Our Digital Future

Multidisciplinary Perspectives on Long Term Data Preservation and Access

Data Preservation project at the LHCb experiment at CERN

ANA TRISOVIC

TRISOVIC@HEP.PHY.CAM.AC.UK

15 MARCH 2016

CAMBRIDGE

Agenda

LHCb experiment at CERN

Data preservation project and motivation

CERN Analysis Preservation portal

LHCb data and software

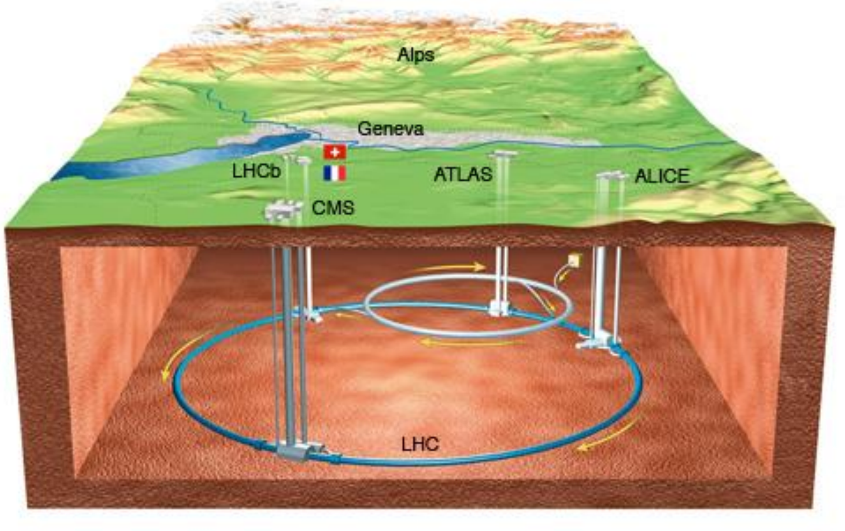
Data dependency database

LHCb experiment at CERN

One of the four particle detectors at the Large Hadron Collider at CERN

Investigate asymmetry between matter and antimatter

840 people from 60 scientific institutes



Data Preservation project

The experimental & simulated data

Software & documentation

Analysis & publications

Motivation for Data Preservation

1. Research reproducibility – reanalysing data in search of new signals or to improve current measurements.
2. Scientific potential – Analyse old data to search for a signal predicted by a new theory.
3. Social reasons – CERN is funded by the world community; therefore data should be preserved and made available to general public.

