The LHCb NTuple Wizard

Sebastian Neubert[1]

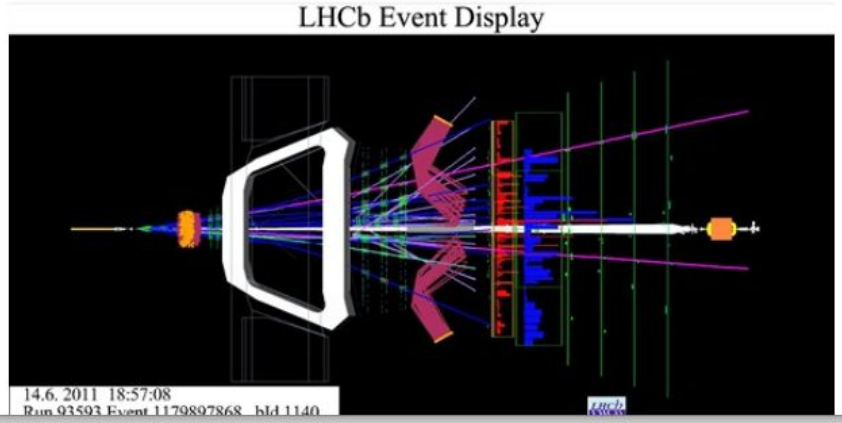
[1]HISKP Bonn

Mini Workshop Best Practices in
Model preservation
14.12.23

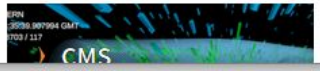
LHCb releases first set of data to the public

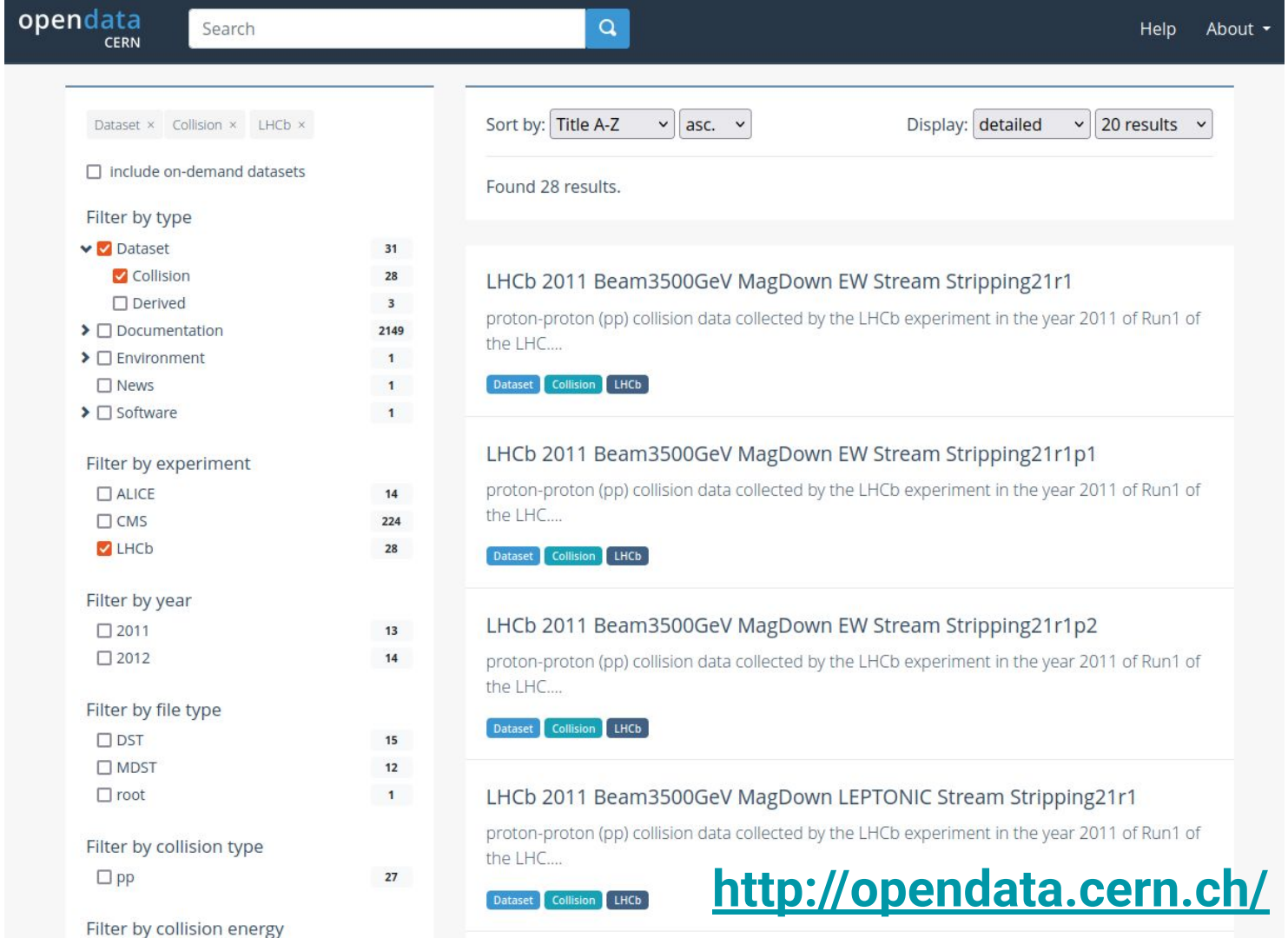
The LHCb collaboration has released data from Run 1 of the LHC to the public for the first time, allowing research to be conducted by anyone in the world

