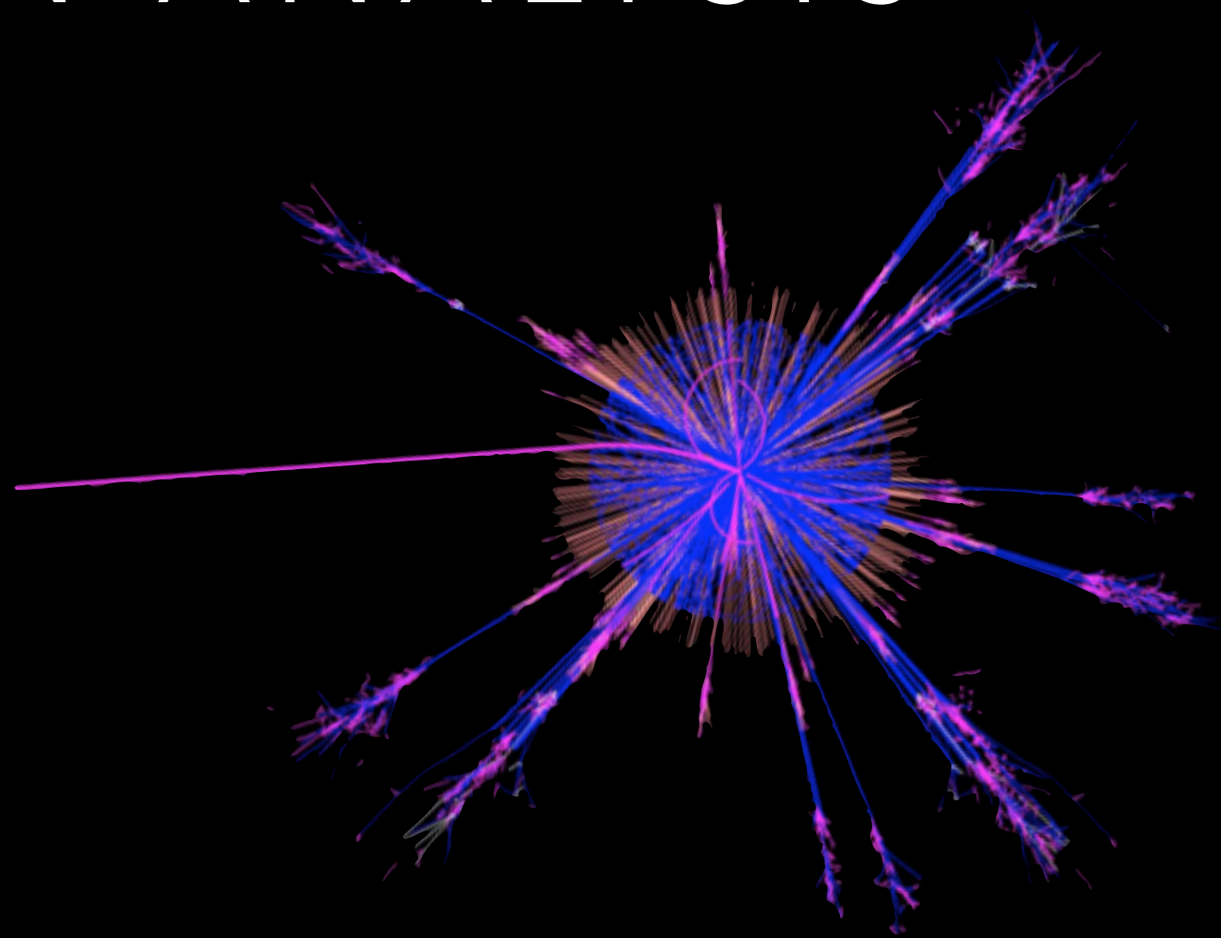
NYU Center for Data Science

Center for Cosmology and particle physics

INFRASTRUCTURE FOR REINTERPRETATION

RECAST & CERN ANALYSIS PRESERVATION



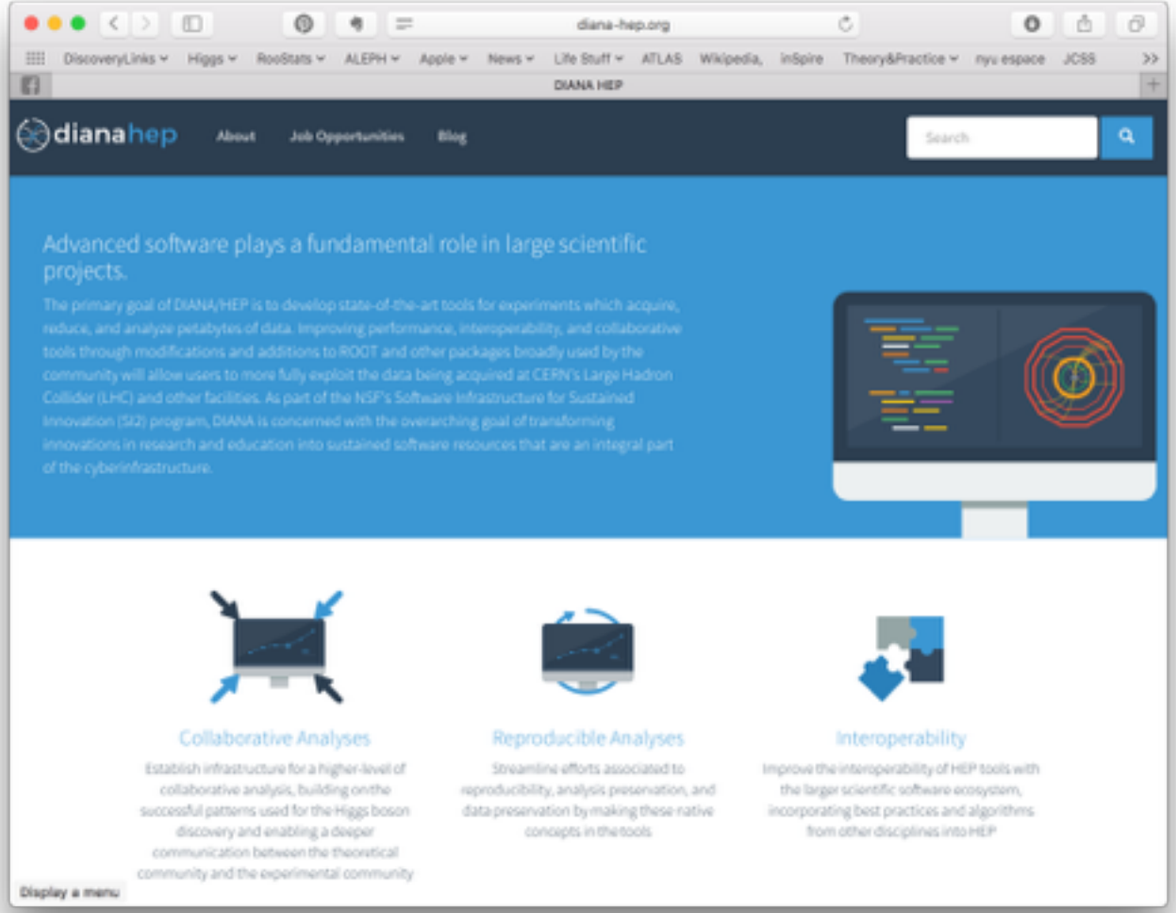
@KyleCranmer
New York University
Department of Physics
Center for Data Science

DASPOS AND DIANA

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

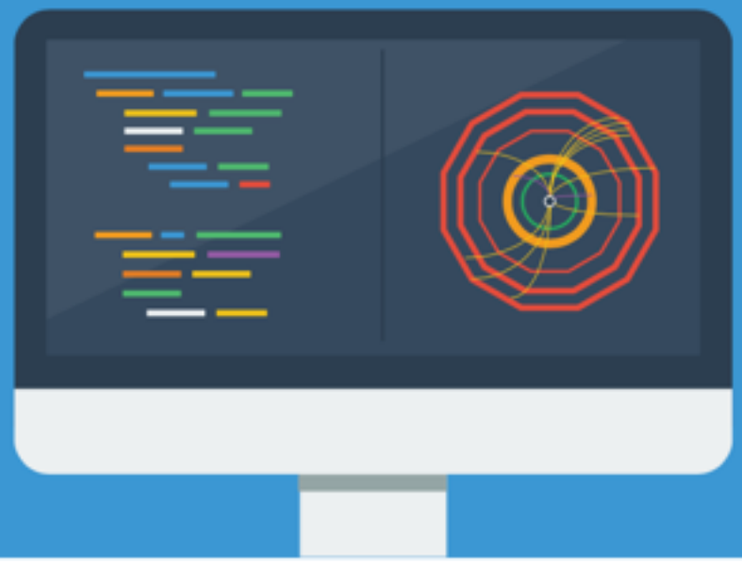
We are working closely with CERN Analysis Preservation (CAP) portal, INSPIRE, and HEPData to build infrastructure for High Energy Physics

- I will focus on infrastructure to support reinterpretation / recasting



Advanced software plays a fundamental role in large scientific projects.