Different approach for short term and long term data preservation

CERN ANALYSIS PRESERVATION

and

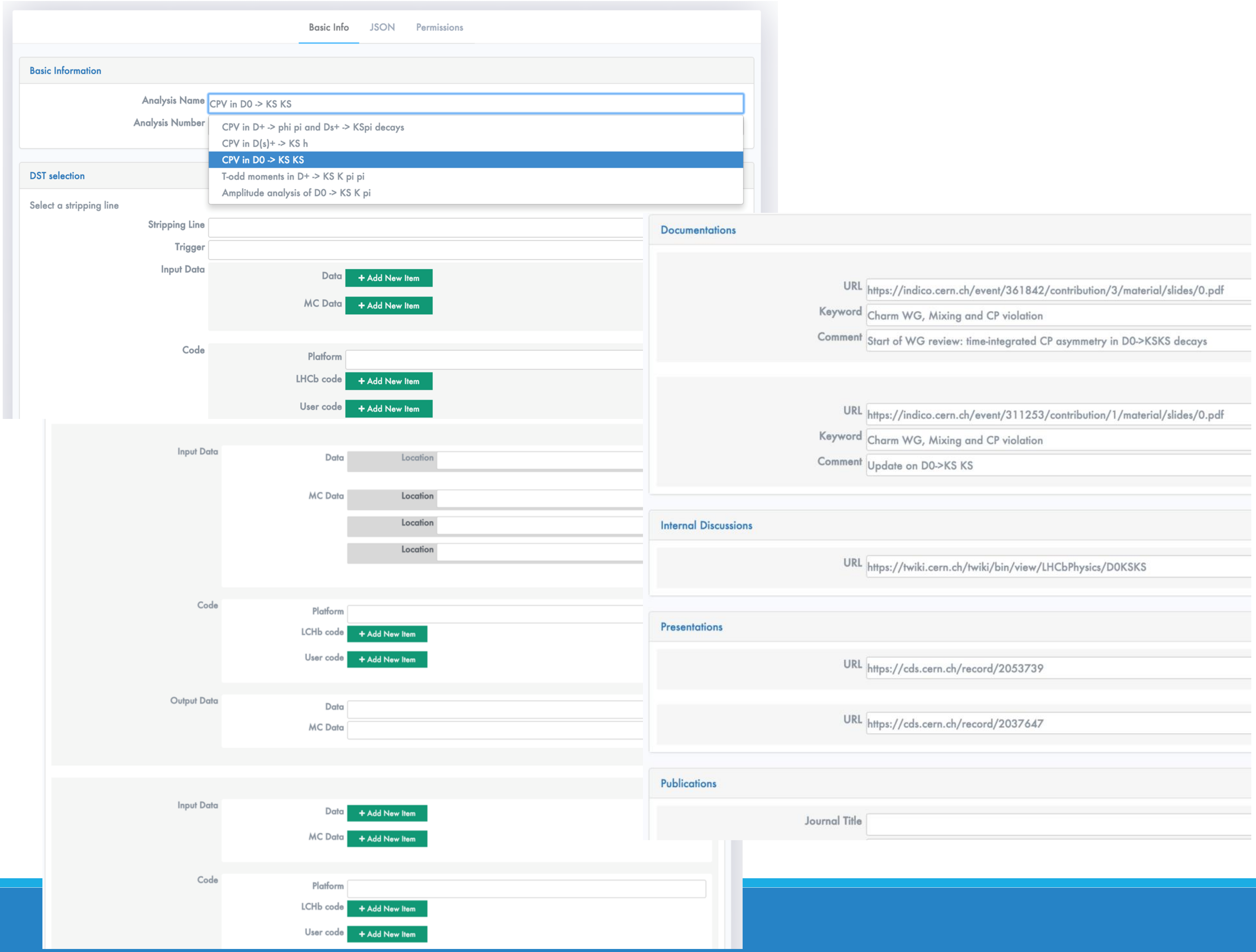
opendata CERN

CERN ANALYSIS PRESERVATION portal

Web portal for physicists, fellows, interns, doctoral and summer students to log information about their analyses

- Input data, n-tuples, code, papers, other publication, peer review and Q&A

Team: Sünje Dallmeier-Tiessen, Anxhela Dani, Tibor Sinko, Javier D. Fernandez, Pamfilos Fokianos, Patricia S. Herterich



Basic Information

Analysis Name: CPV in D0 -> KS KS

Analysis Number:

- CPV in D+ -> phi pi and Ds+ -> KSpi decays
- CPV in D(s)+ -> KS h
- CPV in D0 -> KS KS
- T-odd moments in D+ -> KS K pi pi
- Amplitude analysis of D0 -> KS K pi

DST selection

Select a stripping line

Stripping Line:

Trigger:

Input Data
- Data
- MC Data

Code
- Platform:
- LHCb code
- User code

Input Data
- Data: Location
- MC Data: Location
- Location
- Location

Code
- Platform:
- LCHb code
- User code

Output Data
- Data:
- MC Data:

Input Data
- Data
- MC Data

Code
- Platform:
- LCHb code
- User code

Documentations

URL: https://indico.cern.ch/event/361842/contribution/3/material/slides/0.pdf

Keyword: Charm WG, Mixing and CP violation

Comment: Start of WG review: time-integrated CP asymmetry in D0->KSKS decays

URL: https://indico.cern.ch/event/311253/contribution/1/material/slides/0.pdf

Keyword: Charm WG, Mixing and CP violation

Comment: Update on D0->KS KS

Internal Discussions

URL: https://twiki.cern.ch/twiki/bin/view/LHCbPhysics/D0KSKS

Presentations

URL: https://cds.cern.ch/record/2053739

URL: https://cds.cern.ch/record/2037647

Publications

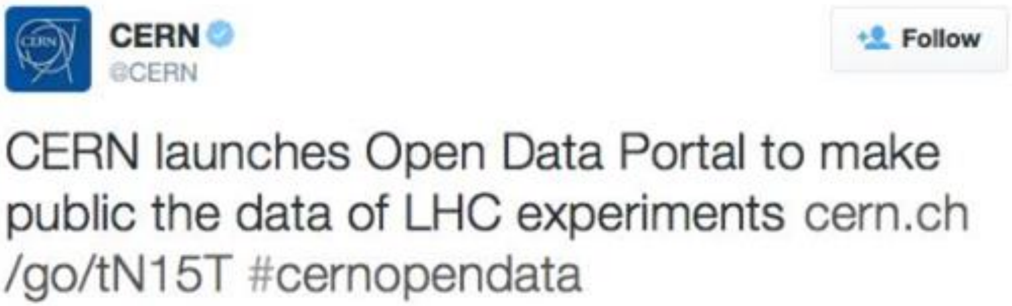
Journal Title:

CERN Open data portal

Access point to data produced by the research and experiments conducted at CERN

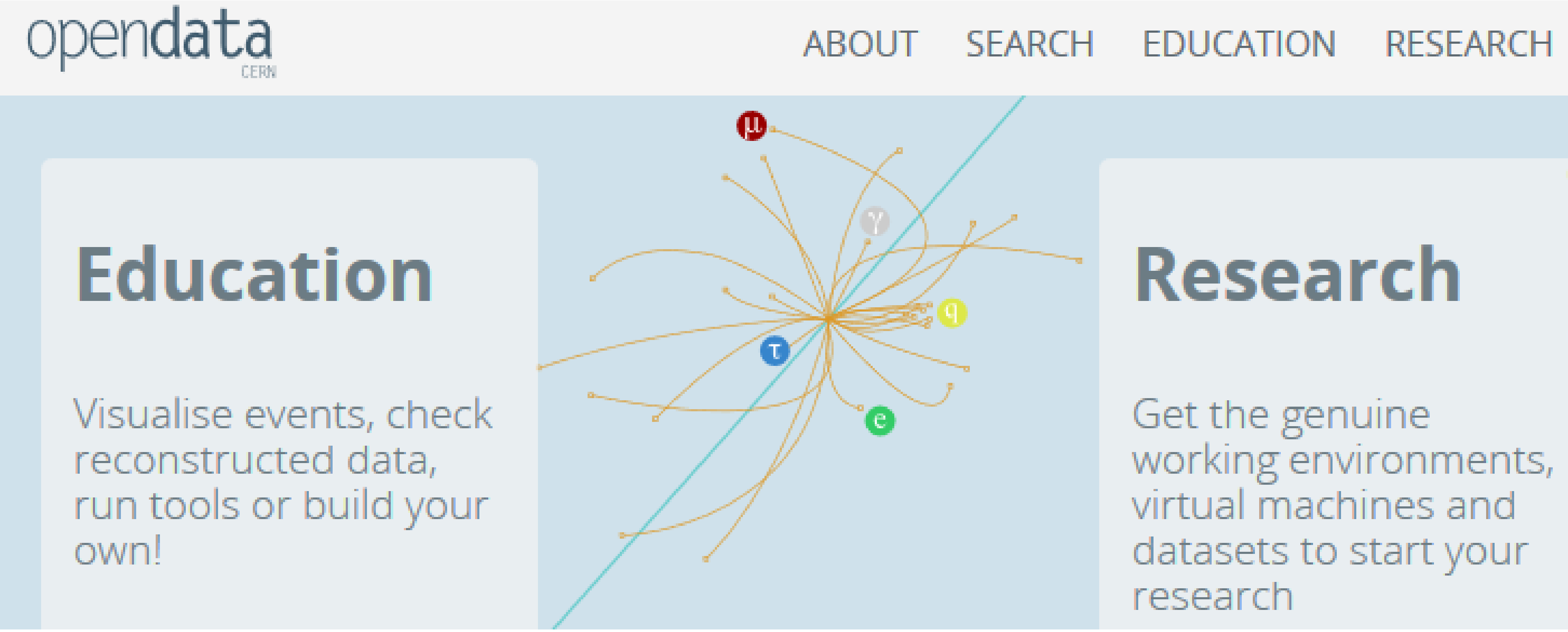
Provides the data, software and documentation

Data is going to be preserved if it is available online and used by scientists worldwide



CERN launches Open Data Portal to make public the data of LHC experiments cern.ch /go/tN15T #cernopendata

