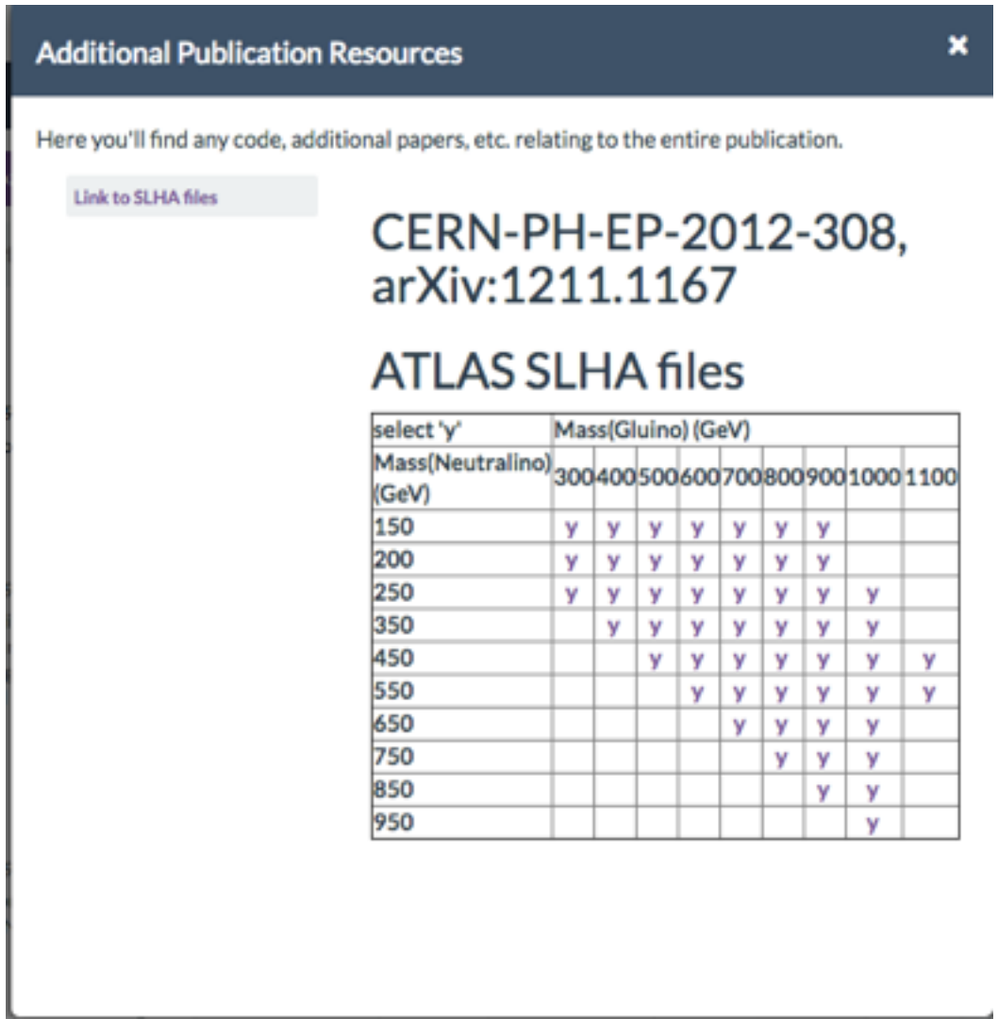
Records already lots additional information but mostly unstructured if it can't be put into table format

- SLHA files as collection of links
- HistFactory likelihoods as uploaded XML/ROOT files
- Rivet Analyses as links

With large # of records in LHC era, machine-readability is key, needs structure of stored data beyond the tables



VS



select 'y'	Mass(Gluino) (GeV)								
Mass(Neutralino) (GeV)	300	400	500	600	700	800	900	1000	1100
150	y	y	y	y	y	y	y		
200	y	y	y	y	y	y	y		
250	y	y	y	y	y	y	y	y	
350		y	y	y	y	y	y	y	
450			y	y	y	y	y	y	y
550				y	y	y	y	y	y
650					y	y	y	y	
750						y	y	y	
850							y	y	
950								y	

Possible new data types, beyond the table

- smarter **Table**
  - add metadata for table, indep vars, dep vars
    - already have units, and error label,
  - useful additions:
    - data type (per column) : -> data, simulated events (MC), unfolded data
    - table type: measured distribution, limit contour, covariance matrix, efficiency table
- **Model information** (generalizes SLHA resources)
  - upload UFO model
  - parameter cards for probed parameter points, run cards, etc.
- **Analysis Code References**:
  - Link to Rivet Analyses
  - Link to CERN Analysis Preservation Portal (contains code via RECAST)
  - other possibilities, CheckMate / ATOM / etc.
- **Native Likelihood information**:
  - expand existing work on HistFactory integration
    - HF XML schema easily translatable to HepData-native YAML

Technically easy to integrate iteratively as new data types emerge using JSON schemas, which HD already uses for tables.