The primary goal of DIANA/HEP is to develop state-of-the-art tools for experiments which acquire, reduce, and analyze petabytes of data. Improving performance, interoperability, and collaborative tools through modifications and additions to ROOT and other packages broadly used by the community will allow users to more fully exploit the data being acquired at CERN's Large Hadron Collider (LHC) and other facilities. As part of the NSF's Software Infrastructure for Sustained Innovation (SI2) program, DIANA is concerned with the overarching goal of transforming innovations in research and education into sustained software resources that are an integral part of the cyberinfrastructure.



Collaborative Analyses

Establish infrastructure for a higher-level of collaborative analysis, building on the successful patterns used for the Higgs boson discovery and enabling a deeper communication between the theoretical community and the experimental community

Reproducible Analyses

Streamline efforts associated to reproducibility, analysis preservation, and data preservation by making these native concepts in the tools

Interoperability

Improve the interoperability of HEP tools with the larger scientific software ecosystem, incorporating best practices and algorithms from other disciplines into HEP

e.g. Python & Mathematica

ANALYSIS PRESERVATION ACTIVITIES

LHC experiments are putting effort into "analysis preservation" in order to

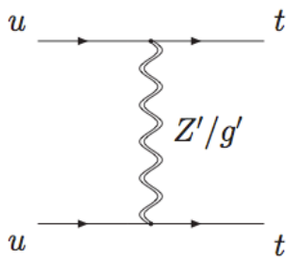
- ensure reproducibility of published results,
- streamline extension of analysis to new data as graduate students transition,
- reinterpret existing analysis in the context of new theories (aka "recasting")

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

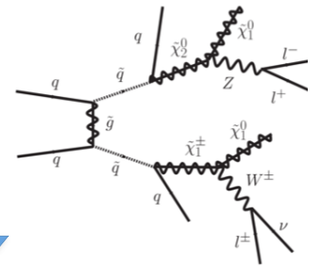
Level-1. Published results

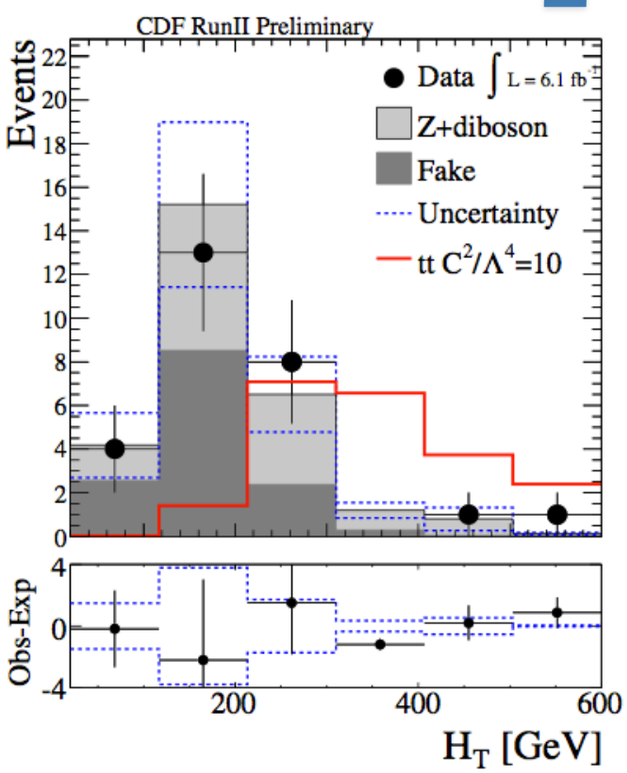
All scientific output is published in journals, and preliminary results are made available in Conference Notes. All are openly available, without restriction on use by external parties beyond copyright law and the standard conditions agreed by CERN.

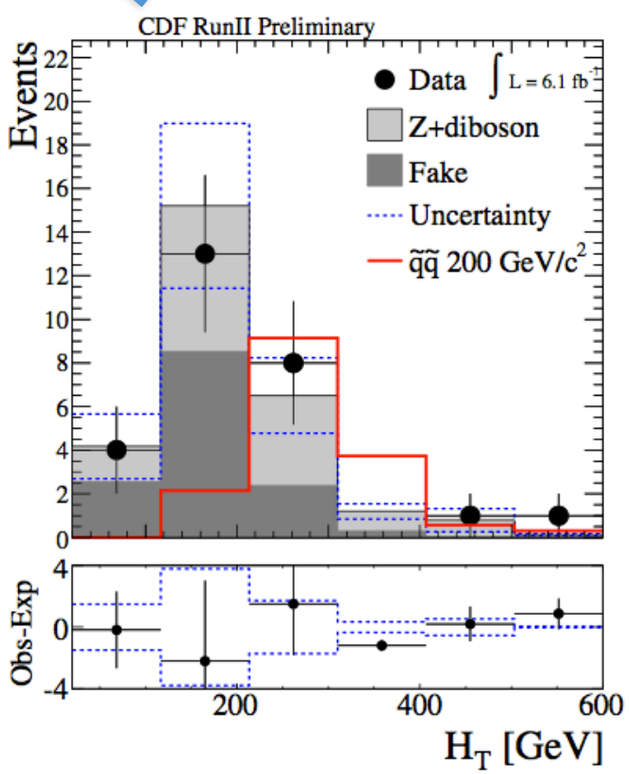
Data associated with journal publications are also made available: tables and data from plots (e.g. cross section values, likelihood profiles, selection efficiencies, cross section limits, …) are stored in appropriate repositories such as HEPDATA[2]. ATLAS also strives to make additional material related 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







CERN ANALYSIS PRESERVATION

Significant effort by CERN to provide service for the experiments to aid analysis preservation.

- many of the same people involved in the CERN Open Data portal, but the focus here is different. This is a service for the experiments, not expected to be open.

- close collaboration with representatives from LHC experiments and DASPOS

Two main approaches being pursued in parallel:

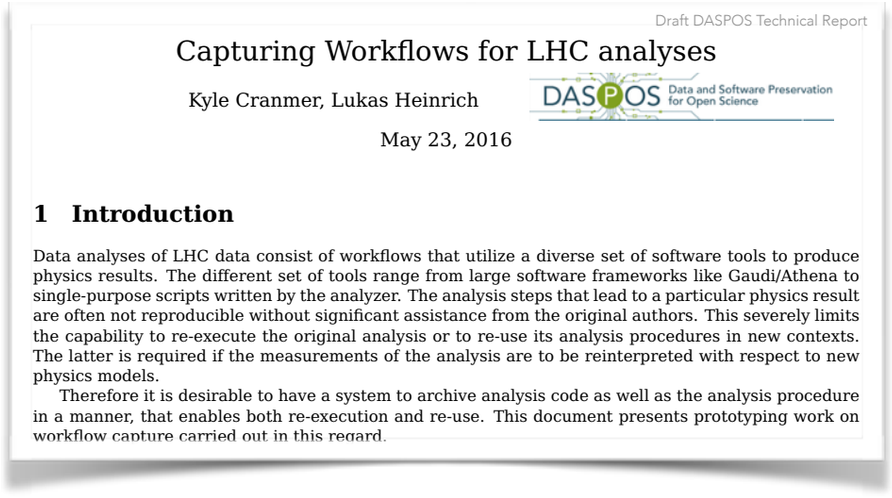
1) a meta-data model to describe all aspects of an analysis

   - including but not limited to cuts, triggers, etc. (similar to "Les Houches Analysis Description Accord")

2) directly capture computational workflows for reproducibility

   - i.e. the code for the published analysis, this is more relevant for Recast

CAPTURING AN ANALYSIS


Draft DASPOS Technical Report

Capturing Workflows for LHC analyses

Kyle Cranmer, Lukas Heinrich

DASPOS Data and Software Preservation for Open Science

May 23, 2016

1 Introduction

Data analyses of LHC data consist of workflows that utilize a diverse set of software tools to produce physics results. The different set of tools range from large software frameworks like Gaudi/Athena to single-purpose scripts written by the analyzer. The analysis steps that lead to a particular physics result are often not reproducible without significant assistance from the original authors. This severely limits the capability to re-execute the original analysis or to re-use its analysis procedures in new contexts. The latter is required if the measurements of the analysis are to be reinterpreted with respect to new physics models.

Therefore it is desirable to have a system to archive analysis code as well as the analysis procedure in a manner, that enables both re-execution and re-use. This document presents prototyping work on workflow capture carried out in this regard.