CERN Open data portal



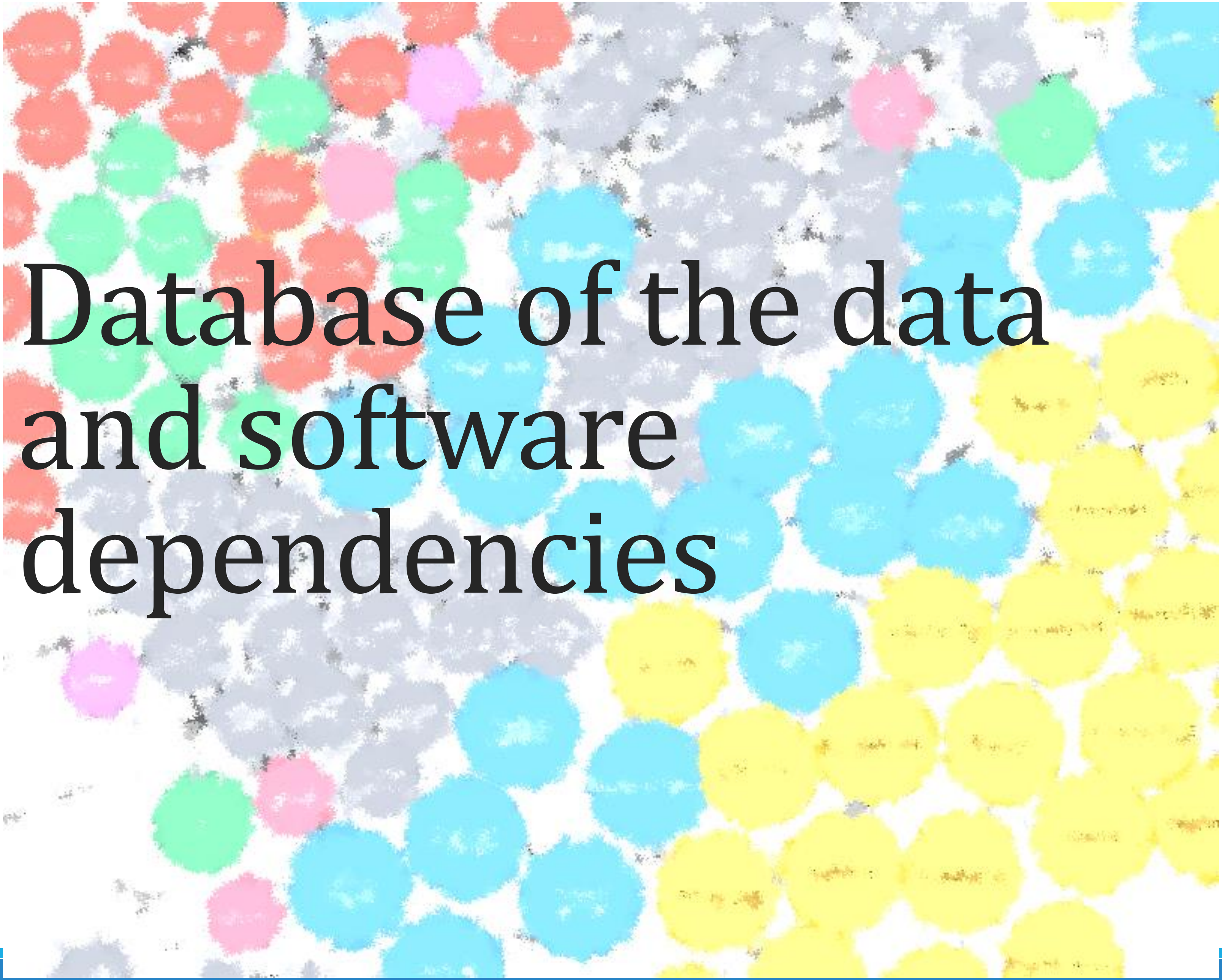
ABOUT SEARCH EDUCATION RESEARCH

Education

Visualise events, check reconstructed data, run tools or build your own!

Research

Get the genuine working environments, virtual machines and datasets to start your research