8 DECEMBER, 2022 | By LHCb collaboration



Related Articles





http://opendata.cern.ch/

DPHEP Levels of Data Complexity

[https://arxiv.org/abs/1205.4667]

1. **Published results + additional information**
   - supplemental data tables, ntuples
   - HEPData entries, rivet plugins
   - notes, technical information
   - documentation, slides
   - analysis code, jupyter notebooks

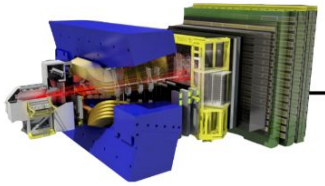
2. **Education and Outreach**
   - simplified data formats, e.g. highly preprocessed ntuples

3. **Reconstructed data + analysis level software**
   - Calibrated reconstructed data with the level of detail useful for algorithmic, performance and physics studies
   - preservation of analysis level experiment-specific software

4. **Raw data + reconstruction software**
   - **Not released for LHC data**

LHCb Level 3 Data

Release policy: 50% @ 5yrs, 100% @ 10yrs
after end of running period



- Level 3 data in LHCb **defined as the output of the stripping**
- Same level of abstraction accessed by LHCb members
- Organized in ~10 streams, according to physics signature
- Software needed to access data (DaVinci) is open source, available via CVMFS (or container)
- Documentation: LHCb Starterkit openly available

Data to be released very soon: full 2011/12 Stripping Output

Mindaugas Sarpis, Bonn
UNIVERSITÄT BONN

BHADRON.MDST
BHADRONCOMPLETEEVENT.DST
CHARM.MDST
CHARMCOMPLETEEVENT.DST
DIMUON.DST
EW.DST
LEPTONIC.MDST
RADIATIVE.DST
SEMILEPTONIC.DST

Already released
~ 200TB

Ready to be released
(preparing press announcement)
~ 800TB

All this data comes with meta-data and documentation.
The data we are releasing now contains almost 7500 Pre-Selections (stripping lines)

- **This includes common particles used to build more complex decays**
- **DST files contain the full events - data mining possible**
- **MDST files only contain the particles selected by the stripping lines (rest of event is discarded)**

But it still is in DST/MDST Format and requires the LHCb Software to read.

