



HEPDATA AND IT'S ROLE IN HEP'S

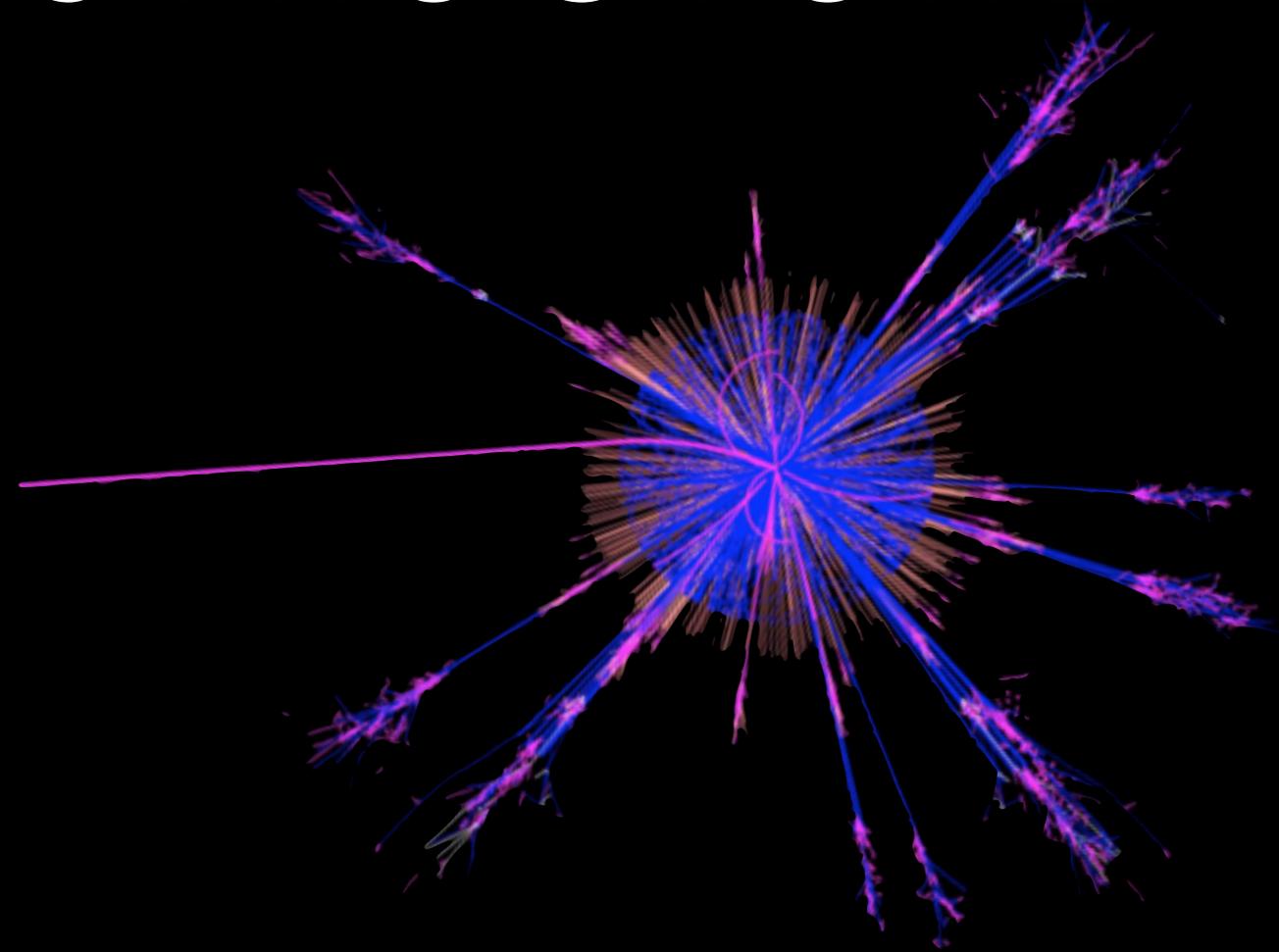
# CYBER INFRASTRUCTURE

**@KyleCranmer**

New York University

Department of Physics

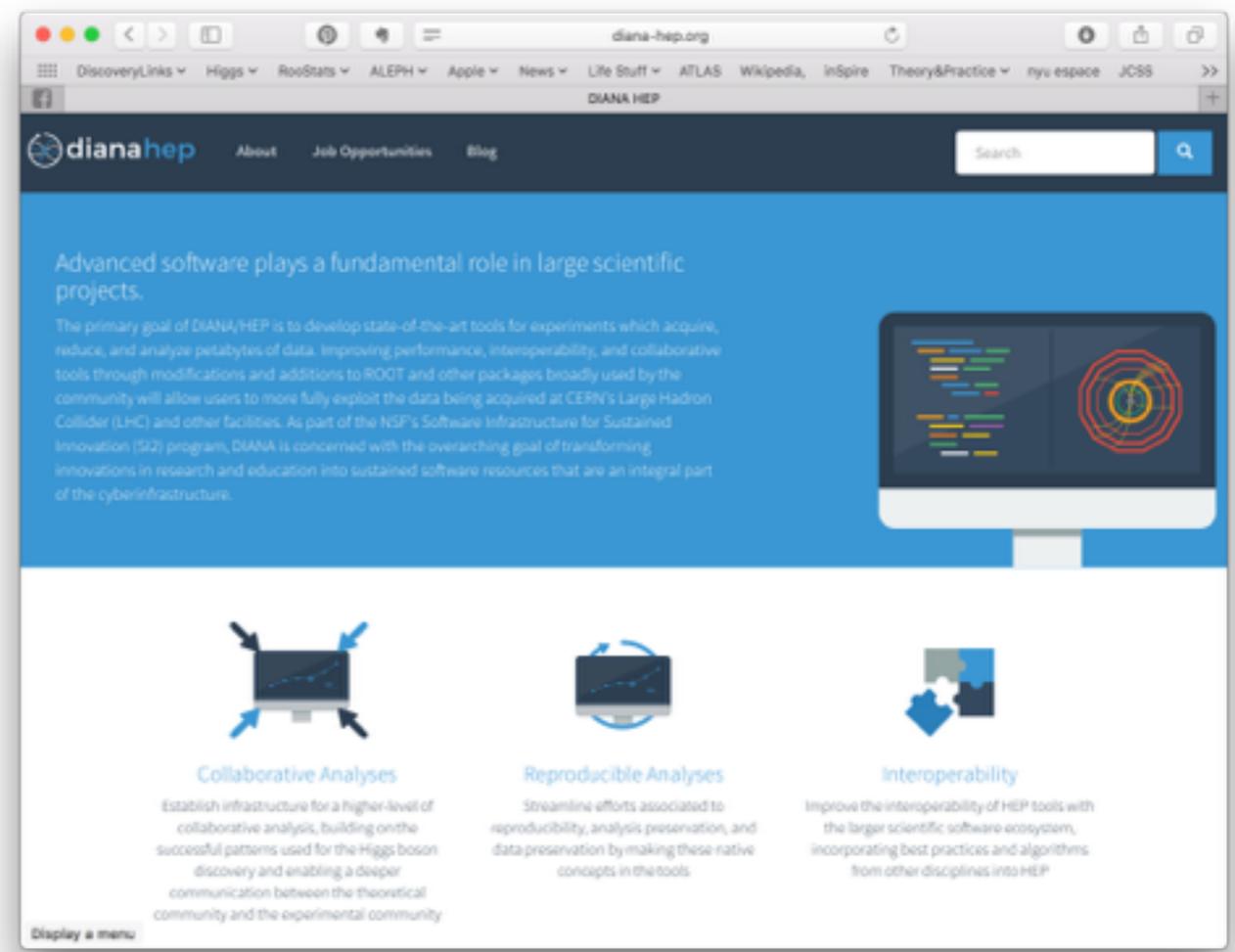
Center for Data Science



# 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.

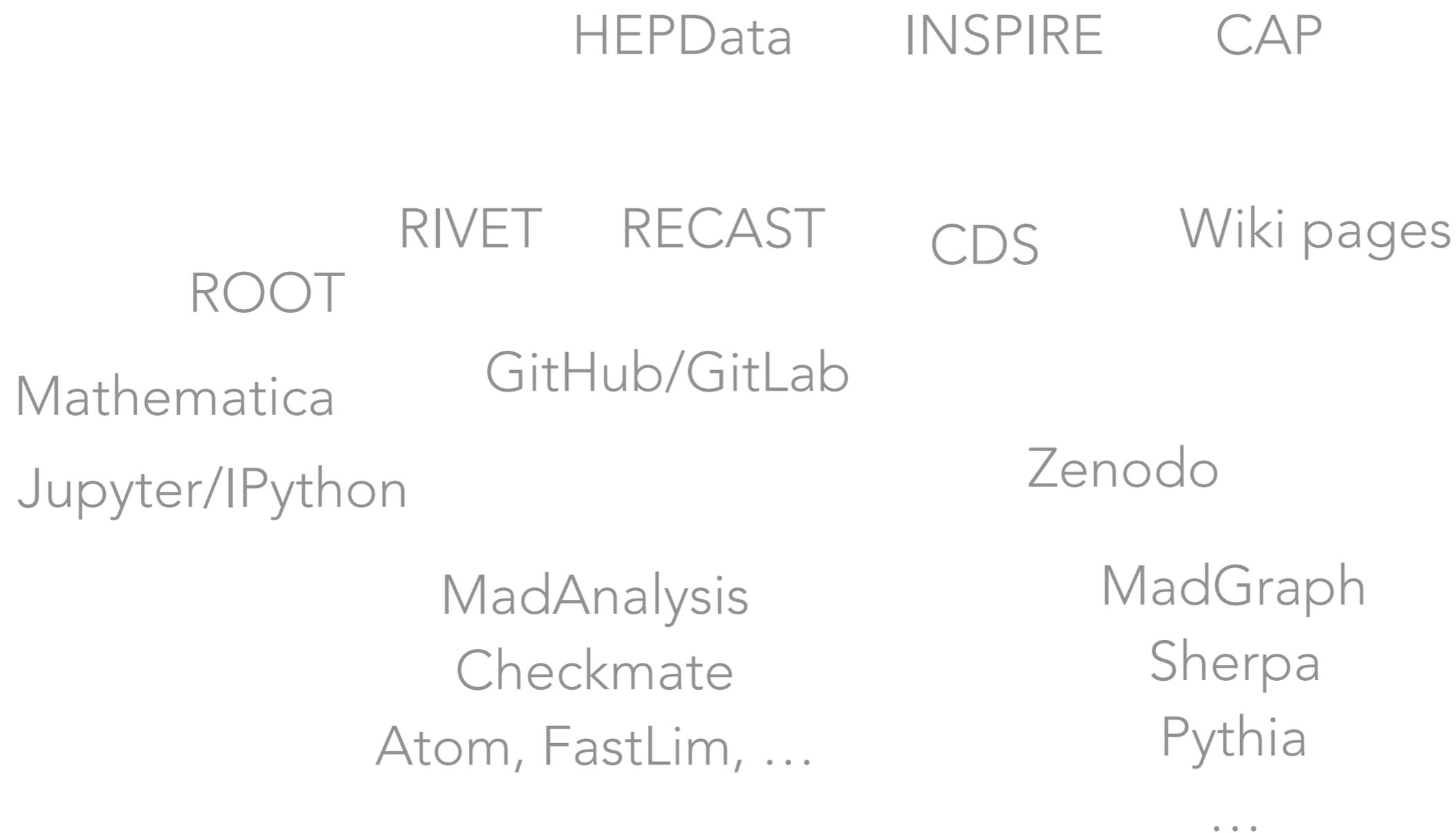
Working closely with CERN Analysis Preservation (CAP) portal, INSPIRE, and HEPData



# ECOSYSTEM OF TOOLS

Goal is to have integrations between these tools.

What do we want to do? What is missing?



# 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

Search for new phenomena in final states with large jet multiplicities and missing transverse momentum with ATLAS using  $\sqrt{s} = 13$  TeV proton-proton collisions

Table 1

$E_T^{\text{miss}}/\sqrt{|H_T|}$  distribution in validation region 7e[50]0b. Two benchmark signal models are overlaid on the plot for comparison. Labelled "pMSSM" and "2-step", they show signal distributions from the example SUSY models (as described in the paper): a pMSSM slice model with  $(m_{\tilde{g}}, m_{\tilde{u}_L}) = (1300, 200)$  GeV and a cascade decay model with  $(m_{\tilde{g}}, m_{\tilde{u}_L}) = (1300, 200)$  GeV.

cmenergies

observables

SQRTR	DATA	BACKGROUND	pMSSM-1300-200	2Step-1300-200
0.00-0.25	14752	97302.51	0.002	0
0.25-0.50	240304	263827.42	0.031	0.748

Observed 95% CL limit for the pMSSM grid when the signal cross section is increased by one standard deviation.

Table 8

SQRTR	13000.0-GeV
$m(j)$ [GeV]	$m(j)$ [GeV]
1441.3	177.9
1448.8	193.6
1451.2	197.7
1459.4	209.4
1465.3	225.1