Two complementary strategies for capturing an analysis

- meta-data describing the analysis at a high-level
- code, environment, etc. needed to re-execute the computational workflow

In addition to re-executing the workflow on the exact same inputs (reproducibility), we also want to be able to **reuse** the workflow on new inputs or with different settings

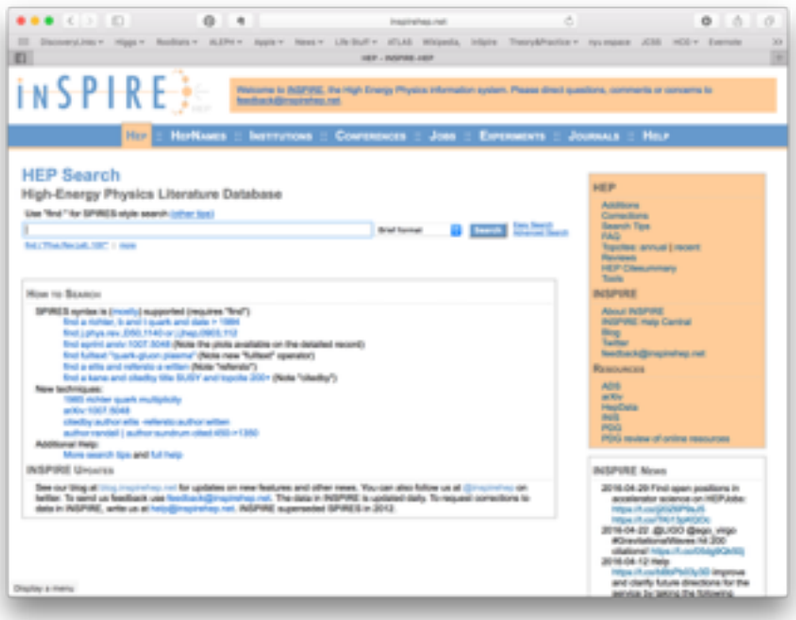
- we call this a parametrized workflow, needed for reinterpretation / “recasting”

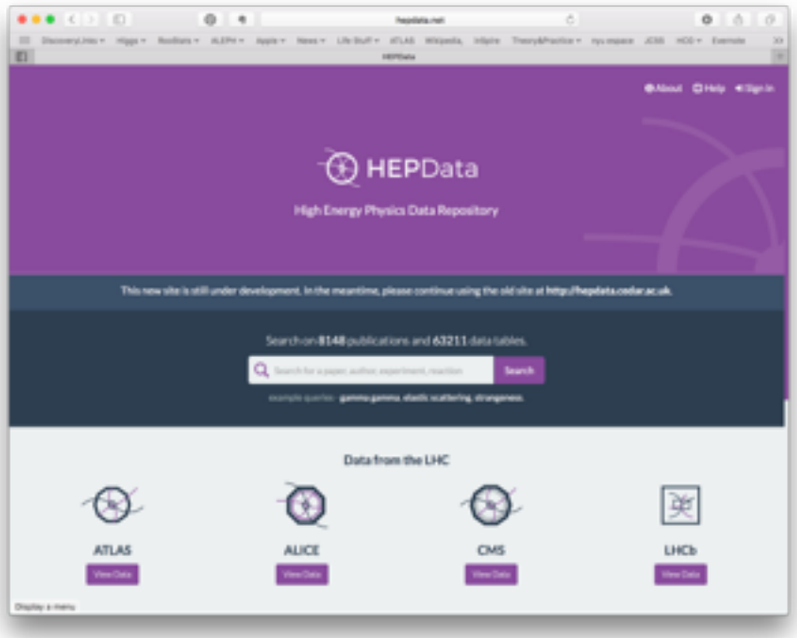
We have developed JSON schemas to capture two types of ingredients

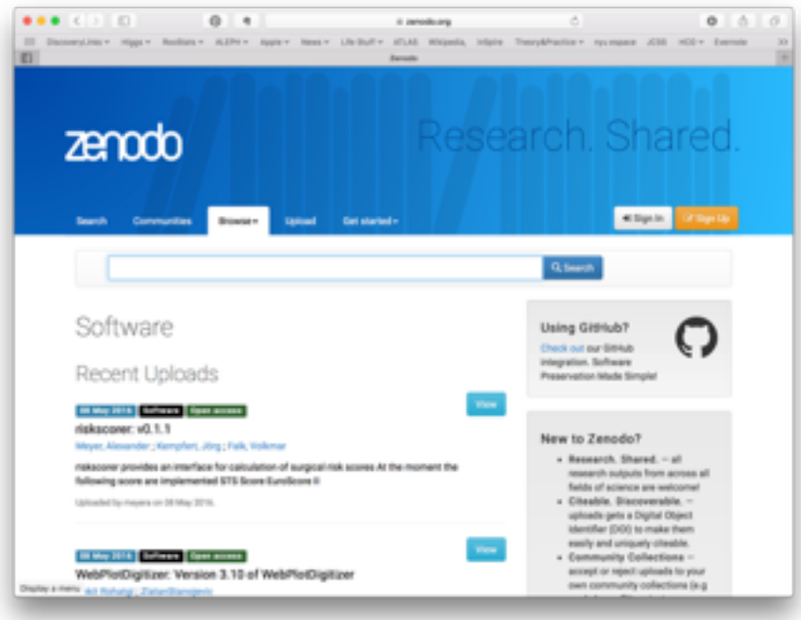
- **packtivity** to describe individual processing stages (docker container, options, …)
- **yadage** to describe how to connect the pieces together into a parametrized workflow
- CERN ANALYSIS PRESERVATION can now store and serve up analyses preserved this way

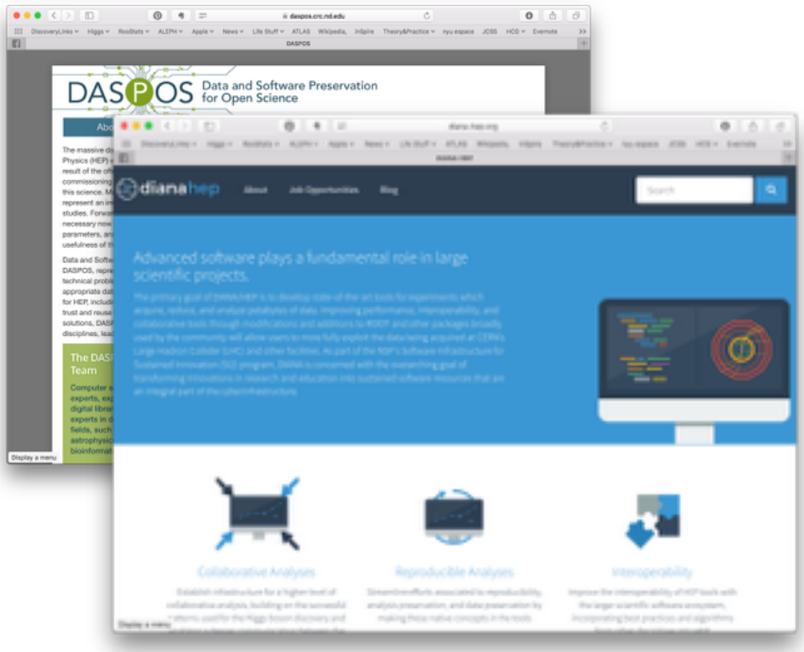
We leverage docker so that each processing stage can have its own computing environment. Recast backend can run new theory through this workflow for reinterpretation.