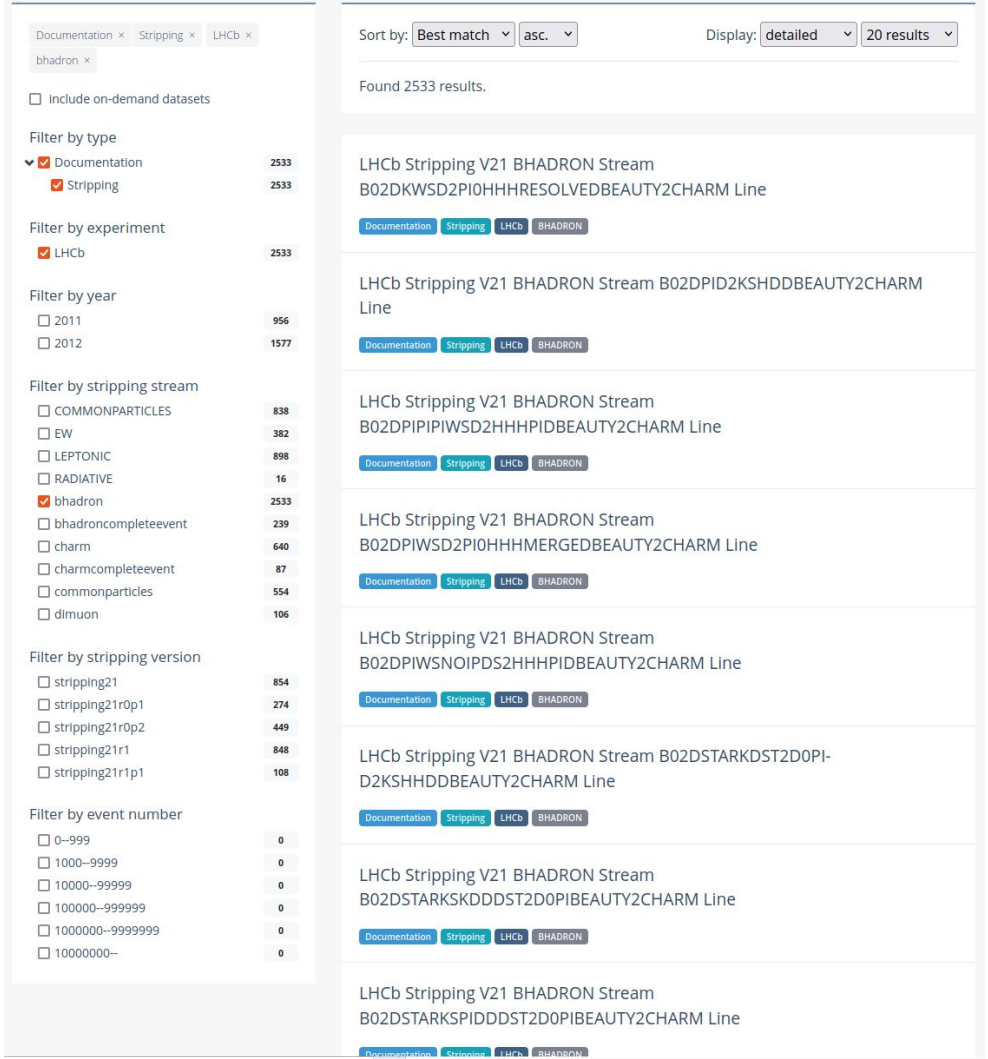
LHCb Preselections

All the info is there but you need to know how to look for it. E.G,:

The decay B^0 -> D pi
with D -> K_s h
where the K_s is reconstructed from two downstream tracks

is selected by the

B02DPiD2KSHDDBeauty2CharmLine



Typical LHCb Selection “Stripping” documentation

LHCb Stripping V21 BHADRON Stream B02DPIPIPIWSD2HHHPIDBEAUTY2CHARM Line

Documentation Stripping

[stripping21 lines]

StrippingB02DPiPiPiWSD2HHHPIDBeauty2CharmLine

Properties:

OutputLocation	Phys/B02DPiPiPiWSD2HHHPIDBeauty2CharmLine/Particles
Postscale	1.0000000
HLT	(HLT_PASS_RE('Hlt2Topo.*Decision') | HLT_PASS_RE('Hlt2IncPhi.*Decision'))
Prescale	1.0000000
L0DU	None
ODIN	None

Filter sequence:

LoKi::VoidFilter/StrippingB02DPiPiPiWSD2HHHPIDBeauty2CharmLineVOIDFilter

Code	(recSummaryTrack(LHCb.RecSummary.nLongTracks, TrLONG) \< 500)

CheckPV/checkPVmin1

MinPVs	1
MaxPVs	-1

GaudiSequencer/SeqD2HHHBeauty2Charm

GaudiSequencer/SEQ:D+2HHHBeauty2CharmFilter

LoKi::VoidFilter/SelFilterPhys_StdAllNoPIDsPions_Particles

Code	CONTAINS('Phys/StdAllNoPIDsPions/Particles')>0

FilterDesktop/PiInputsBeauty2CharmFilter

Code	(TRCHI2DOF\<3.0) & (PT>100*MeV) & (P>1000*MeV) & (MIPCHI2DV(PRIMARY)>4.0) & (TRGHP\<0.4)
Inputs	['Phys/StdAllNoPIDsPions']
DecayDescriptor	None
Output	Phys/PiInputsBeauty2CharmFilter/Particles