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

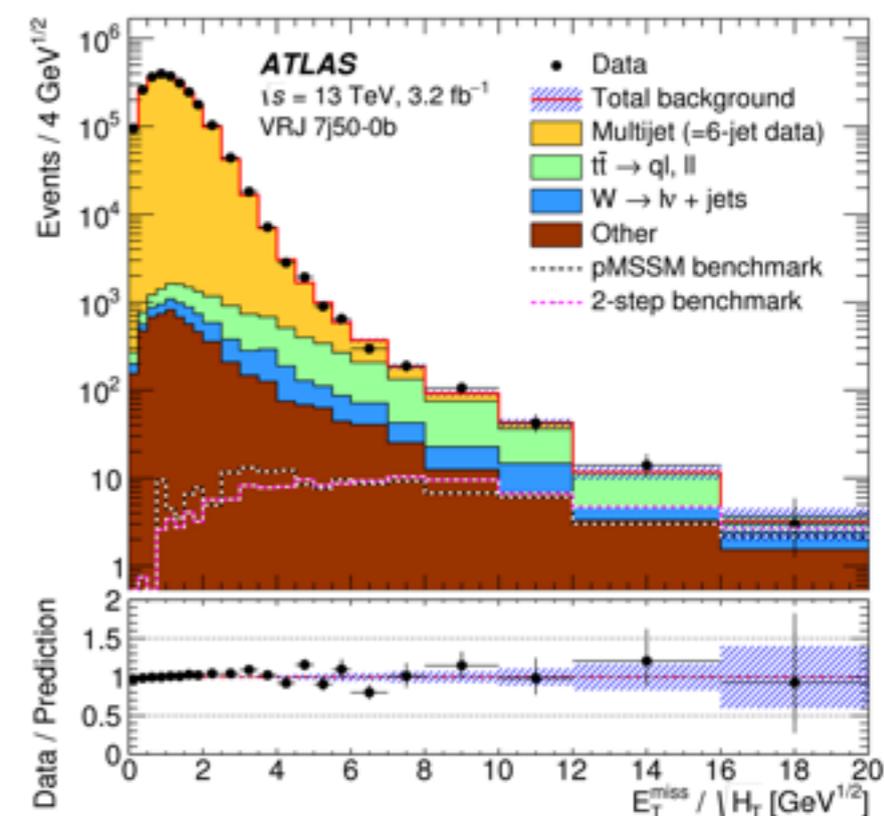
Use case: Find all 95% contours for a given simplified model

# IMPROVING REPRESENTATION OF PLOTS

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

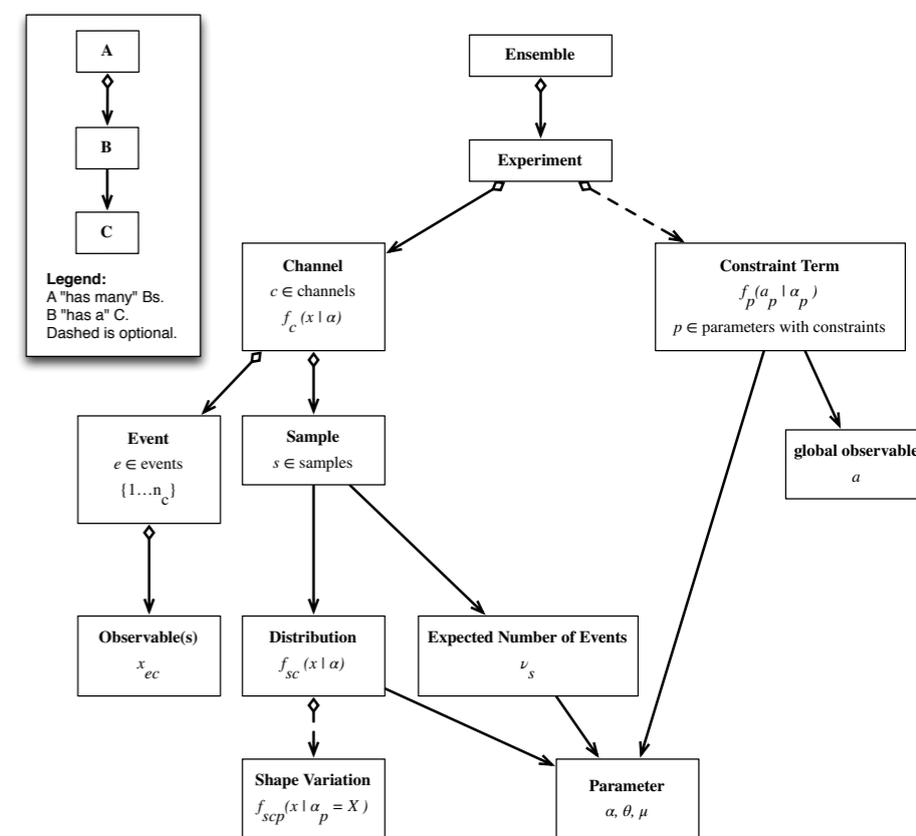
But typical plot has many background components, complicated correlated uncertainties, variational histograms, etc.. Does not fit HEPData tables

- 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.