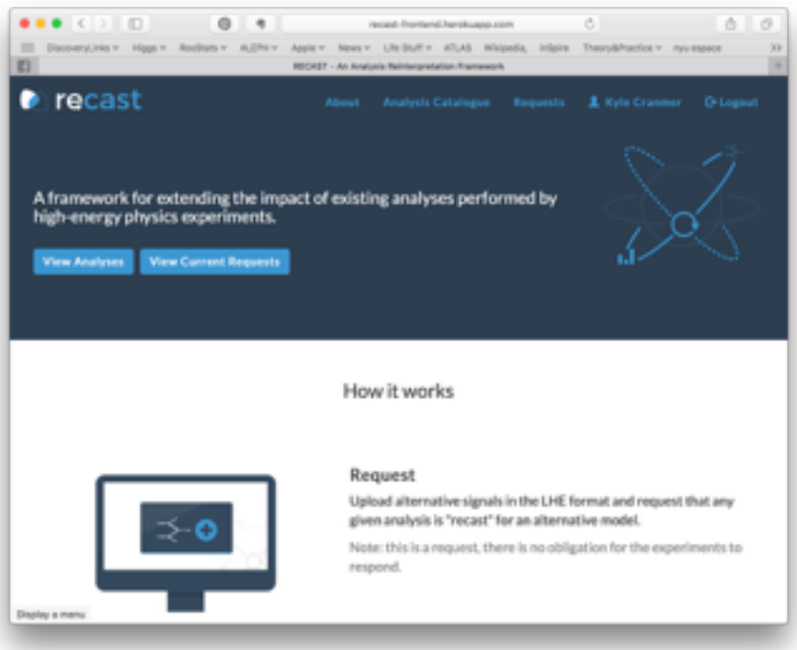
INFRASTRUCTURE FOR DATA AND ANALYSIS PRESERVATION







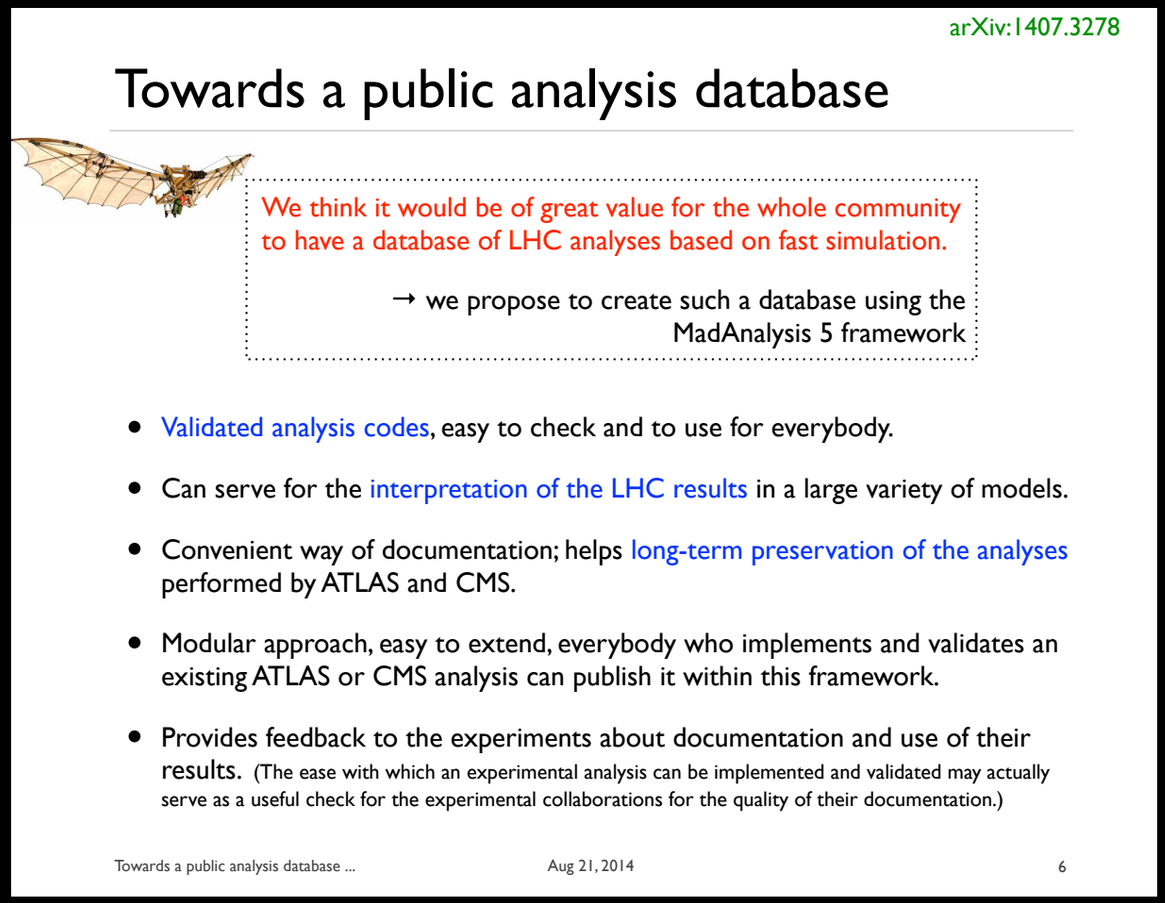




PHENO RECASTING SOFTWARE

- 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
  - Gambit
  - SModelS
  - XQCAT
  - CheckMate
  - unofficial contributions to Rivet

Sabine Kraml



- As I'll show, it is possible to interface RECAST infrastructure with these unofficial pheno recasting tools.

Current status of ATLAS policy

JUST TO CLARIFY

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.

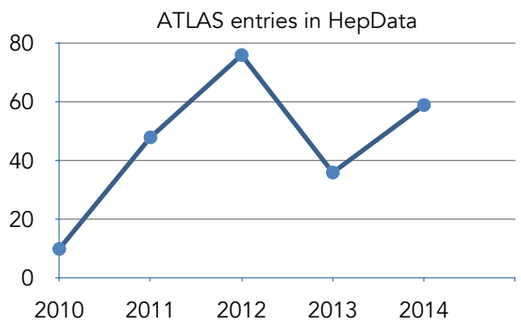
CHEP 2015 POSTER

Analysis Preservation in ATLAS

ATLAS Data Access Policy

> *"ATLAS has fully supported the principle of open access in its publication policy."*

Data associated with journal publications are made available: tables and data from plots



ATLAS also strives to make additional material related to the paper available to allow for a reinterpretation of the data in the context of new theoretical models. For example:

- Information on signal acceptances of searches is also entered in **HepData** to allow reinterpretation of these searches in a limited context
- Simplified, portable and self-contained formats for educational and public understanding purposes
- **RIVET** for encapsulation of unfolded measurements
- ATLAS is also exploring how to provide the capability for reinterpretation of searches in the future via a service such as **RECAST.** RECAST allows theorists to evaluate the sensitivity of a published analysis to a new model they have developed by submitting their model to ATLAS.

link to policy document: http://bitly.com/ZTvcWi

Scope and Purpose

Reproducibility ← → **Replicability**

"**Reproducibility**" is defined as repeating the analysis of the same data using the original procedures, software and tools.

- **Primary Technology:** virtualization, containerization
- **Timescale**: short/medium term
- **Use case:** confirmation & clarification if questions arise, reinterpretation of existing result for new physics model

"**Replicability**" is defined as repeating the analysis of new data or new versions of old data (e.g. after a reprocessing), potentially with newer versions of software and tools.

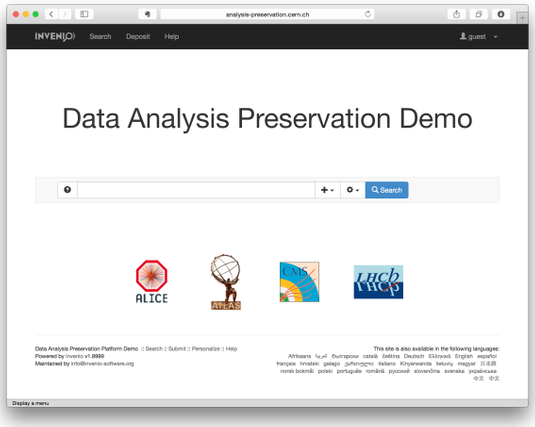
- **Primary Technology:** migration, regression testing
- **Timescale:** medium/long term
- **Use case:** extend analysis with new data, facilitate migration to new groups or similar signatures

*(this approach overlaps with our **DPHEP** efforts)*

An Eye On The Future

ATLAS is now reviewing the concept of analysis preservation with the aim to bring coherence and robustness to the process and with a clearer view of the level of reproducibility that is reasonably achievable.

- ATLAS is working with **CERN-IT** and **DPHEP** to develop a tool to **capture** provenance, derived data, and analysis code at various levels
- ATLAS members of **DASPOS** are exploring generic tools (CDE, PTU, parrot, docker, LXC, etc.) to automatically capture provenance & computing environment that can be preserved & distributed
- ATLAS is prototyping and evaluating a **RECAST** backend that leverages the preserved analysis to provide a service for reinterpretation



