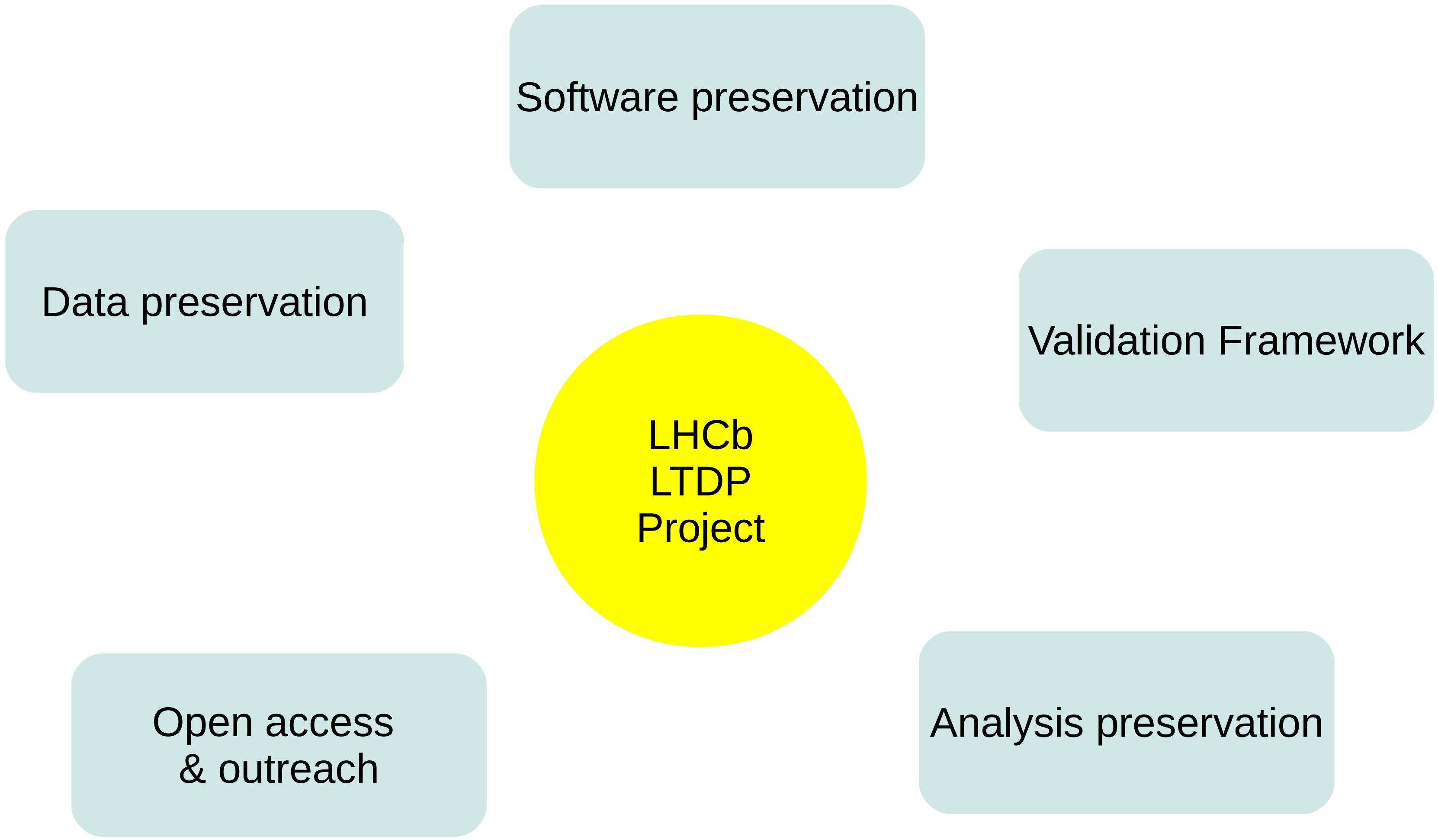
Update on DP activities in LHCb

- June 09th , 2015 -

S.Amerio (U. Padova)

