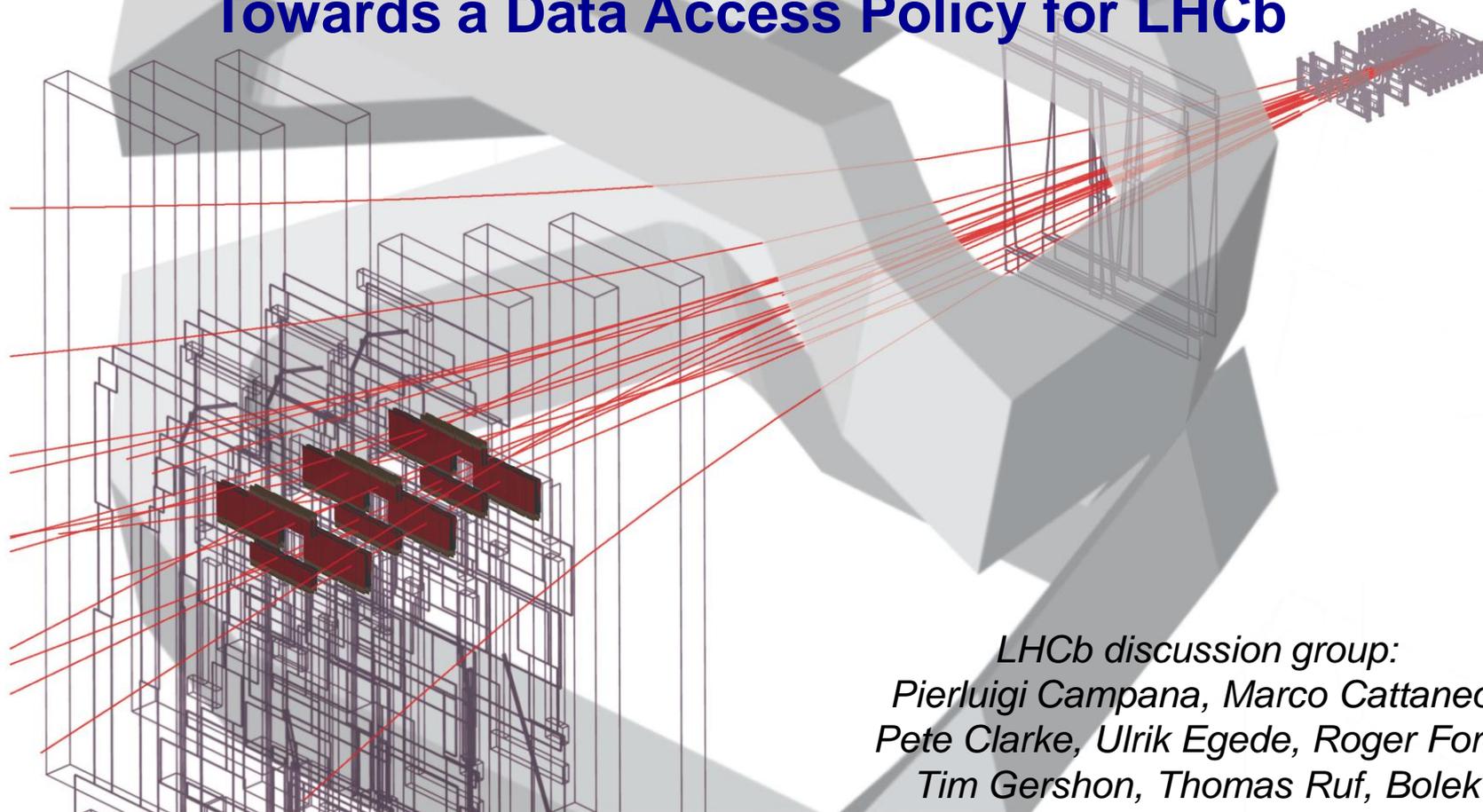


# Data Preservation & Open Data Access

## Towards a Data Access Policy for LHCb



*LHCb discussion group:  
Pierluigi Campana, Marco Cattaneo,  
Pete Clarke, Ulrik Egede, Roger Forty,  
Tim Gershon, Thomas Ruf, Bolek  
Pietrzyk*

# Data Preservation & Open Data Access

- ❑ LHCb has begun to address the related (but different issues) of
  - **Long term preservation of experimental** data for the collaboration itself
  - **Open Data Access** for people external to the collaboration
  
- ❑ A “discussion group” within LHCb has been participating in
  - The CERN+LHCb experiments group
  - DPHEP meetings
  
- ❑ First presentation of the subject to the wider Collaboration in May 2012
  - A draft (strawman) policy was presented

# Long Term Preservation of Data

## □ Principle motivation is internal to collaboration

- We need to ensure that we can access and analyse our own data many years from now

## □ LHCb will endeavour to:

- Soon: task force to identify what needs to be done to develop the capability with respect to our current position
- Later: Identify additional resources\*\* to implement
- This is a technical+physics challenge - it is not just a computing issue

\*\*Important: LHCb is resource limited, and does not at present foresee the people to do this within current resources.