http://data-demo.cern.ch/

Lukas Heinrich, for the ATLAS Collaboration

ATLAS EXPERIMENT

Poster for CHEP 2015

The Data Analysis Preservation Demo has been renamed

CERN ANALYSIS PRESERVATION

Status of Recast

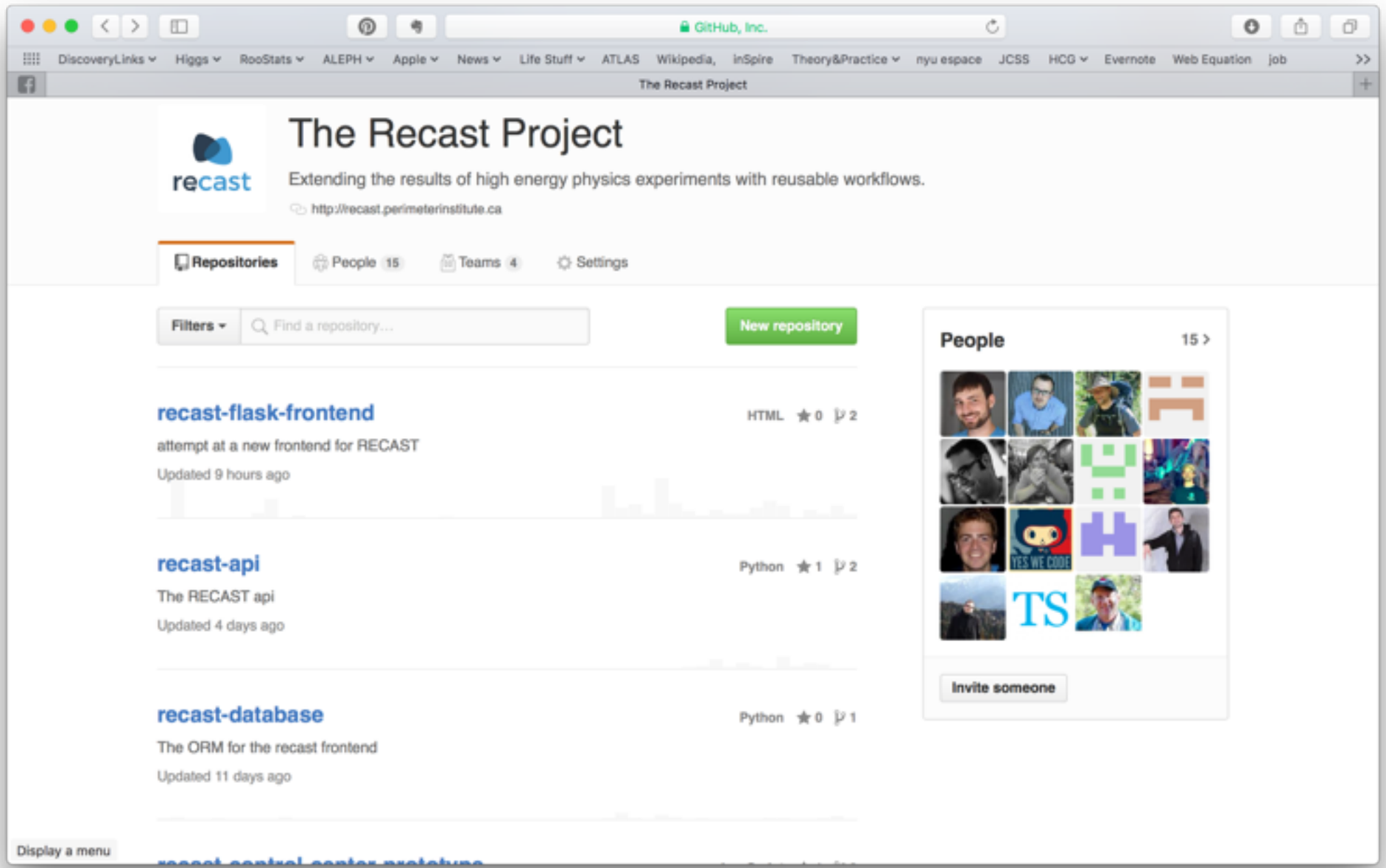
recast

Many people contributing now. Contributions from CERN, DASPOS, DIANA, GitHub, Moore-Sloan Data Science Environment at NYU, Notre Dame, Nebraska, …

Using **yadage** and **packtivity** JSON schemas developed by Lukas Heinrich and described in draft DASPOS technical report for packaging realistic LHC analyses

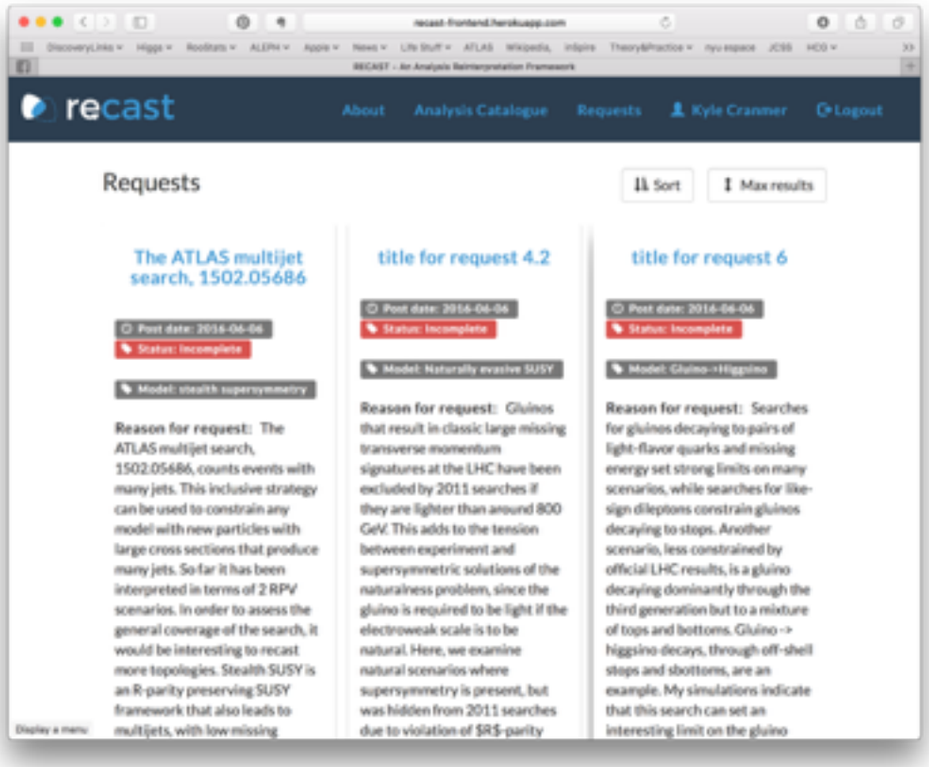
CERN Analysis Portal (CAP) is able to store and serve up analysis workflows stored in this format.

New front-end webpage thanks to Christian Bora (Nebraska, DASPOS) and Eamonn Maguire (CERN)



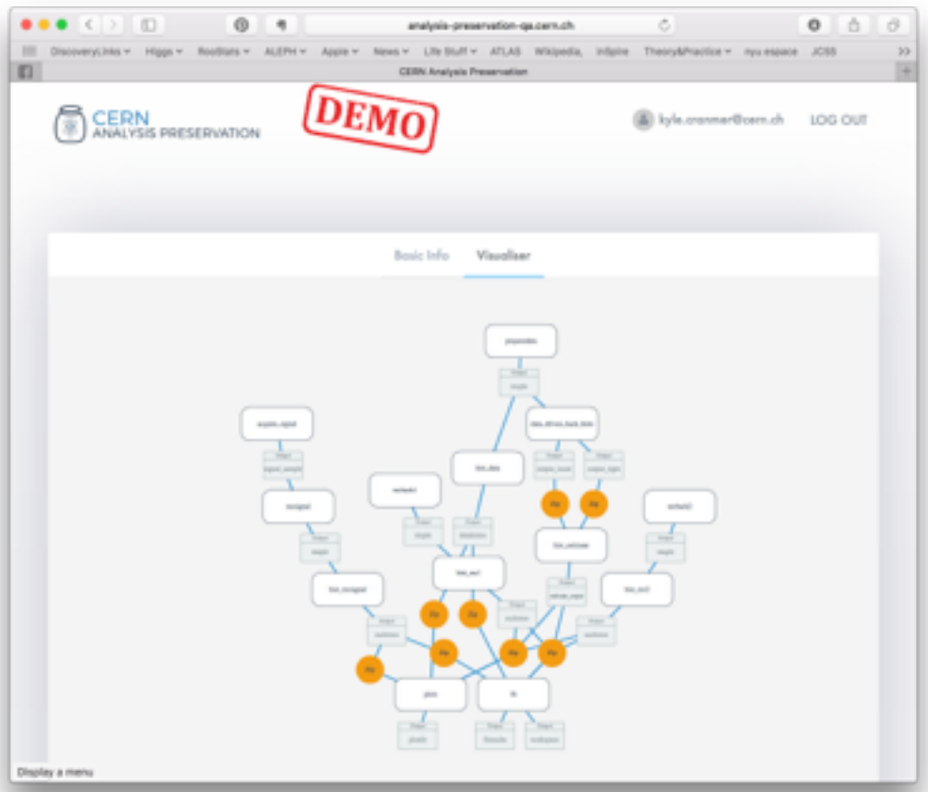
RECAST PROGRESS

1)