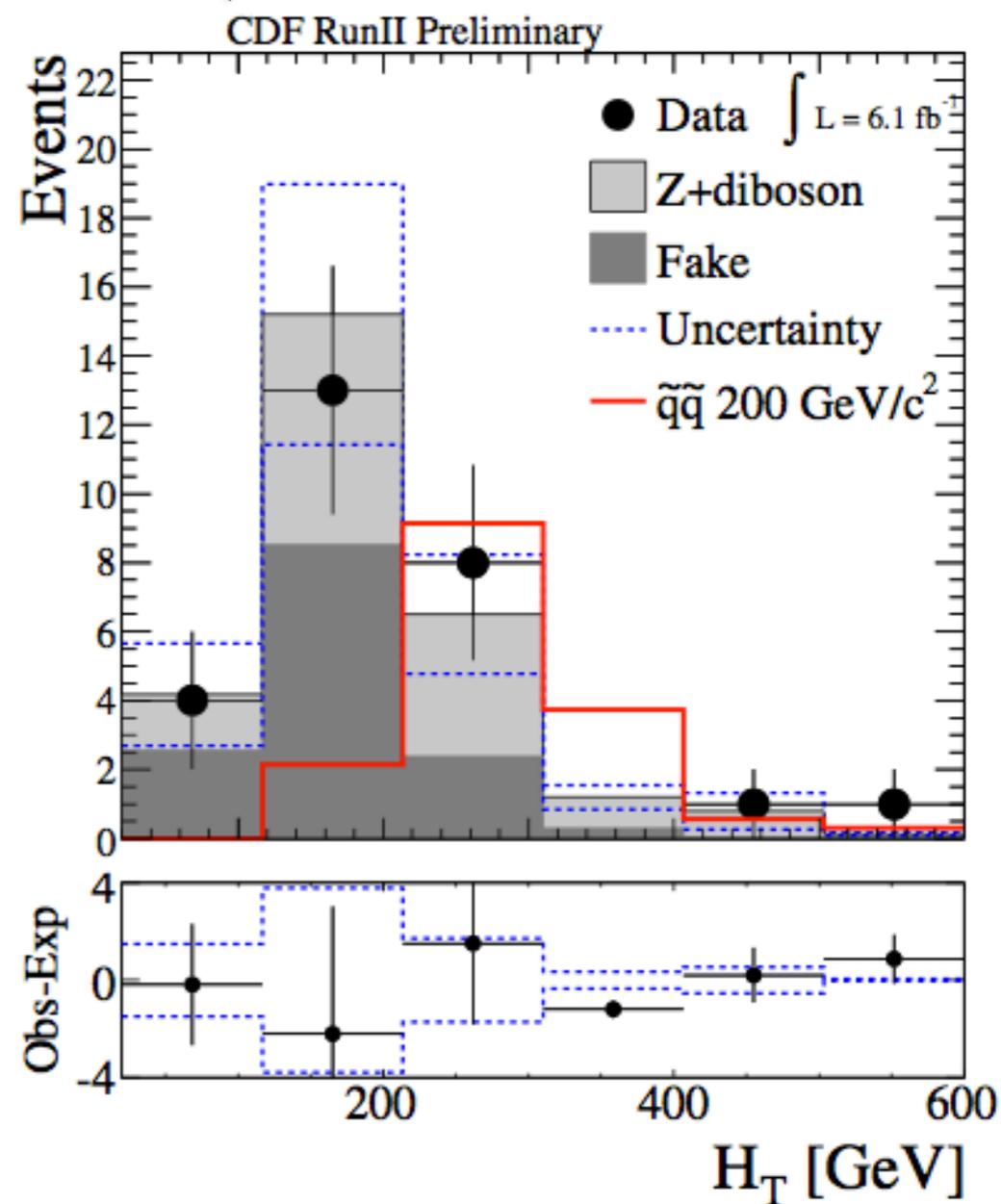
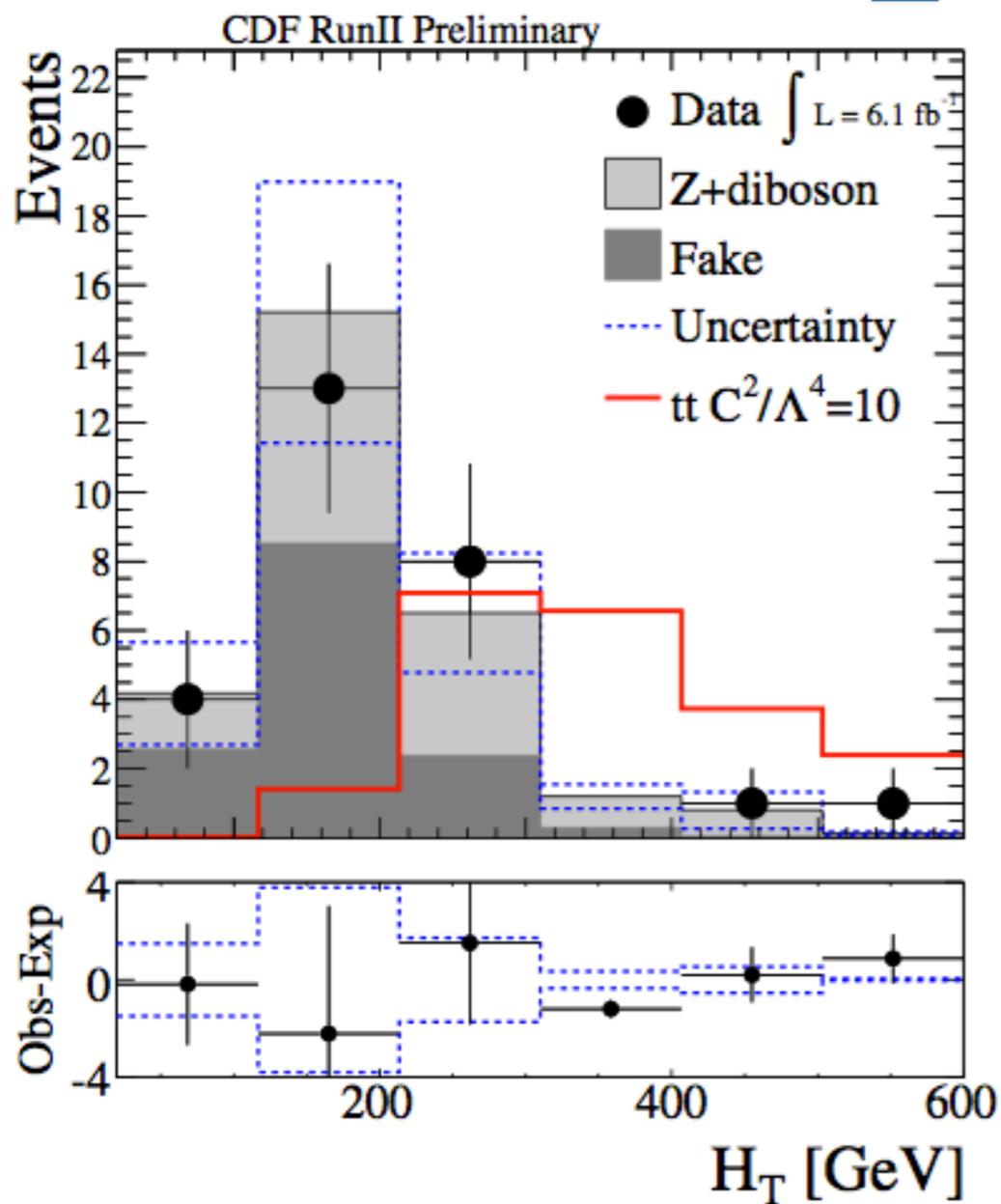
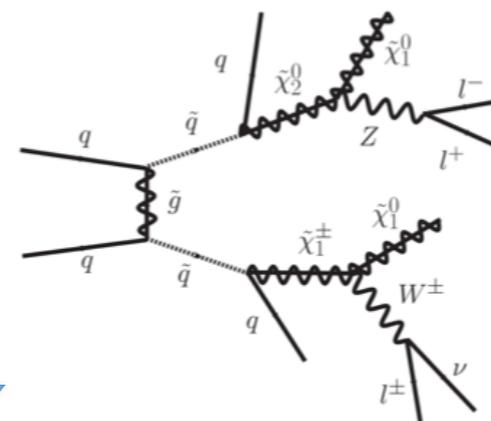
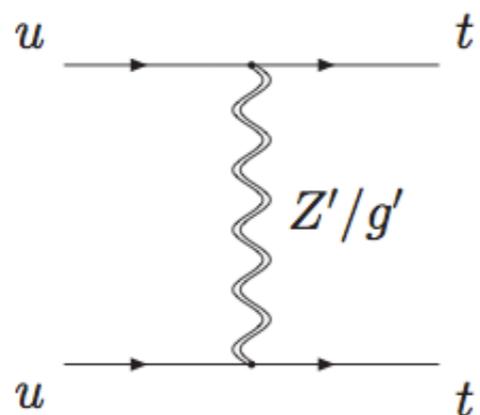
## Level-1. Published results

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)