# Open data access

- LHCb recognises the drivers to consider open data access
  - Data re-use by non collaboration members at some point in the future
  - Align with CERN and other LHC experiments
  - Having even a modest policy is MUCH better than not having one at all
  - If we are proactive we can influence funders own policies in a sensible way
  
- We hope to develop ODA capability off the back of our internal data preservation efforts
  
- To this end a draft policy was developed
  - Presented to collaboration in May
  - Now being iterated

## **Draft Policy THIS IS NOT AGREED OR ENDORSED BY THE LHCB CB YET**

1. Data preservation is fundamentally important for the collaboration itself <.....> LHCb will seek to develop such a data preservation capability as soon as practical.
2. LHCb supports the principle of open data access. We can envisage providing some such data access based upon the work needed internally for data preservation (point 1 above). However, as for other modern high-energy physics experiments, the data are complex, and making data available meaningfully requires substantial resources, therefore in the immediate future any provision in this sector will be modest.
3. Overall the collaboration expects to follow the policy developed by CERN and the LHC experiments jointly on these matters, after appropriate approval by the LHCb Collaboration Board.
4. LHCb is resource limited at present, and would therefore welcome the availability of additional resources from funders targeted at this important sector in order to realise aspirations of this document.
5. Access to its data by people outside the collaboration can be considered at four levels of increasing complexity, listed below, with associated conditions.....

## **Draft Policy THIS IS NOT AGREED OR ENDORSED BY THE LHCB CB YET**

### **Level-1. Published results**

All scientific output is published in journals, preliminary results made available in Conference Reports.

All are Open Access, without restriction on use beyond the standard conditions agreed by CERN.

Data associated to the publications will also made available: tables and data from plots (e.g. including likelihood profiles).

### **Level-2. Outreach and Education**

LHCb already participates in outreach activities and will continue to do so. This includes event displays of selected events, ntuple or similar level data for illustrating the calculation of invariant mass distributions, lifetimes, CP asymmetries, etc. Such activity is strongly encouraged.

The data are provided for educational purposes only, and are not considered suitable for publication. Only a limited fraction of the complete LHCb data-set may be used.

## Draft Policy THIS IS NOT AGREED OR ENDORSED BY THE LHCB CB YET

### Level-3. Reconstructed data (i.e. DST)

Here are the general ideas (wording is under discussion)

- Software is already openly available.
- DST Data at CERN may be made available X years after it is taken.  
X  $\geq$  5 but many think this should be longer
- Documentation on an “as is” basis.  
Unfortunately at present there is no resource to dedicate to ODE
- Any publication that results from data analysis by non-members of the collaboration will require a suitable acknowledgement and disclaimer to be included
- No review of such publications will be undertaken by the collaboration.
- Members of the LHCb collaboration are not permitted to sign such papers.

## **Draft Policy THIS IS NOT AGREED OR ENDORSED BY THE LHCB CB YET**

### **Level-4. Raw data**

- It is practically impossible to make the full raw data-set from scientific endeavours of the scale of high-energy physics easily usable in a meaningful way outside of the collaboration.
- It should be noted that access to the raw data is not even permitted to individuals within the collaboration, and that instead the production of reconstructed data (as discussed in 3. above) is performed centrally.
- LHCb will not devote any resources to providing access to the full raw data-set for non-members of the collaboration.
- However, access to representative smaller samples of raw data might be considered if well motivated. This would be case by case.
- LHCb will of course not hinder access demanded by any applicable compliance legislation. We are simply saying that we cannot proactively develop anything for this eventuality.

# Summary

- ❑ Data preservation is upon us – and its common sense we initiate at least a requirements study – and preferably identify some effort to take it forward.
- ❑ LHCb is participating in the CERN+LHC wide discussions + DPHEP
- ❑ Open data access is out there – it may not be imminent in terms of requests – but it has the weight of funders and legislation
- ❑ It is far better to have a policy than not have one.
- ❑ We have presented a draft policy to focus the next step in iteration of the LHCb position.