On behalf of LHCb Data Preservation group

Overview of LHCb DP project



Software preservation

Data preservation

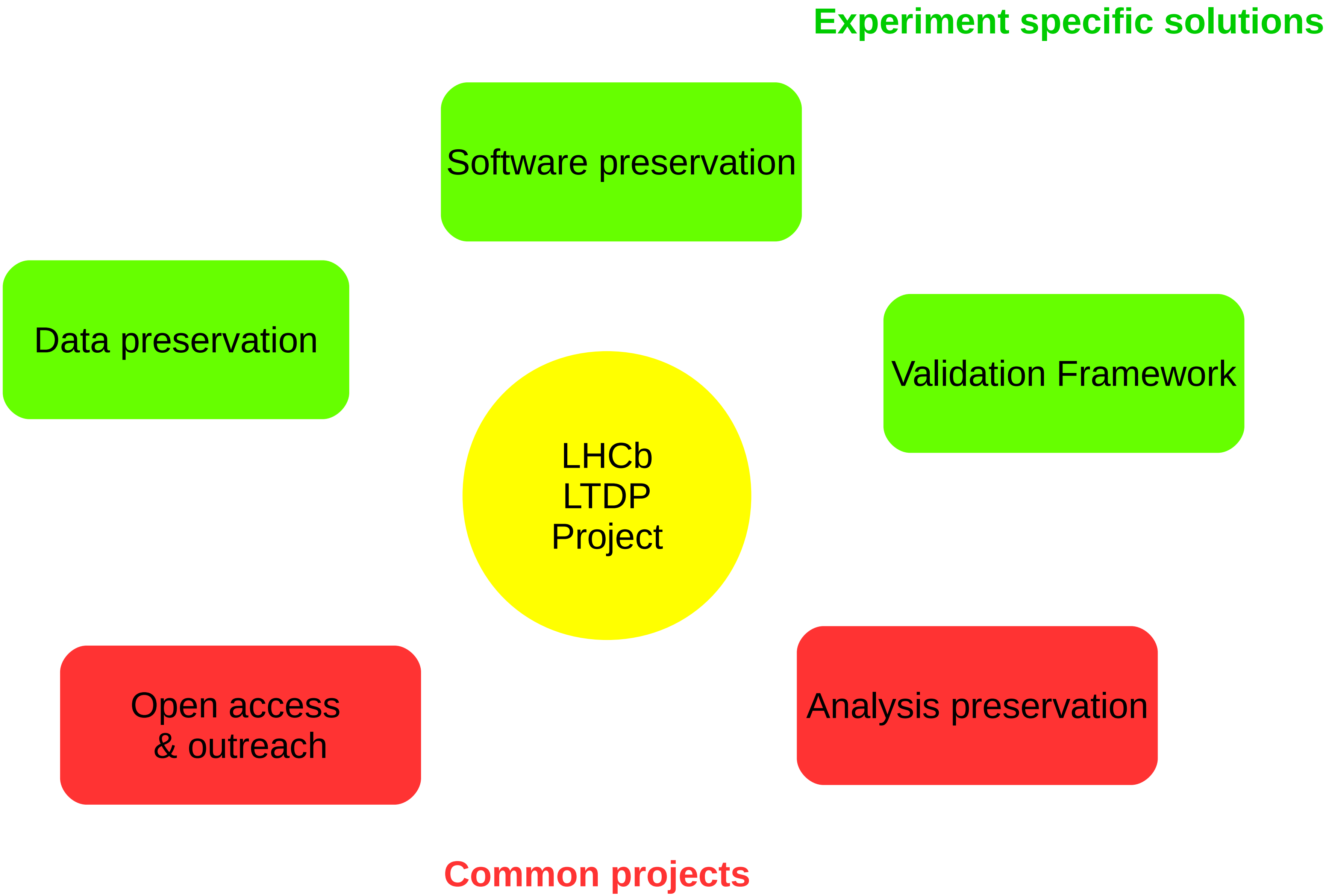
Validation Framework

LHCb LTDP Project

Open access & outreach

Analysis preservation

Overview of LHCb DP project



Experiment specific solutions

Software preservation

Data preservation

Validation Framework

LHCb
LTDP
Project

Open access
& outreach

Analysis preservation

Common projects

LHCb and the OpenData Portal

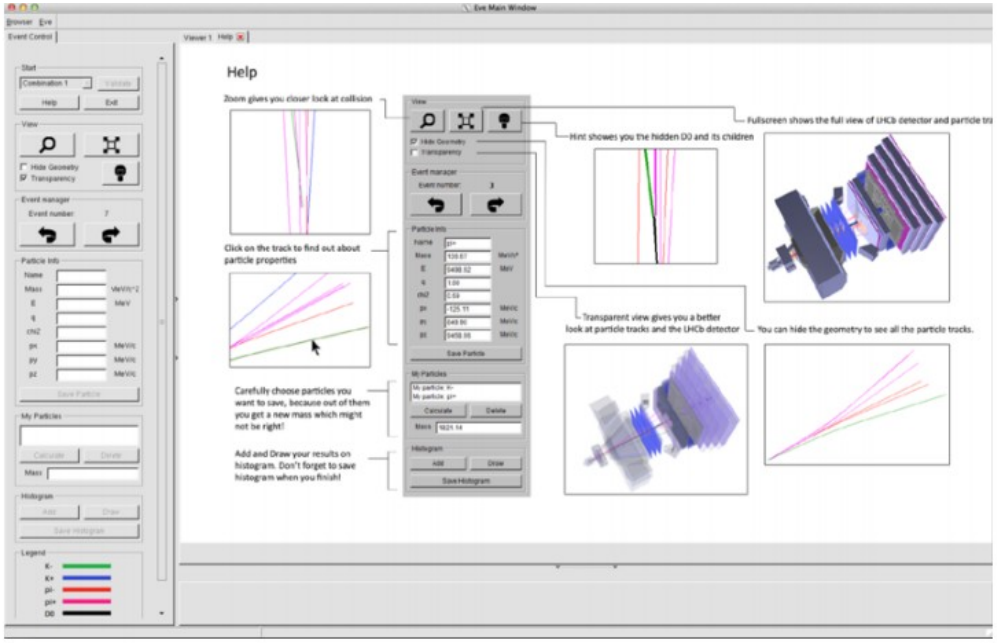
Current status:

DATA

- Event display data (5k events from 2011 data taking)
- D0 → K pi data (60k events from 2011 data taking)
- Format: ROOT trees

SOFTWARE

- LHCb virtual machine image
- Event display and D0 lifetime analysis software (ROOT)



Help

Zoom gives you closer look at collision

Click on the track to find out about particle properties

Carefully choose particles you want to save, because out of them you get a new mass which might not be right!

Add and Draw your results on histogram. Don't forget to save histogram when you finish!

Fullscreen shows the full view of LHCb detector and particle tra

Hint showes you the hidden D0 and its children

Transparent view gives you a better look at particle tracks and the LHCb detector

You can hide the geometry to see all the particle tracks.