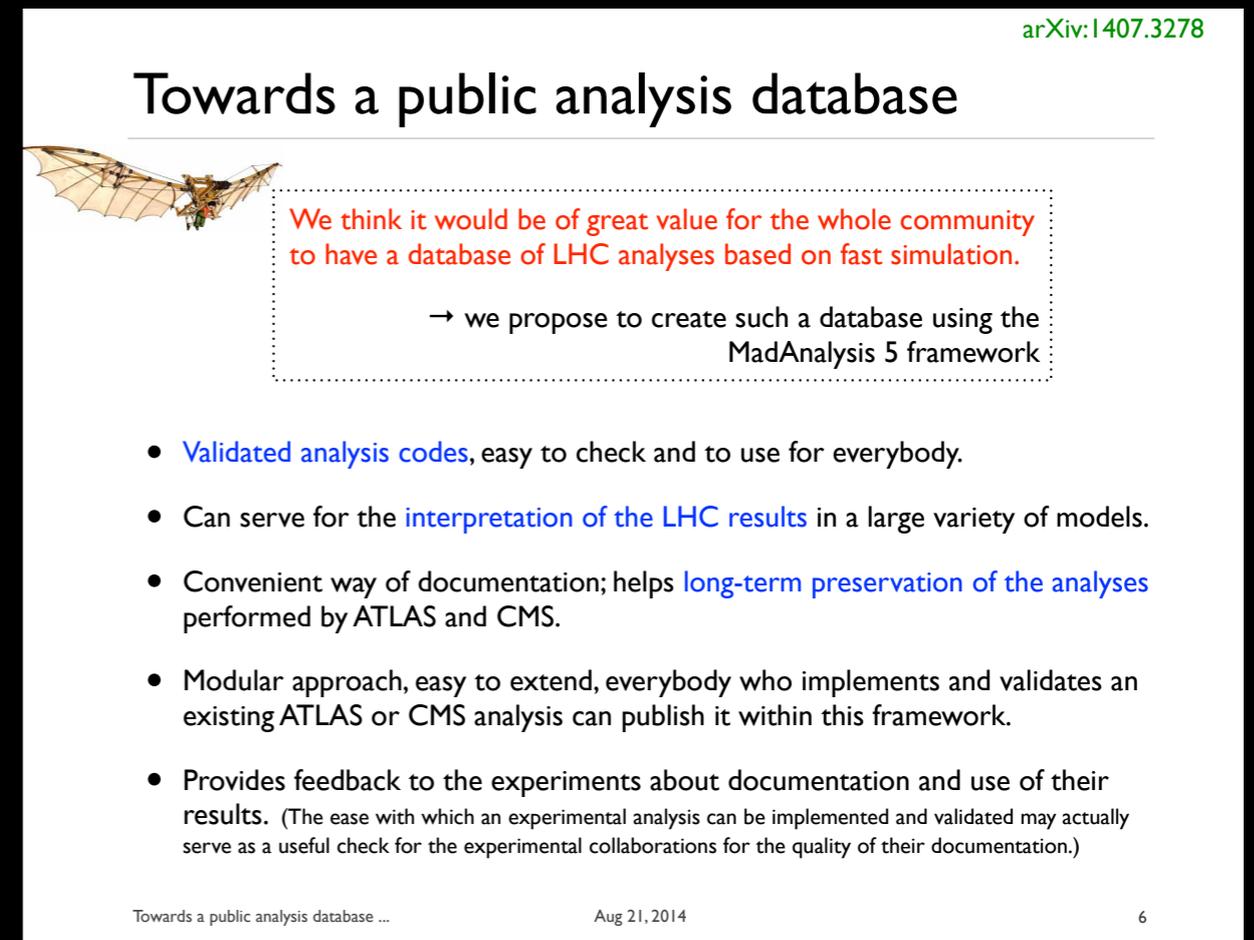
# PHENO RECASTING SOFTWARE

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

Sabine Kraml

arXiv:1407.3278

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



## Towards a public analysis database

We think it would be of great value for the whole community to have a database of LHC analyses based on fast simulation.

→ we propose to create such a database using the MadAnalysis 5 framework

- Validated analysis codes, easy to check and to use for everybody.
- Can serve for the interpretation of the LHC results in a large variety of models.
- Convenient way of documentation; helps long-term preservation of the analyses performed by ATLAS and CMS.
- Modular approach, easy to extend, everybody who implements and validates an existing ATLAS or CMS analysis can publish it within this framework.
- Provides feedback to the experiments about documentation and use of their results. (The ease with which an experimental analysis can be implemented and validated may actually serve as a useful check for the experimental collaborations for the quality of their documentation.)

Towards a public analysis database ... Aug 21, 2014 6

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

# RECAST

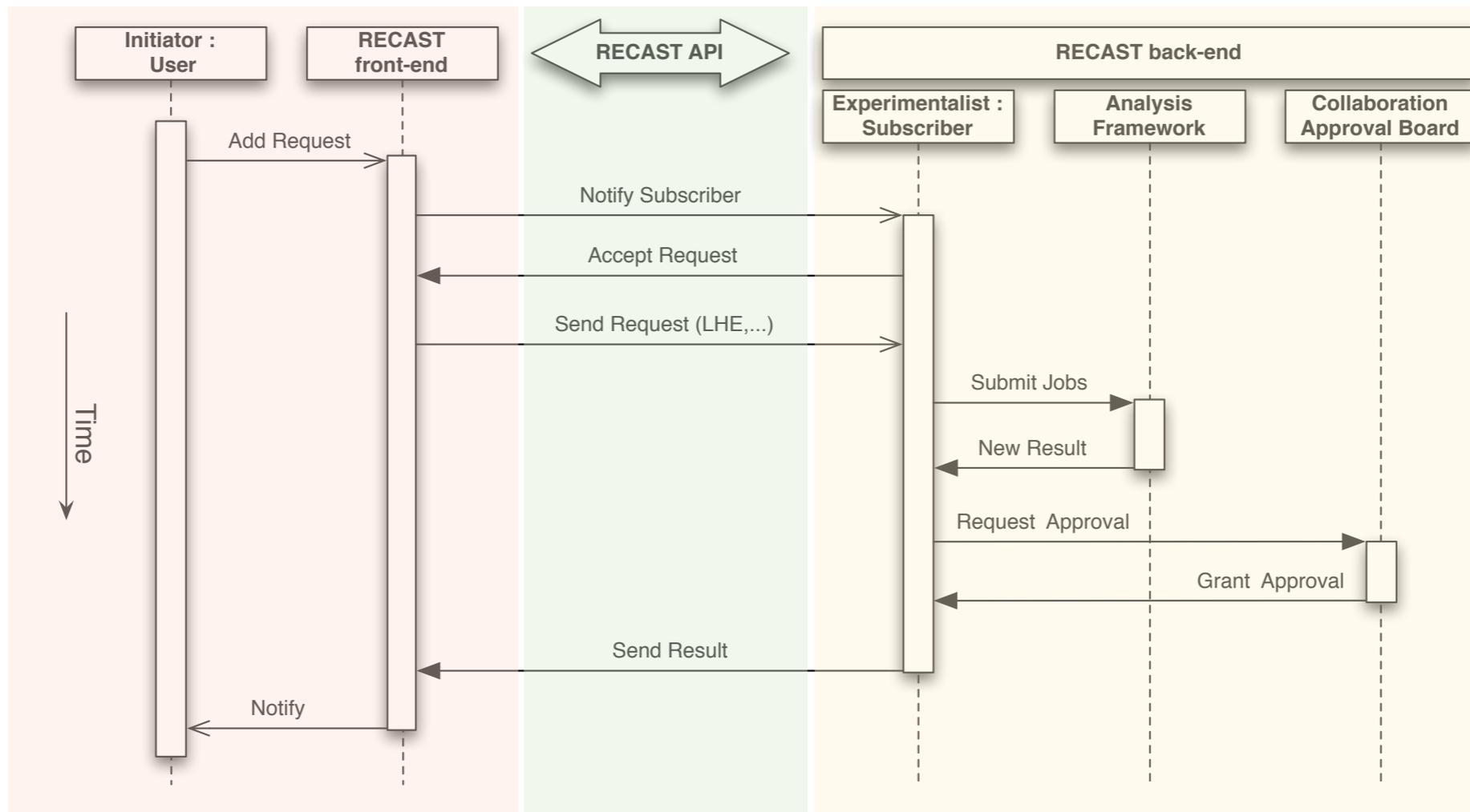
Many people contributing now. Using DASPOS's JSON schema developed by Lukas Heinrich for packaging realistic LHC analyses