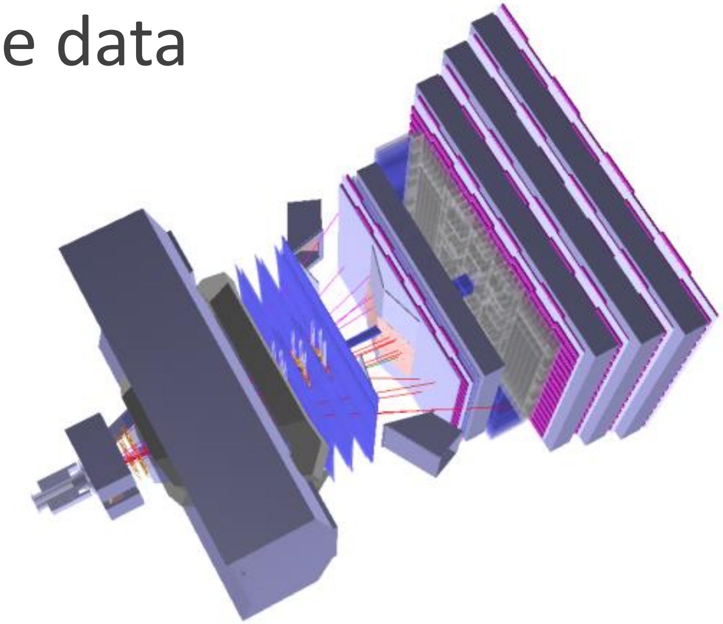


Database of the data and software dependencies

Experimental data

During the run, there are 40 million collisions per second

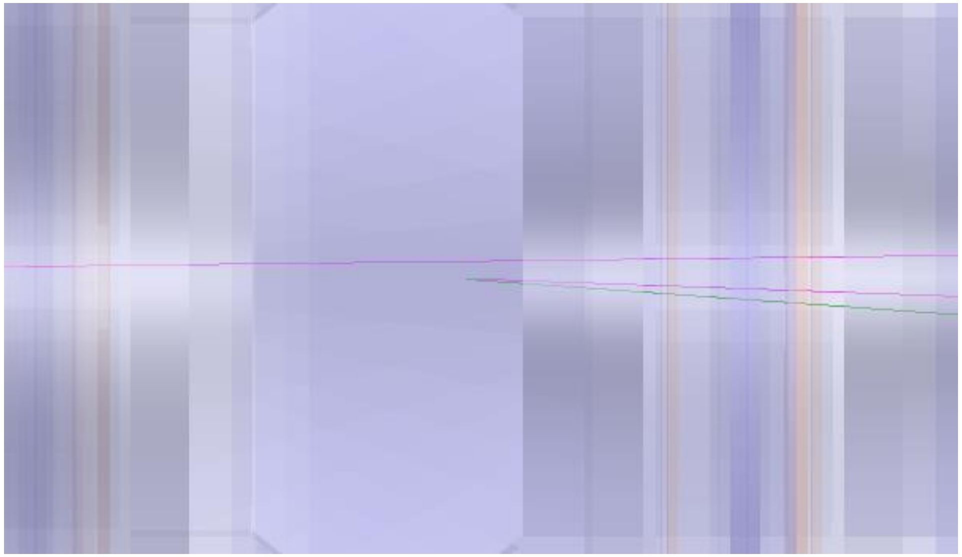
The mechanism called the trigger identifies interesting events and saves them, discarding the other 99.9% of the data

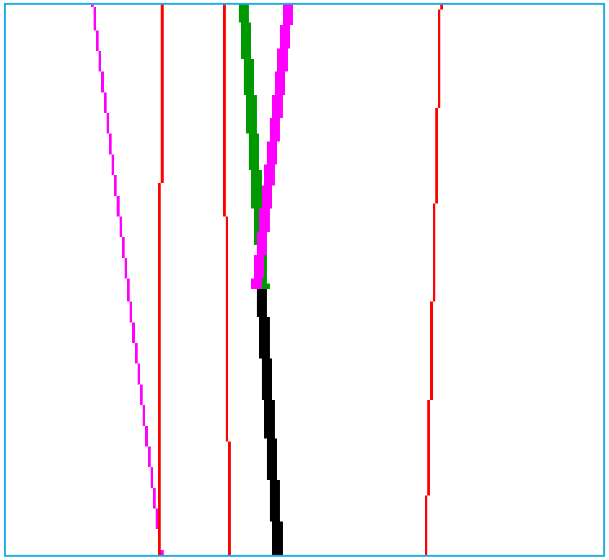


Experimental data

Elementary particles collide creating unstable particles that decay quickly

Necessary to reconstruct an “image” of the event





Simulated data

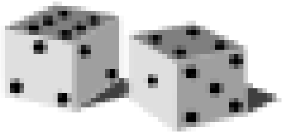
Simulation mimic what happens in the LHCb detector

Comparing the simulated with the real data helps us interpret the results

The volume of the simulated data is bigger than the real data

LHCb data management

Flow of the real and simulated data:



Reconstruction

- From raw data format to readable data
- Heavy software for processing

Data reduction

- Producing streams corresponding to activity of the working groups

User analysis

http://cds.cern.ch/record/1460575

Size of the LHCb data

What do we save?