CombineParticles/ProtoD+2HHHBeauty2Charm

Inputs	['Phys/PiInputsBeauty2CharmFilter']
DaughtersCuts	
CombinationCut	(ASUM(PT)>1800*MeV) & (in_range(1769.62*MeV,AWM('pi+','pi+','pi-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('pi+','pi+','K-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('K+','pi+','pi-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('K+','pi+','K-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('pi+','K+','pi-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('K+','K+','pi-'),2068.49*MeV)) & (AHASCHILD((ISBASIC & HASTRACK & (TRCHI2DOF\<2.5) & (PT > 500*MeV) & (P > 5000*MeV))|((ABSID=='KS0') & (PT > 500*MeV) & (P > 5000*MeV) & (BPVVDCHI2 >

DecayDescriptor	None
Output	Phys/PiInputsBeauty2CharmFilter/Particles

CombineParticles/ProtoD+2HHHBeauty2Charm

Inputs	['Phys/PiInputsBeauty2CharmFilter']
DaughtersCuts	
CombinationCut	(ASUM(PT)>1800*MeV) & (in_range(1769.62*MeV,AWM('pi+','pi+','pi-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('pi+','pi+','K-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('K+','pi+','pi-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('K+','pi+','K-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('pi+','K+','pi-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('K+','K+','pi-'),2068.49*MeV)) & (AHASCHILD((ISBASIC & HASTRACK & (TRCHI2DOF\<2.5) & (PT > 500*MeV) & (P > 5000*MeV))|((ABSID=='KS0') & (PT > 500*MeV) & (P > 5000*MeV) & (BPVVDCHI2 > 1000)))) & (ACUTDOCA(0.5*mm,'LoKi::DistanceCalculator'))
MotherCut	(VFASPF(VCHI2/VDOF)\<10) & (BPVVDCHI2>36) & (BPVDIRA>0)
DecayDescriptor	None
DecayDescriptors	['D+ -> pi+ pi+ pi-']
Output	Phys/ProtoD+2HHHBeauty2Charm/Particles

SubPIDMMFilter/D+2HHHSubPIDSelBeauty2Charm

Code	ALL
Inputs	['Phys/ProtoD+2HHHBeauty2Charm']
DecayDescriptor	None
Output	Phys/D+2HHHSubPIDSelBeauty2Charm/Particles
MaxMM	2068.4900
MinMM	1769.6200
PIDs	[['pi+' , 'pi+' , 'pi-'] , ['pi+' , 'pi+' , 'K-'] , ['K+' , 'pi+' , 'pi-'] , ['K+' , 'pi+' , 'K-'] , ['pi+' , 'K+' , 'pi-'] , ['pi+' , 'K+' , 'K-'] , ['K+' , 'K+' , 'pi-']]

FilterDesktop/D+2HHHBeauty2CharmFilter

Code	in_range(1769.62,MM,2068.49)
Inputs	['Phys/D+2HHHSubPIDSelBeauty2Charm']
DecayDescriptor	None
Output	Phys/D+2HHHBeauty2CharmFilter/Particles
ModeOR	False
IgnoreFilterPassed	False

GaudiSequencer/SEQ:D-2HHHBeauty2CharmFilter

LoKi::VoidFilter/SelFilterPhys_StdAllNoPIDsPions_Particles

Code	CONTAINS('Phys/StdAllNoPIDsPions/Particles')>0

FilterDesktop/PiInputsBeauty2CharmFilter

Code	(TRCHI2DOF\<3.0) & (PT>100*MeV) & (P>1000*MeV) & (MIPCHI2DV(PRIMARY)>4.0) & (TRGHP\<0.4)
Inputs	['Phys/StdAllNoPIDsPions']
DecayDescriptor	None
Output	Phys/PiInputsBeauty2CharmFilter/Particles

CombineParticles/ProtoD-2HHHBeauty2Charm

Inputs	['Phys/PiInputsBeauty2CharmFilter']
DaughtersCuts	
CombinationCut	(ASUM(PT)>1800*MeV) & (in_range(1769.62*MeV,AWM('pi+','pi+','pi-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('pi+','pi+','K-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('K+','pi+','pi-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('K+','pi+','K-'),2068.49*MeV)|in_range(1769.62*MeV,AWM('pi+','K+','K-'),

Level 3 Data - Resource Projections

2020 projections

	ALICE	ATLAS	CMS	LHCb
Run 2	2 PB	0.5 PB	2 PB	10 PB (including Run 1)
Run 3	4 PB	1 PB	4 PB	45 PB
Total	**6 PB**	**1.5 PB**	**6 PB**	**55 PB**

Mitigation Strategies:

- Provide protected access to existing copies of stripping/turbo output via WG-production slots. Needs “ntupling wizard”
- **Provide direct access to data on grid storage**

FAIR Data Principles

[The FAIR Guiding Principles for scientific data management and stewardship. Nature *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18]

Findable: Metadata and data should be easy to find for both humans and computers.