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

The screenshot displays the GitHub repository page for 'The Recast Project'. The page header includes the project name and a description: 'Extending the results of high energy physics experiments with reusable workflows.' Below this, there are navigation tabs for 'Repositories', 'People 14', 'Teams 3', and 'Settings'. A search bar and a '+ New repository' button are also visible. The main content area lists three repositories:

- recast-cli**: Python, 0 stars, 1 fork. Description: 'command line interface for RECAST'. Updated 6 hours ago.
- recast-api**: Python, 1 star, 2 forks. Description: 'The RECAST api'. Updated 8 hours ago.
- recast-backend**: Python, 0 stars, 0 forks. Description: 'backend code for RECAST'. Updated 24 days ago.

On the right side, there is a 'People' section showing 14 contributors with their avatars. Below the avatars is an 'Invite someone' button. A 'DASPOS Data and Software Preservation for Open Science' logo is overlaid in the top right corner of the screenshot.

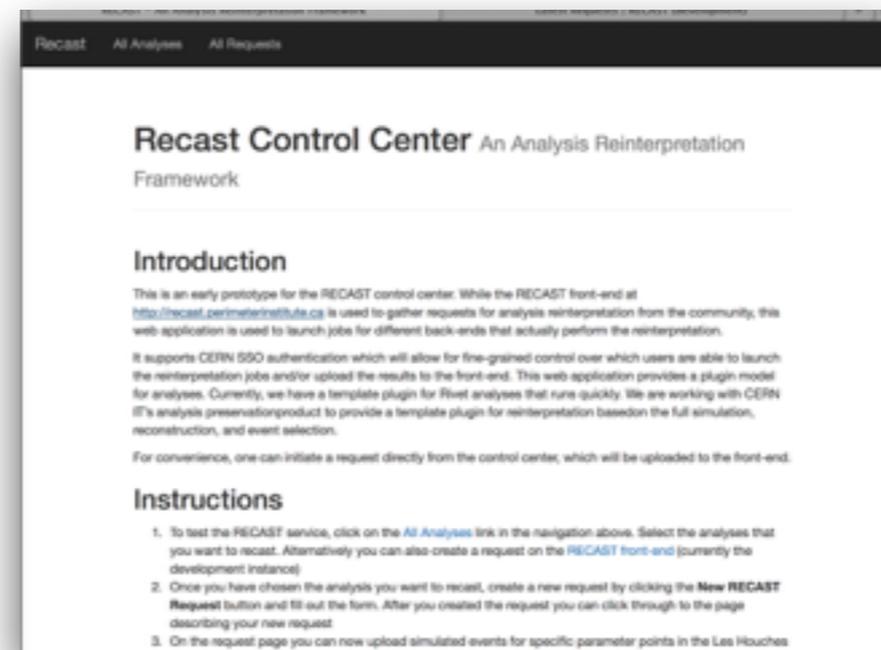


[recast.perimeterinstitute.ca](http://recast.perimeterinstitute.ca)

[recast-demo.cern.ch](http://recast-demo.cern.ch)



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



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**

+Add Parameter Point Upload to RECAST

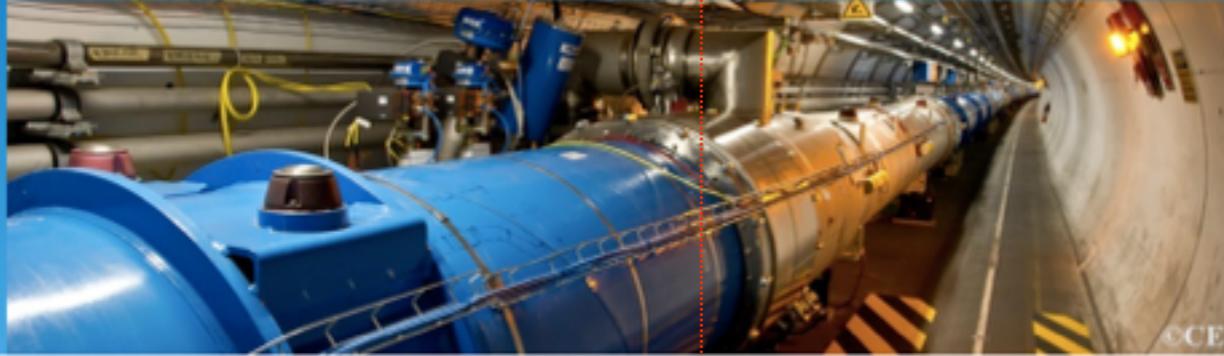
Parameter	Description	Number of Events	Cross-Section
parameter-0	test for UCI	1000	20

**2. process request**

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

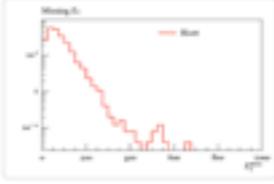
**Request:** test-upload-2  
**ROOT file with TH1:** 20150118095414b5872ab0-1a2b-10a4-c154-5cead413bc8f.zip  
**Status:** Completed

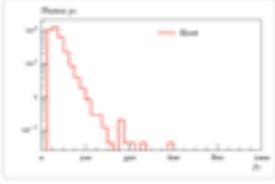
## Results for request 4cdc558c-8f4a-eab4-fdbf-cd91a34db4b2 - parameter-0

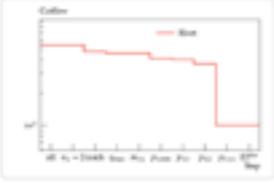
**Efficiency**

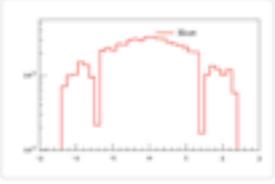
0.18272727272727274

**Plots**

> MET: 

> PhotonPt: 

> Outflow: 

> 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.