LHCb and the OpenData Portal

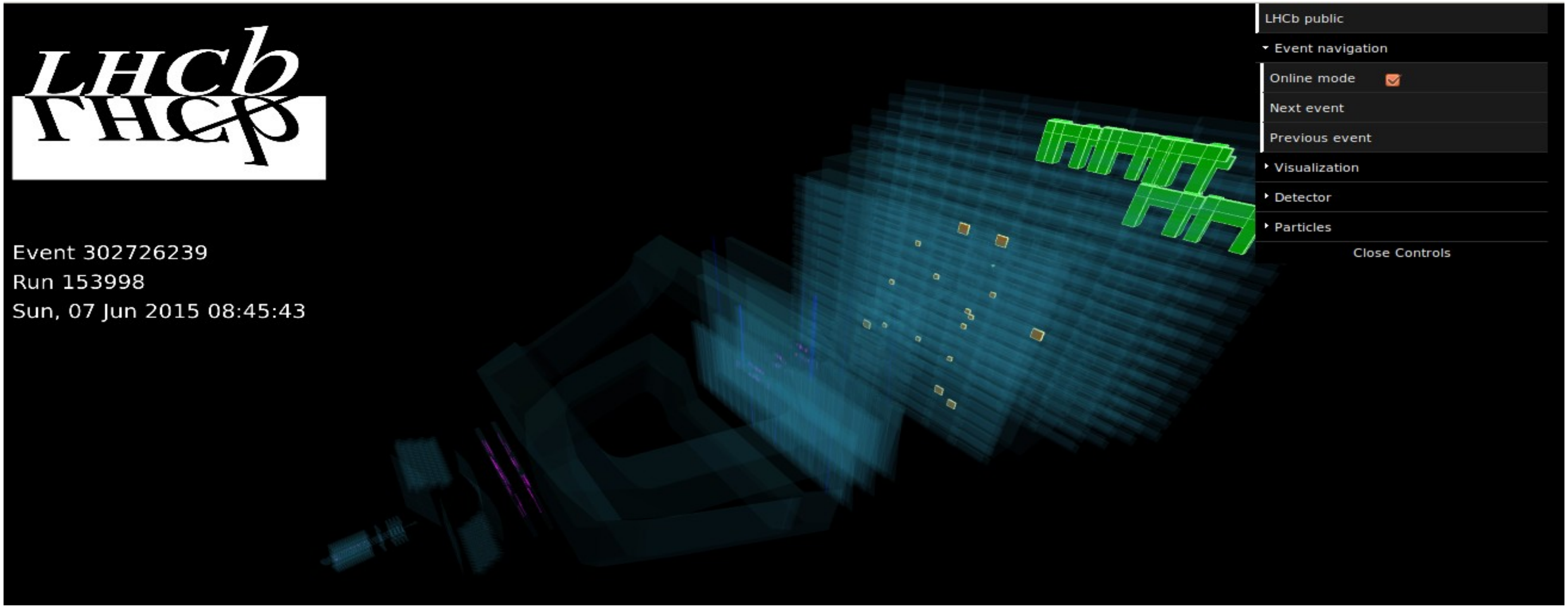
Future plans:

- Enrich the *For Education* area with additional exercises

- Start planning the *For Research area*, in view of our public release:
  - Extract an example analysis (Measurement of CP asymmetry) from *the undergraduate laboratory experiment @ Manchester*
  - Full 2011 data, C++ and ROOT