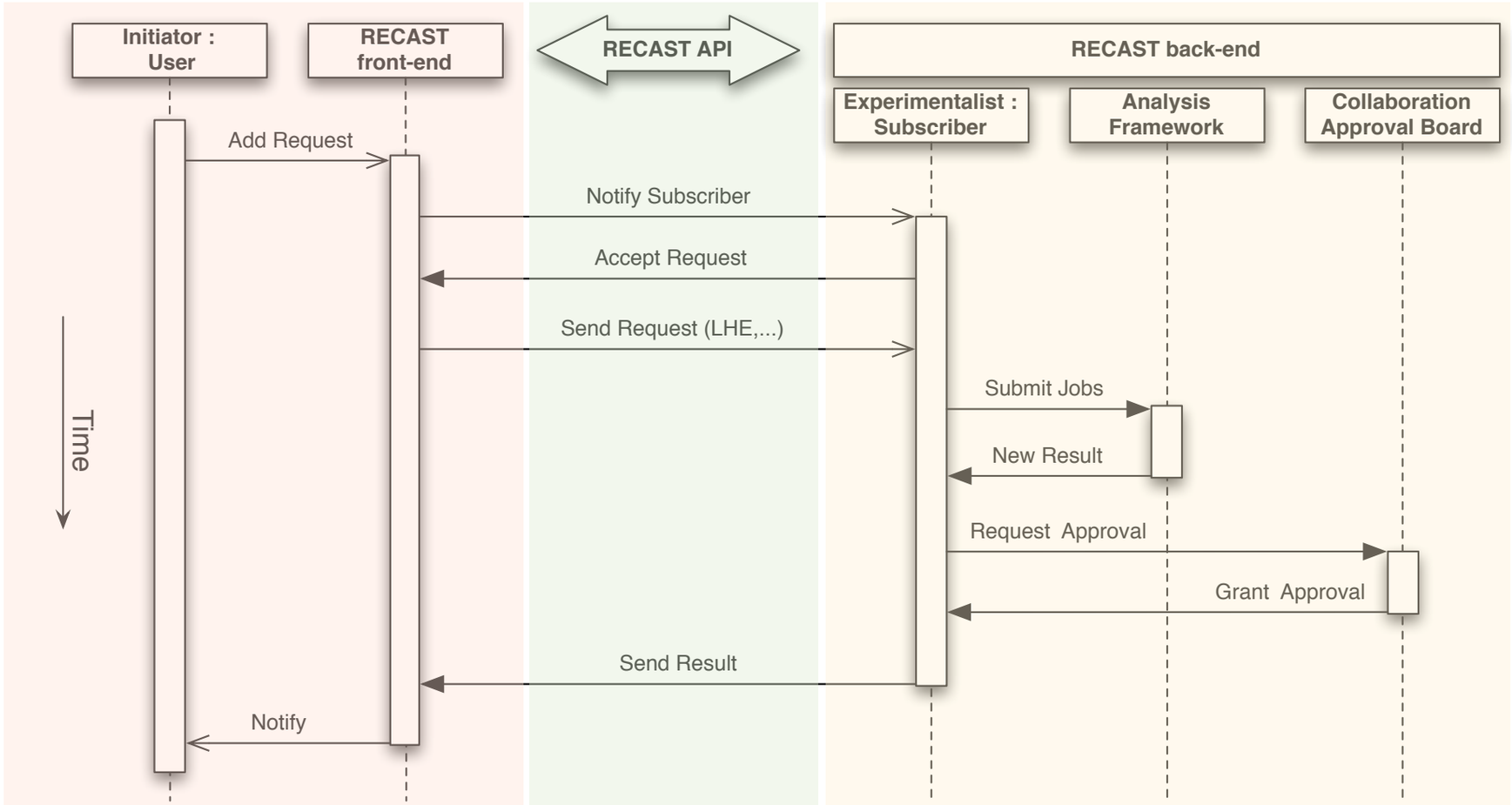


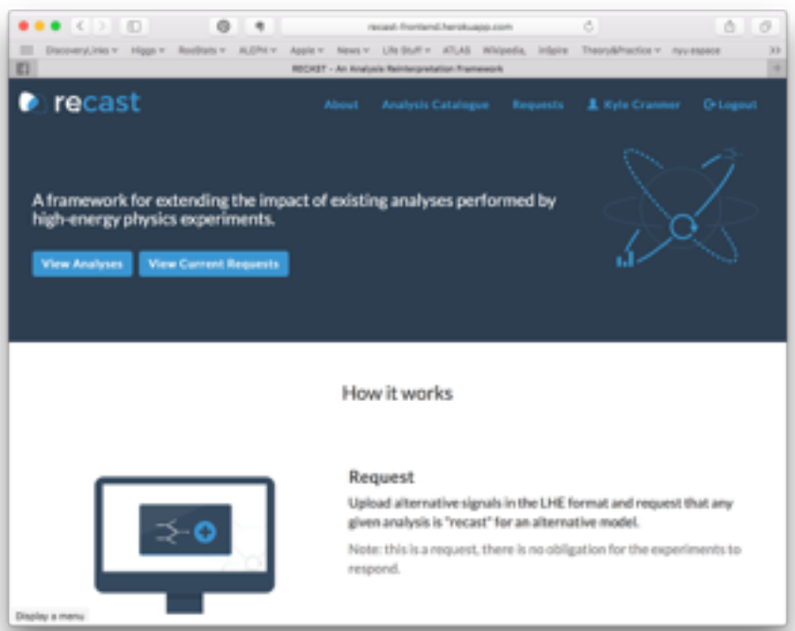
1. Recast front-end user interface developed by Christian Bora at Nebraska-Lincoln (DASPOS)
   - accepts requests for recasting and presents results
   - has a RESTful API and corresponding command-line interface
2. An ATLAS SUSY analysis has been captured using **packtivity** and **yadage** schemas and stored in CERN Analysis Preservation portal (see Kilian Rosbach's talk Thursday)
   - workflow, individual analysis steps, computing environment
3. Recast back-end can pull and re-execute analysis on new theory to reinterpret the original published result

2)

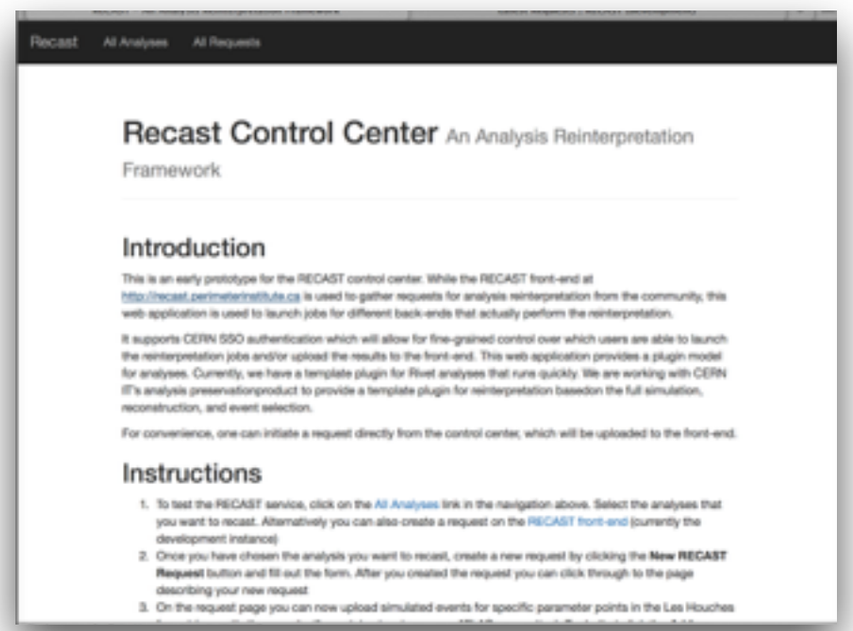


3)

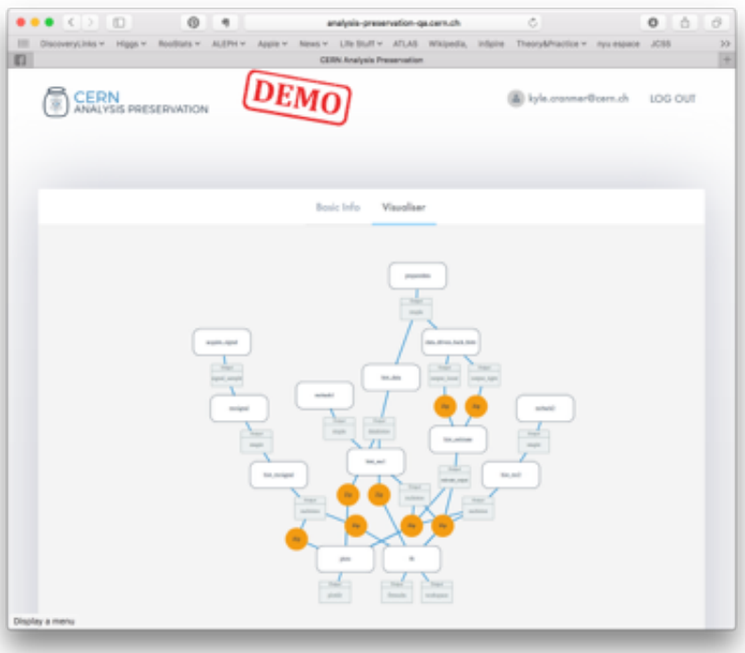




Front-End: public facing
collects requests



Control Center: not public, uses CERN auth.,
oversees processing of jobs on back-end



CERN Analysis Preservation:
Stores workflows, provides back-end
computing resources

front-end (open)

control center (closed to experiment)

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

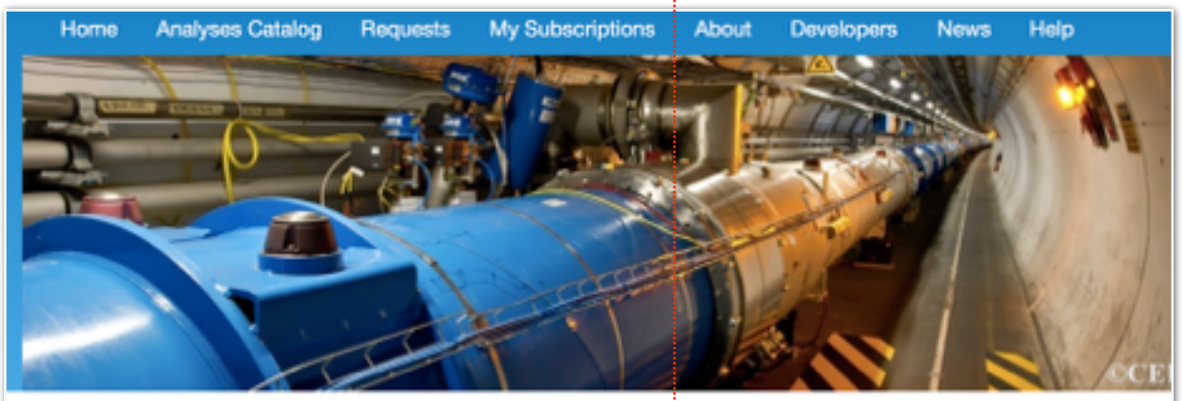
4. upload response

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 » 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