The screenshot shows a Zenodo record page for a recast request response. The URL is <https://sandbox.zenodo.org/record/84#.VUESk9NVhBc>. The record is titled "recast request response 3ee4bfde-739b-c844-99bc-f00b130e1ee3" and was uploaded by Lukas Heinrich on 29 April 2015. The record is marked as "Dataset" and "Embargoed access". The embargoed access ends on 01 January 2016. The DOI is 10.5072/zenodo.84. The license is Creative Commons CCZero. The record is a response to a RECAST request. The preview shows a plot titled "Cutflow" with a red line labeled "Rivet". The x-axis is labeled "Step" and has categories: all,  $n_\gamma = 2$ , crack,  $\eta_{max}$ ,  $m_{\gamma\gamma}$ ,  $PT_{min}$ ,  $PT_1$ ,  $PT_2$ ,  $PT_{\gamma\gamma}$ , and  $E_T^{miss}$ . The y-axis is labeled "Cutflow" and has a tick mark at  $10^3$ . The plot shows a step-like function that decreases as the steps progress, with a sharp drop at the  $PT_{\gamma\gamma}$  step.

29 April 2015 Dataset Embargoed access

## recast request response 3ee4bfde-739b-c844-99bc-f00b130e1ee3

Heinrich, Lukas  
(show affiliations)  
response to a RECAST request

Preview

Page: 1 of 1 - + 110%

Cutflow

Rivet

$10^3$

all  $n_\gamma = 2$  crack  $\eta_{max}$   $m_{\gamma\gamma}$   $PT_{min}$   $PT_1$   $PT_2$   $PT_{\gamma\gamma}$   $E_T^{miss}$   
Step

Files

Publication date: 29 April 2015  
Embargoed  
Files available as Open Access after 01 January 2016  
DOI: [10.5072/zenodo.84](https://doi.org/10.5072/zenodo.84)  
License (for files): [Creative Commons CCZero](https://creativecommons.org/licenses/by/4.0/)  
Uploaded by: [lukasheinrich](#) (on 29 April 2015)

Share

Cite as

Heinrich, Lukas. (2015). recast request response 3ee4bfde-739b-c844-99bc-f00b130e1ee3. Zenodo. [10.5072/zenodo.84](https://doi.org/10.5072/zenodo.84)

Select citation style...

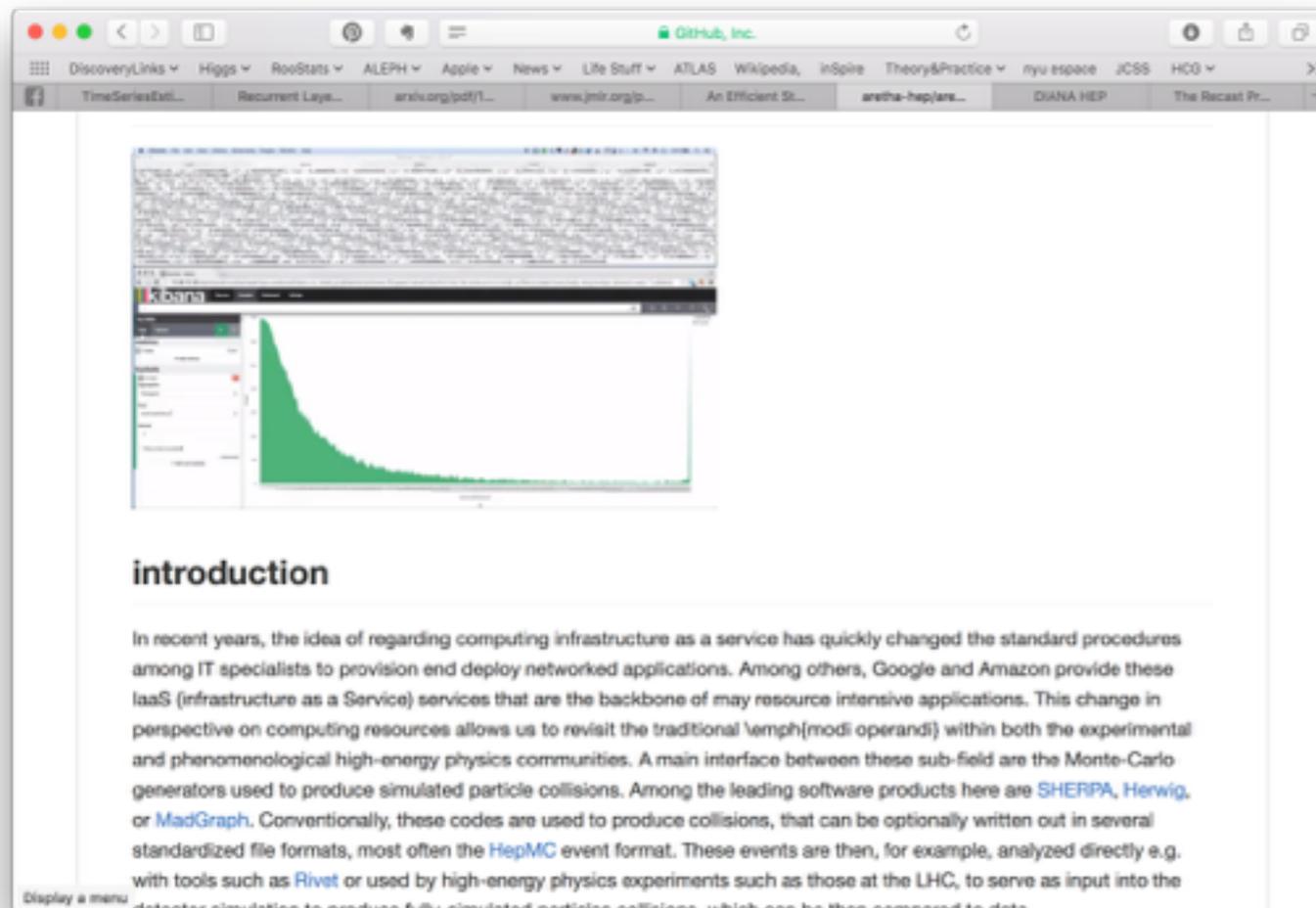
Export

[BibTeX](#), [DataCite](#), [DC](#), [EndNote](#), [NLN](#), [RefWorks](#)  
[MARC](#), [MARCXML](#)