- O(10) PB – raw data
- O(100) TB – processed data
- O(1) TB – users’ data

LHCb software

Gaudi framework provides interfaces and services for event data processing applications

- ***DaVinci*** application – Particles manipulation and measuring physics processes
- ***Brunel*** application – Event reconstruction: particle tracks, particle IDs
- Etc.

Data are compatible with different software versions

Data and software dependences

The database with:

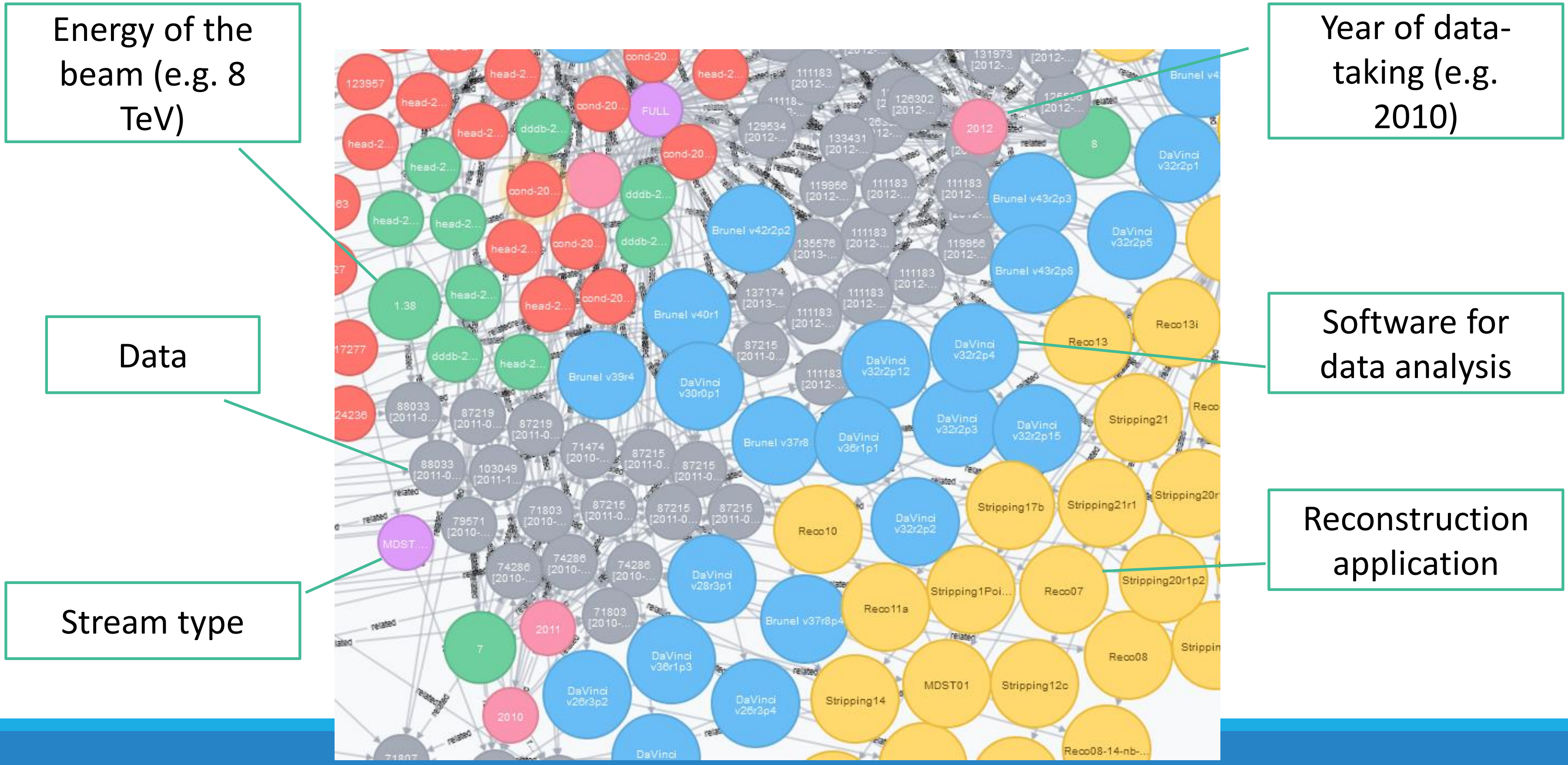
- information about the software, the versions released and their relationships (e.g. what do I need if I want to run DaVinci X)
- information about the data lifecycle, from primitive data files to processed data and their compatibility with the LHCb software