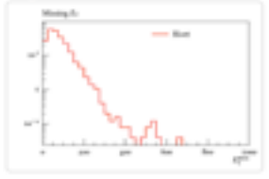
Results for request 4cdc558c-8f4a-eab4-fdbf-cd91a34db4b2 - parameter-0

Efficiency

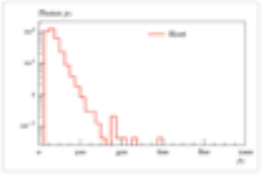
0.18272727272727274

Plots

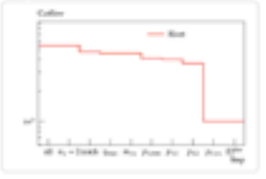
> MET:



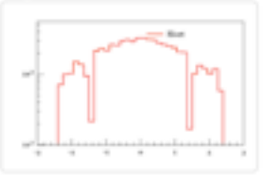
> PhotonPt:



> Cutflow:



> PhotonEta:



INSPIRE & HepData
(stores request/result
& tracks citations)

get citation to
request/response

monitor impact of search

Initiate request

front-end

"Theorist"

review request,
process,
review result,
upload

"Experimentalist"

optional
access control

Control Center

analysis
plugin

analysis
plugin

analysis
plugin

....

Rivet

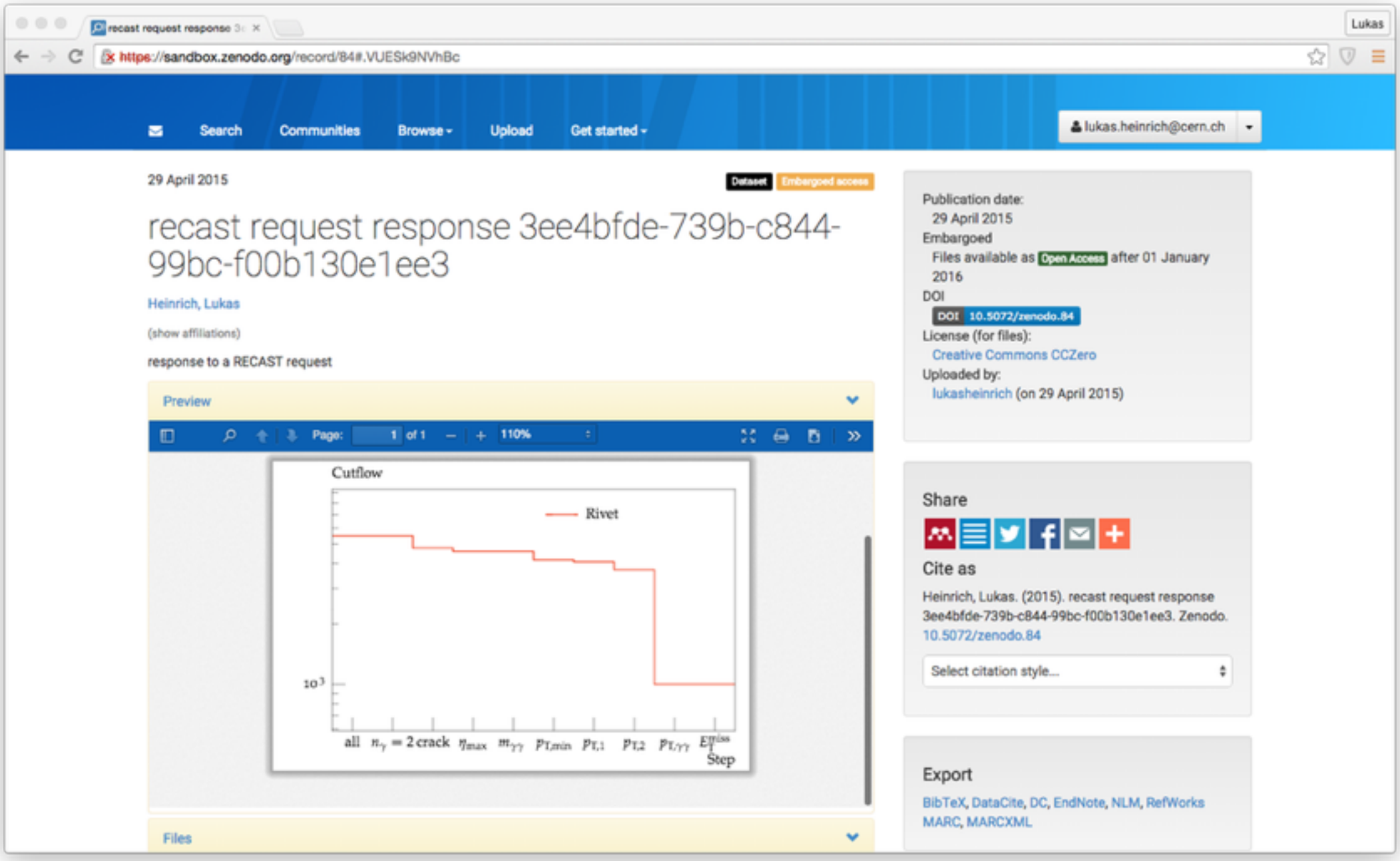
CERN
Analysis
Preservation

Pheno
Tools

EXAMPLE RECAST → HEPDATA / ZENODO

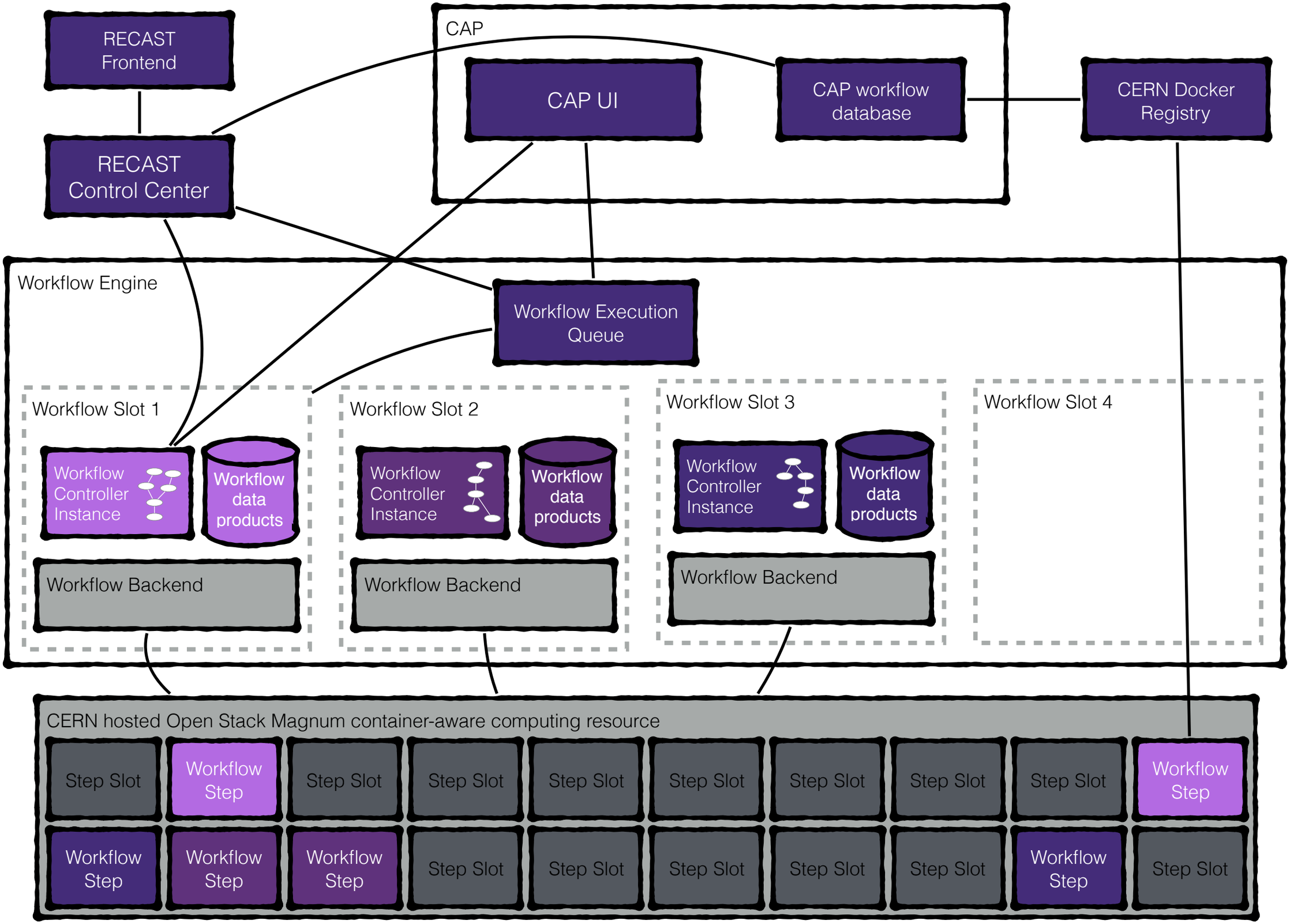
After re-running analysis on new physics model, experiments might want to push result of new interpretation to HEPData. Technically we can do this with Zenodo. Discussing with HEPData and INSPIRE to have API connection to upload result. Both are based on Invenio, so should be easy.