http://cds.cern.ch/record/1994172?ln=en

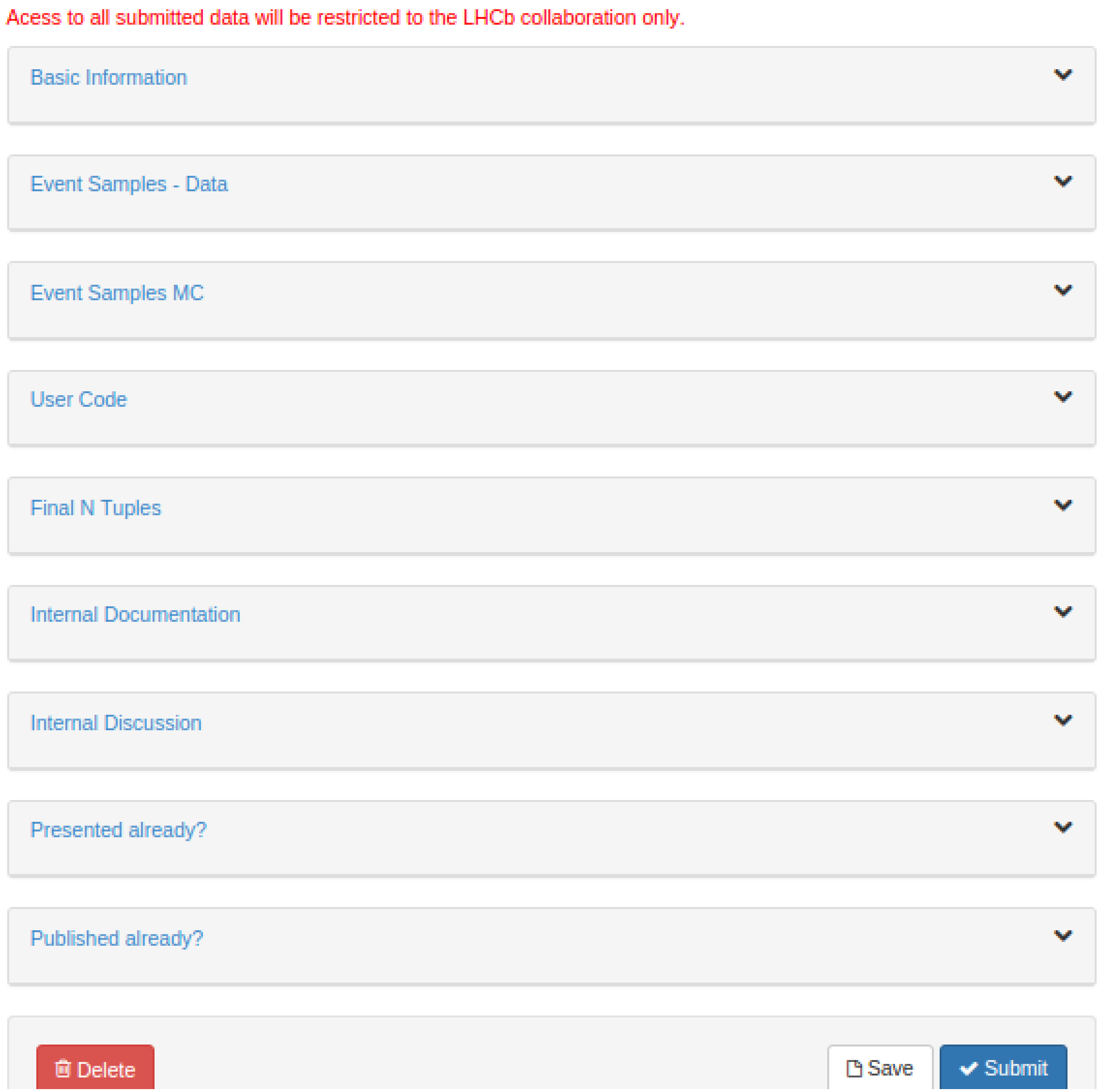
- Add LHCb event previewer exploiting the new 3D WebGL event display (https://lbevent.cern.ch/EventDisplay/index.html)



Event 302726239
Run 153998
Sun, 07 Jun 2015 08:45:43

Analysis preservation framework

Data Analysis Preservation Demo



Acess to all submitted data will be restricted to the LHCb collaboration only.

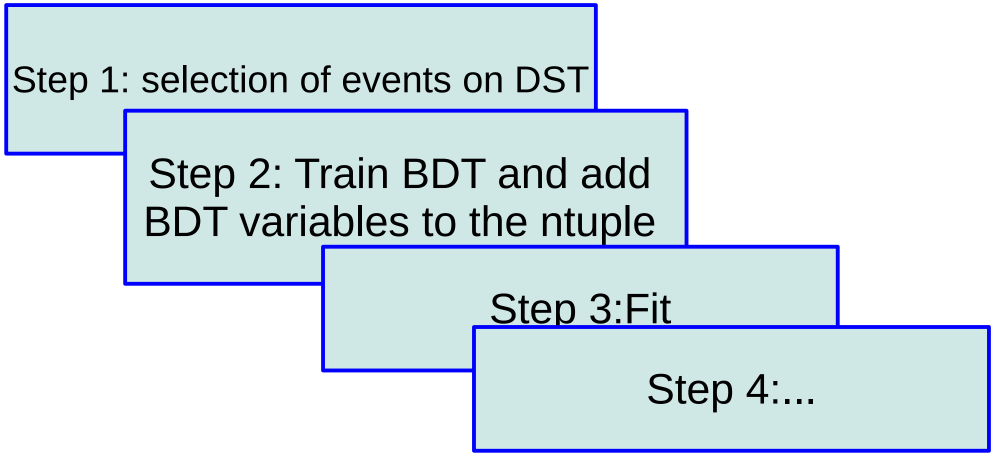
We joined the analysis preservation framework project in 2013.