Use cases

1. Short term future: Software needed to analyse the data from 2012
2. Automatically determining tests that have to be run to guarantee we can still (re)analyse the data
3. Identifying legacy software versions
4. CERN Analysis Preservation (CAP) portal
5. LHCb web pages

Implementation

Implemented in the Graph database Neo4j



Energy of the beam (e.g. 8 TeV)

Data

Stream type

Year of data-taking (e.g. 2010)

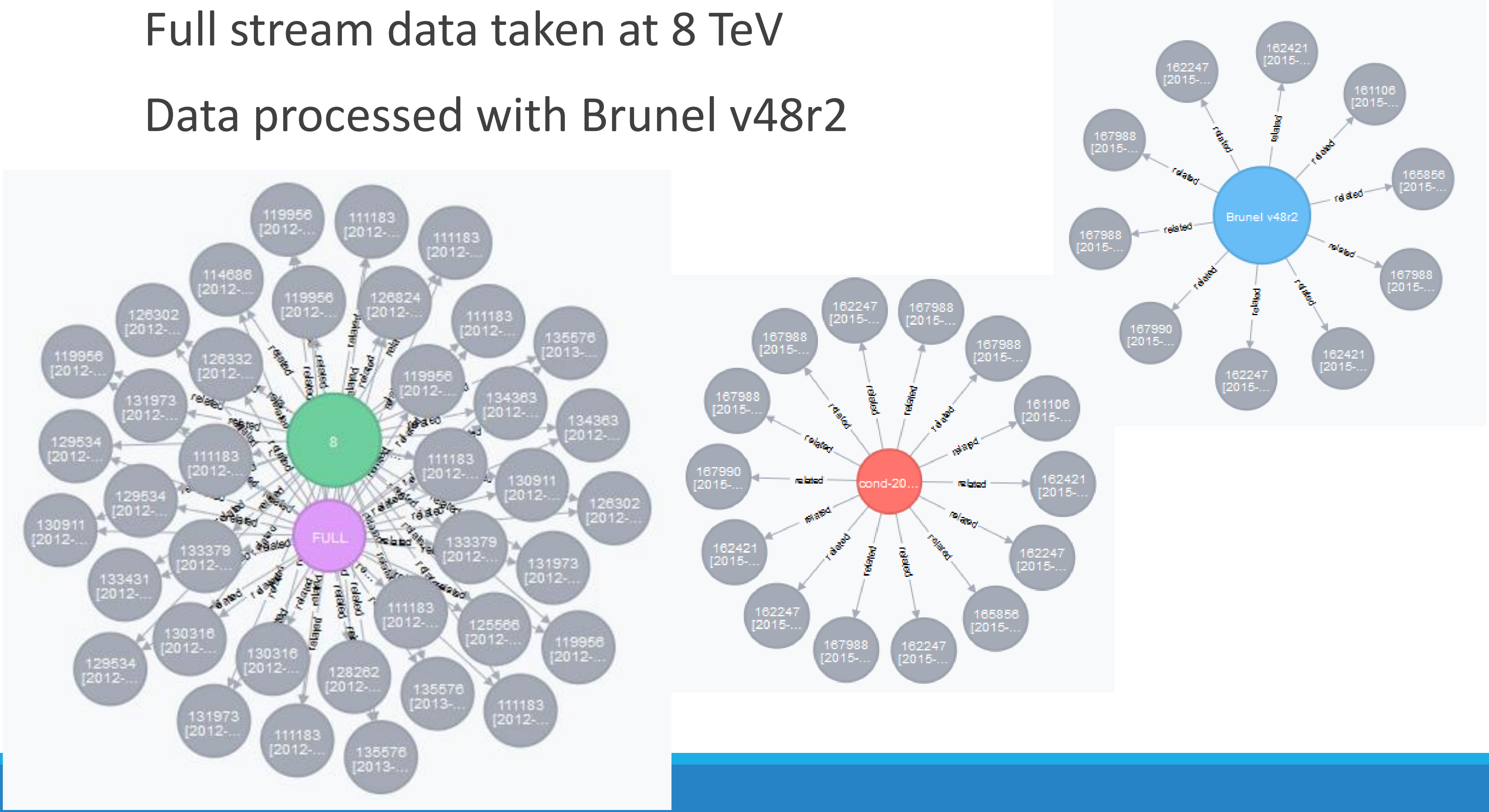
Software for data analysis

Reconstruction application

Examples: easy lookup for the data

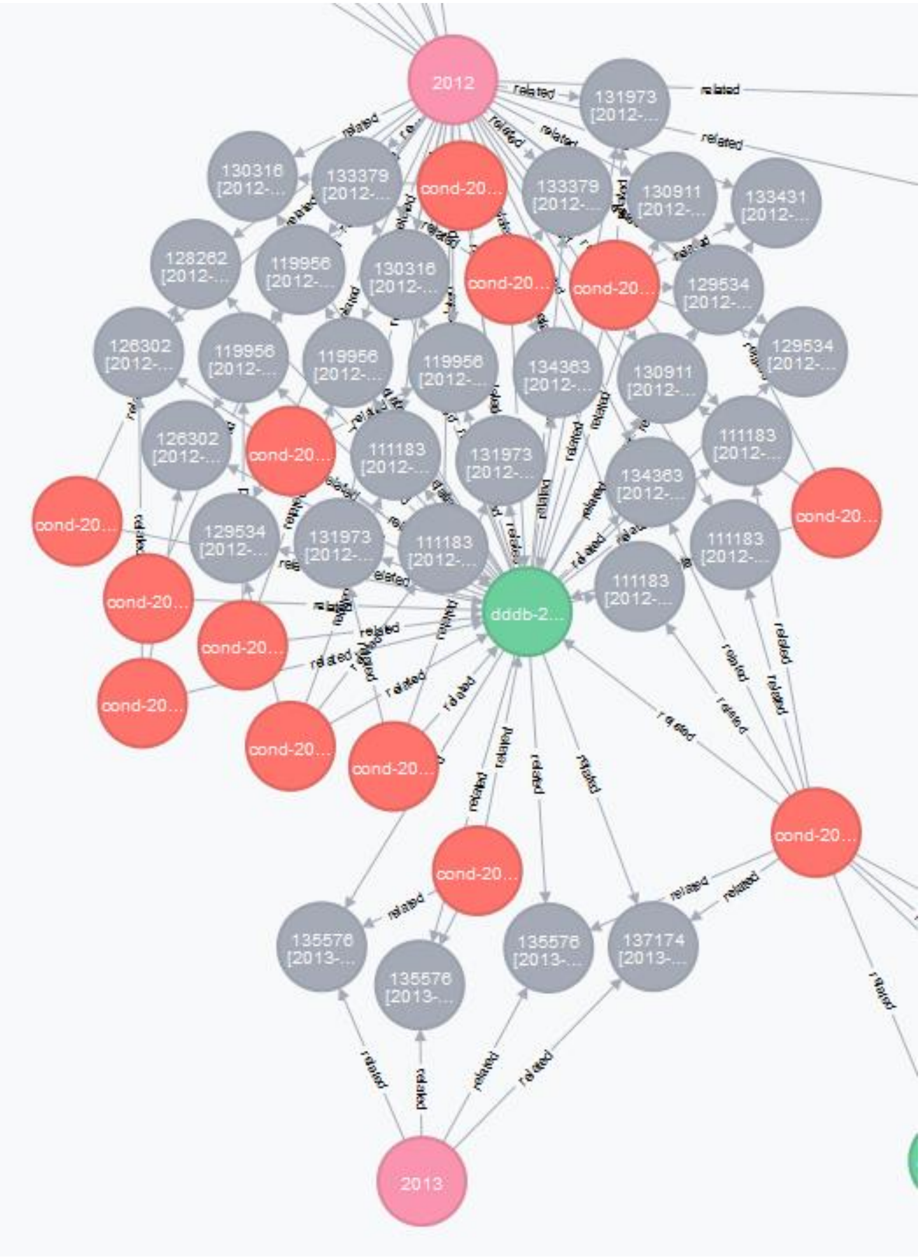
Full stream data taken at 8 TeV

Data processed with Brunel v48r2



Examples:

Data with particular detector conditions



Thank you for your attention!