- this allows for new results to get a DOI and be associated with the original analysis publication



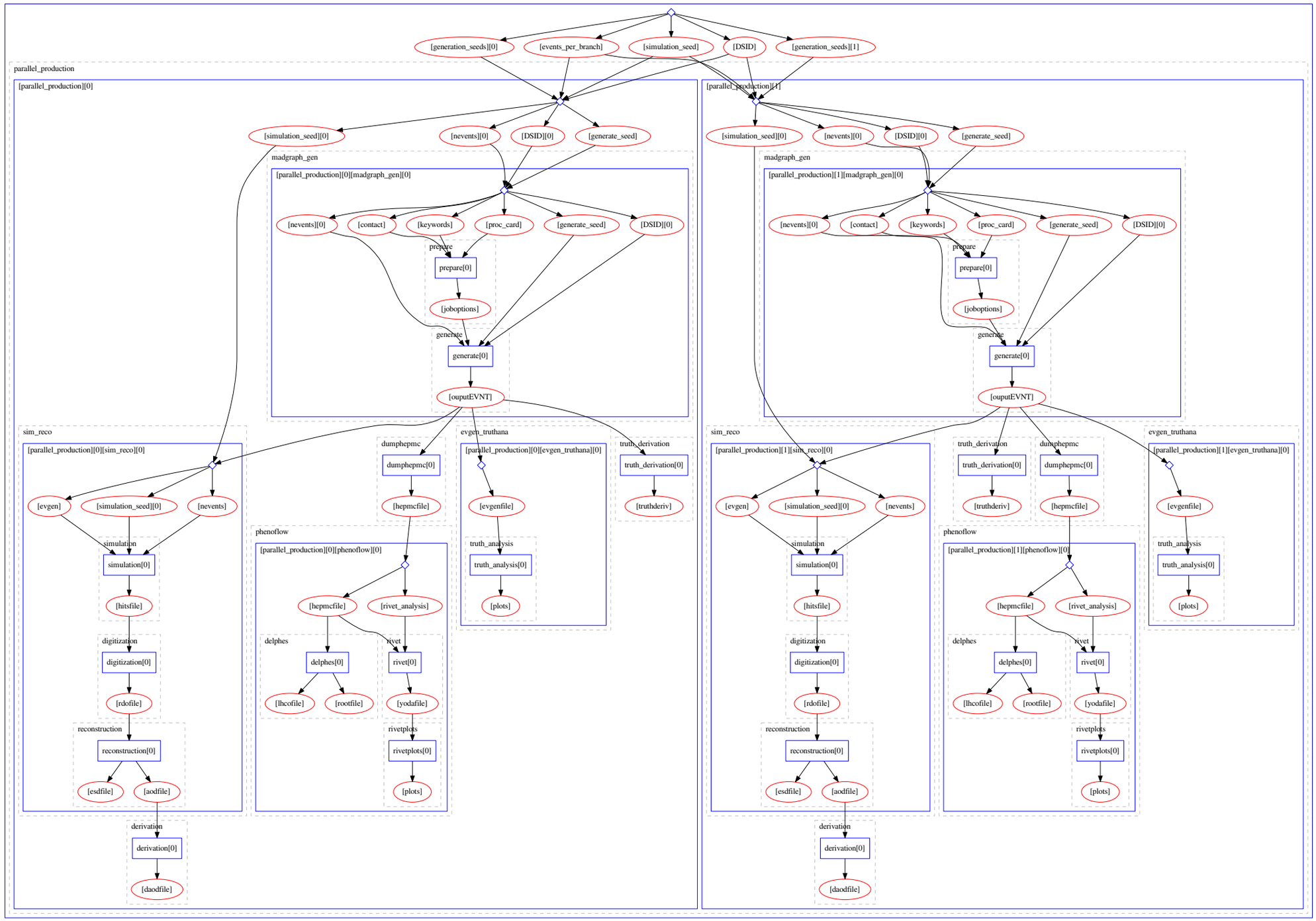
PLANNING FOR CERN-BASED RECAST SERVICE

Schematic of design being developed by CERN / DASPOS / DIANA



A FLEXIBLE WORKFLOW MODEL

A workflow composed of sub-workflows that run Rivet, Delphes, and ATLAS analyses in parallel on the same input



SUMMARY

The experiments are actively engaged in analysis preservation activities that are closely related to reproducible workflows and reinterpretation / recasting

CERN, DASPOS, and DIANA are all contributing to infrastructure for reproducibility and reinterpretation

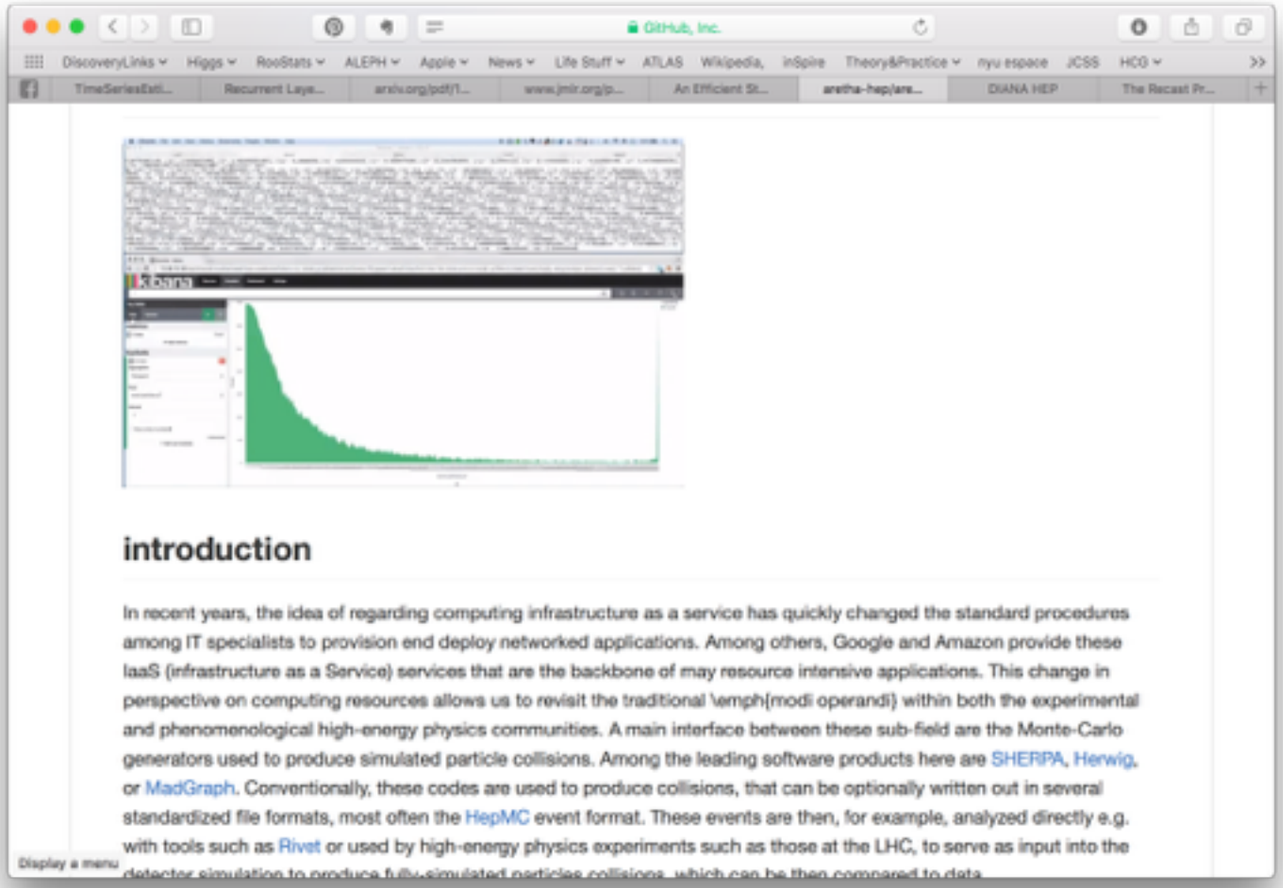
- Plans underway to develop Recast infrastructure integrated with CERN Analysis Preservation Portal

- the Recast infrastructure can be used to run both the analysis code of the experiments, Rivet, and pheno recasting tools

"NETFLIX FOR MONTE CARLO"

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

- ideally, give DOIs to the generator cards and docker container
- can generate more consistent MC on demand