Main motivation: avoid losses of information on analysis (especially final ntuples, code)

First deposition form implemented last summer and tested on one analysis.

Analysis preservation framework

We proposed a new analysis flow to be implemented in the framework, which is more general and should accommodate most of the analysis.
Each analysis is divided into steps:



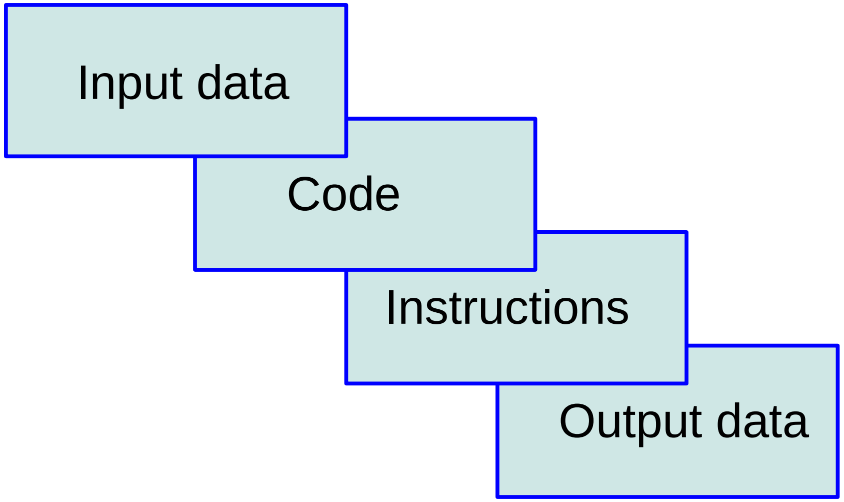
Step 1: selection of events on DST

Step 2: Train BDT and add BDT variables to the ntuple

Step 3:Fit

Step 4:...

... and each step has a defined “ingredients “



Input data

Code

Instructions

Output data

Working on a **self documented ntuple** to save the history of ntuple production.

Feedback from LHCb:

- The problem of analysis preservation (and reproducibility) is being widely discussed within the collaboration.
- LHCb is willing to exploit the DAPF framework to safely archive the published analysis and is considering to make it a requirement for publication.
- For this step it would be useful to have a first production version as soon as possible.

Data and software preservation

Definition of **Run 1 legacy data and software releases**

DATA

- data processed with the latest (legacy) version of the software
- Two copies of raw data and analysis level ntuples, one copy of intermediate format → ~ 12 PB