# 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

Mathematica    GitHub/GitLab

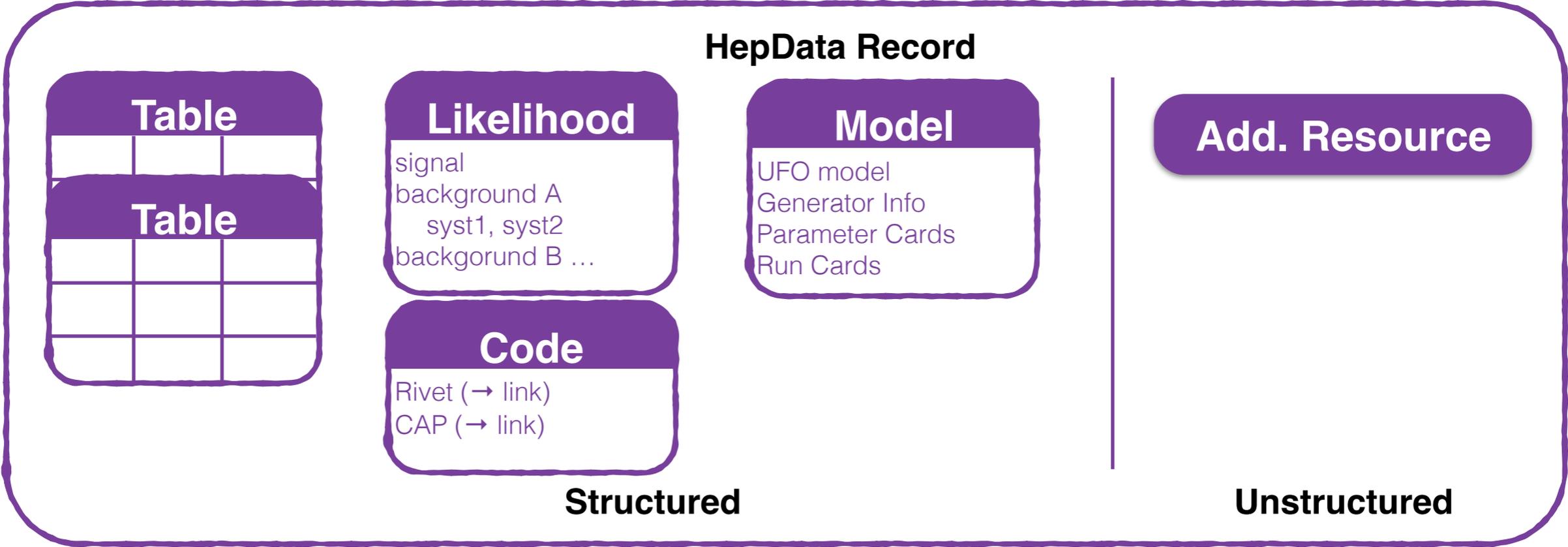
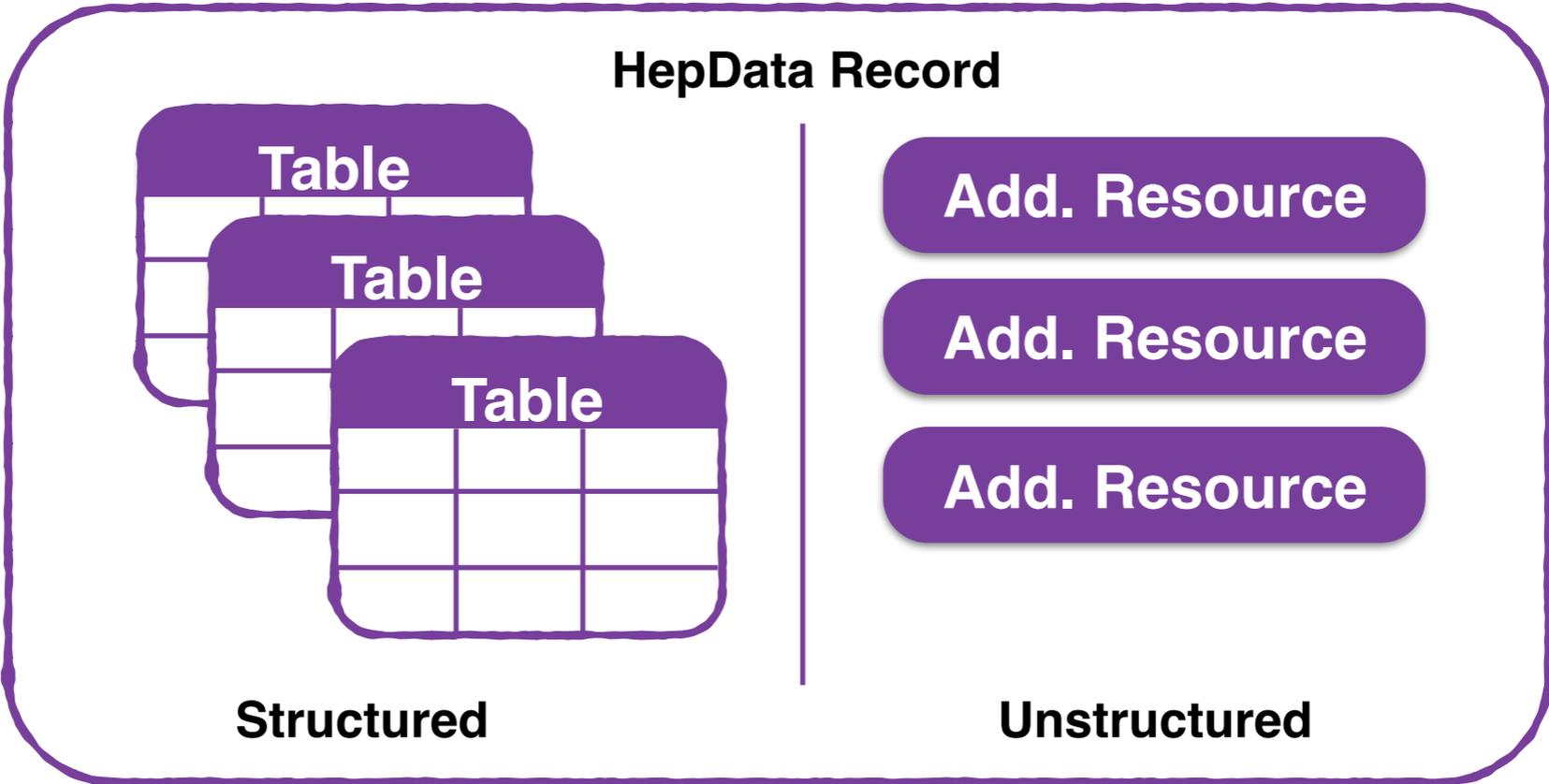
Jupyter/IPython    Zenodo

MadAnalysis    MadGraph

Checkmate    Sherpa

Atom, FastLim, ...    Pythia

...





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)",... )

The screenshot shows the HepData interface. On the left, a list of tables is displayed, with 'Table 1', 'Table 2', 'Table 3', and 'Table 4' highlighted with red lines. Each table entry includes a title, a description, and a URL. On the right, a detailed view of a table is shown, including a filter for 'cmenergies' set to 7000.0. Below this, a table with columns 'RE', 'SQRT(S)', 'DATA', and 'MC' is visible. The 'RE' column contains 'ET(C=MISSING) [GEV]' and the 'DATA' column contains 'Events/10 GeV'. The 'SQRT(S)' column shows '7000.0 GeV' and the 'DATA' column shows '2.3' and '3.5'.

RE	P P --> GAMMA
SQRT(S)	7000.0 GeV
	DATA MC
ET(C=MISSING) [GEV]	Events/10 GeV
0.00 - 10.00	2.3 3.5

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

Additional Publication Resources

Here you'll find any code, additional papers, etc. relating to the entire publication.

Rivet analysis

Extra data files (exclusion limits, acceptance\*efficiency, slha files)

This is a link to an external resource which you can view by clicking the button below.

Download index.shtml

VS

Additional Publication Resources

Here you'll find any code, additional papers, etc. relating to the entire publication.

Link to SLHA files

CERN-PH-EP-2012-308,  
arXiv:1211.1167

ATLAS SLHA files

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.