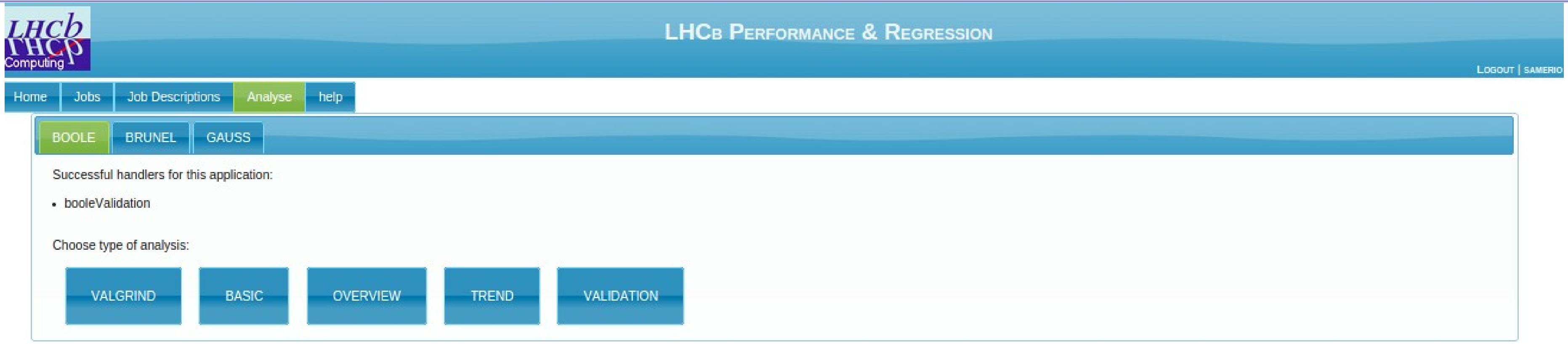
SOFTWARE

- Need to preserve ALL versions of HLT software
- For MC production, need to ensure new generators can be interfaced with legacy reconstruction and ntupling code.
- Documentation
- Validation (see next slide)

Discussion onoing within the collaboration about **non-legacy data** (~ 4 PB)

- Long term future preservation would require a lot of resources , e.g. to keep alive the non-legacy software releases
- Which use cases? Analysis reproducibility?

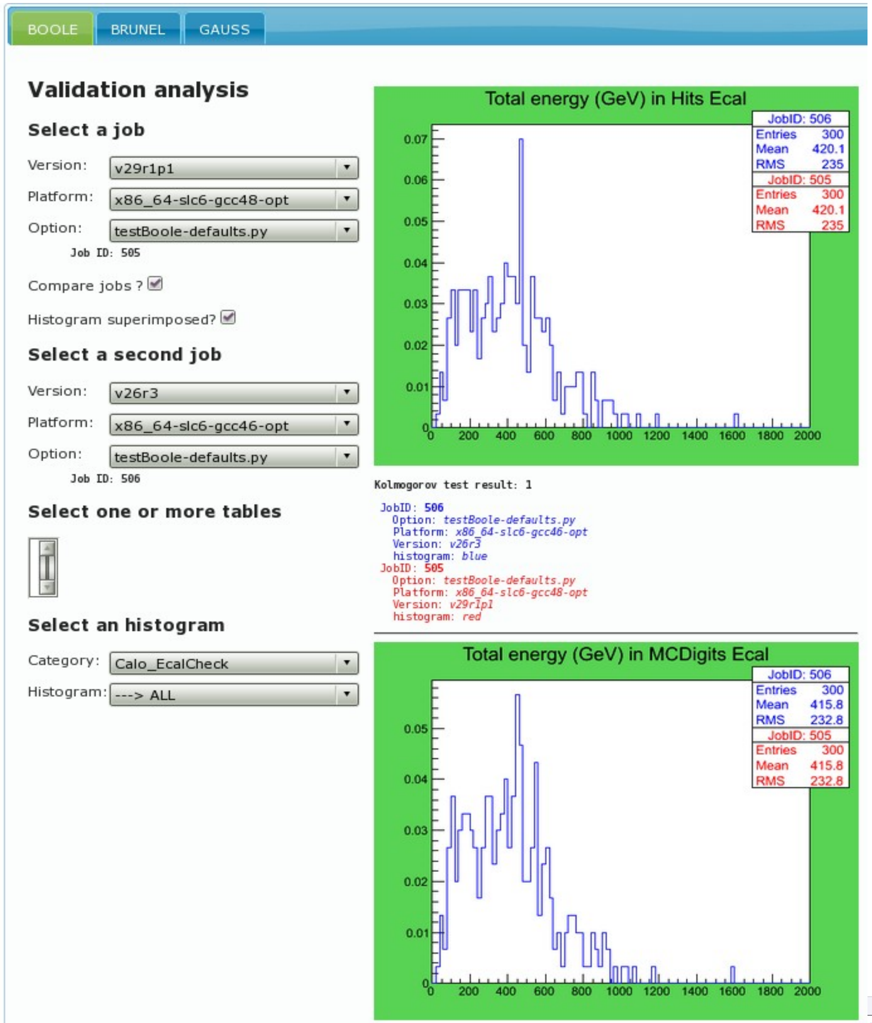
Validation framework



Successful handlers for this application:

- booleValidation

Choose type of analysis:



Validation analysis

Select a job

Version: v29r1p1

Platform: x86_64-slc6-gcc48-opt

Option: testBoole-defaults.py

Job ID: 505

Compare jobs ? ☑

Histogram superimposed? ☑

Select a second job

Version: v26r3

Platform: x86_64-slc6-gcc46-opt

Option: testBoole-defaults.py

Job ID: 506

Select one or more tables

Select an histogram

Category: Calo_EcalCheck

Histogram: ---> ALL

```
Kolmogorov test result: 1

 JobID: 506
   Option: testBoole-defaults.py
   Platform: x86_64-slc6-gcc46-opt
   Version: v26r3
   histogram: blue
 JobID: 505
   Option: testBoole-defaults.py
   Platform: x86_64-slc6-gcc48-opt
   Version: v29r1p1
   histogram: red
```

- Exploit *existing LHCb Performance and Regression framework*
- Define the references (tables and plots) necessary for the validation
- Run job, e.g. to regularly check the legacy release, or increase statistics of a legacy MC sample.
- Compare the physics distributions of new samples with the legacy ones.

Summary

Activities and fruitful collaboration with Cern-IT and other LHC experiments on the Open data portal and the analysis preservation framework → working to enrich the portal with more educational applications and analysis level data. Collaboration with CMS for the WebGL display.

Interest in the analysis preservation framework is increasing in the collaboration → we hope to have soon a first production version to be tested on real analysis.

Working on the Run 1 legacy data and software releases.
Discussion ongoing about non-legacy data.

BACKUP

Legacy data: storage resources

For Run 1 data: **(Sim08)/Reco14/Stripping21**

DATA

	Size (TB)		
	RAW	FULL.DST	DST
Data (2010/11/12/13)	2583	4041	787
MC (Sim08+older)	--	--	794
Total	**2583**	**4041**	**1581**

NB: estimates done with Stripping20. To be conservative, we considered also older Sim0X versions.

If we keep **2 copies of Raw data and DST and only one FULL.DST** --> **12.4 PB** in total

Non-legacy data: storage resources

	Version		ALL.DST, ALL.MDST
			Size (TB)
COLLISION12	Reco13/Stripping18		12.912
	Reco13/Stripping19		0.488
	Reco13a/Stripping19a		1.592
	Reco13c/Stripping19b		1.079
	Reco13e/Stripping19c		0.009
		SUBTOTAL	***16.08***
COLLISION11	Reco09/Stripping13		37.694
	Reco10/Stripping13b		198.064
	Reco11/Stripping15		72.083
	Reco11a/Stripping16		93.969
	Reco12/Stripping17		196.821
	Reco12/Stripping17b		118.102
		SUBTOTAL	***716.733***
COLLISION10	Reco08/Stripping12b		30.477
	Reco08/Stripping12c		10.421
	Reco08/Stripping14		5.393
		SUBTOTAL	***46.291***

	Total older versions		**779.104**

		ALLSTREAMS.DST, DST	Productions
		Size (TB)	
MC11a		596.617	Sim01
		1.531	
	SUBTOTAL	*598.148*	
MC10		201.59	Sim05
		7.725	
	SUBTOTAL	*209.315*	
	TOTAL	**807.463**	

NB: DST only

Single copy:
- 0.8 PB for data
- 0.8 PB for MC
  - **1.6 PB in total**

Two copies:
- 1.6 PB for data
- 1.6 PB for MC
  - **3.2 PB in total**