Accessible: The exact conditions under which the data is accessible should be provided in such a way that humans and machines can understand them.

Interoperable: The (meta)data should be based on standardized vocabularies, ontologies, thesauri etc. so that it integrates with existing applications or workflows.

Reusable: Metadata and data should be well-described so that they can be replicated and/or combined in different research settings.

https://go-fair.org

Solved by
https://opendata.cern.ch

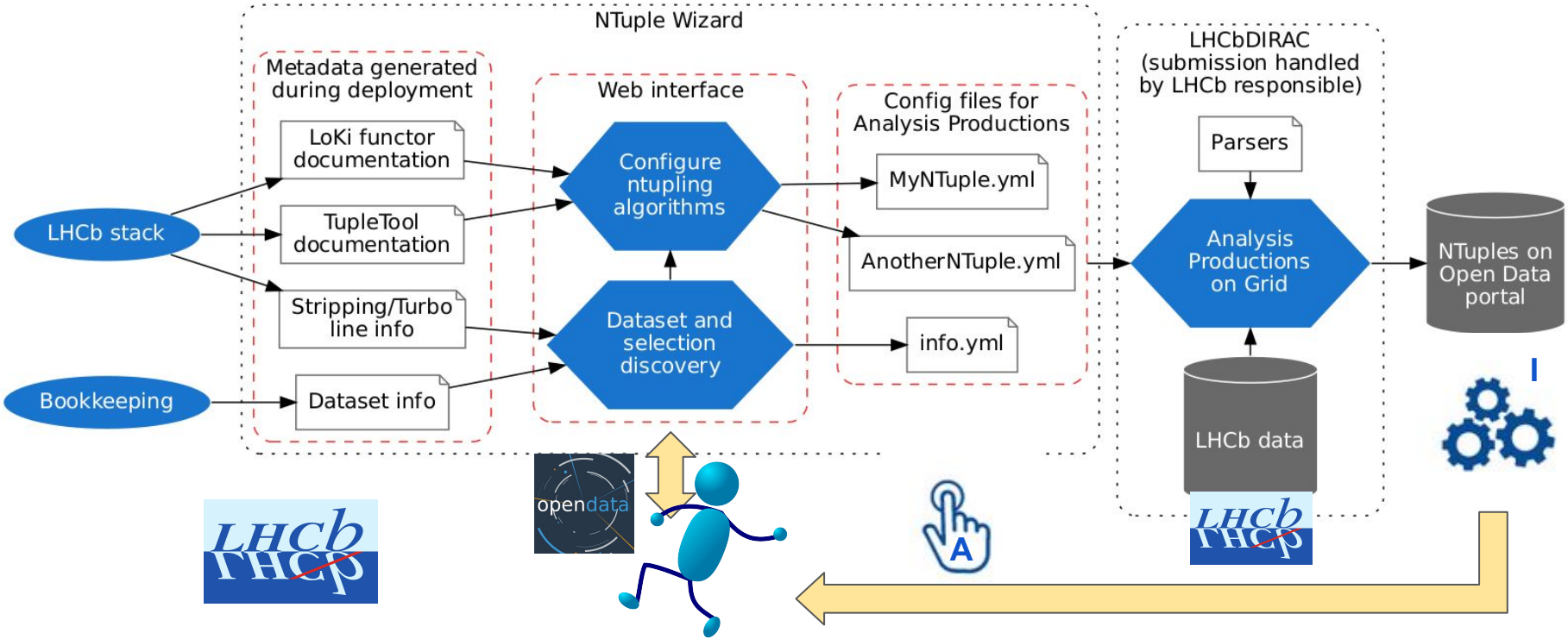
needs continuous curation by the collaborations

Needs dedicated work by the experimental collaborations (link to NFDI, PUNCH4NFDI)

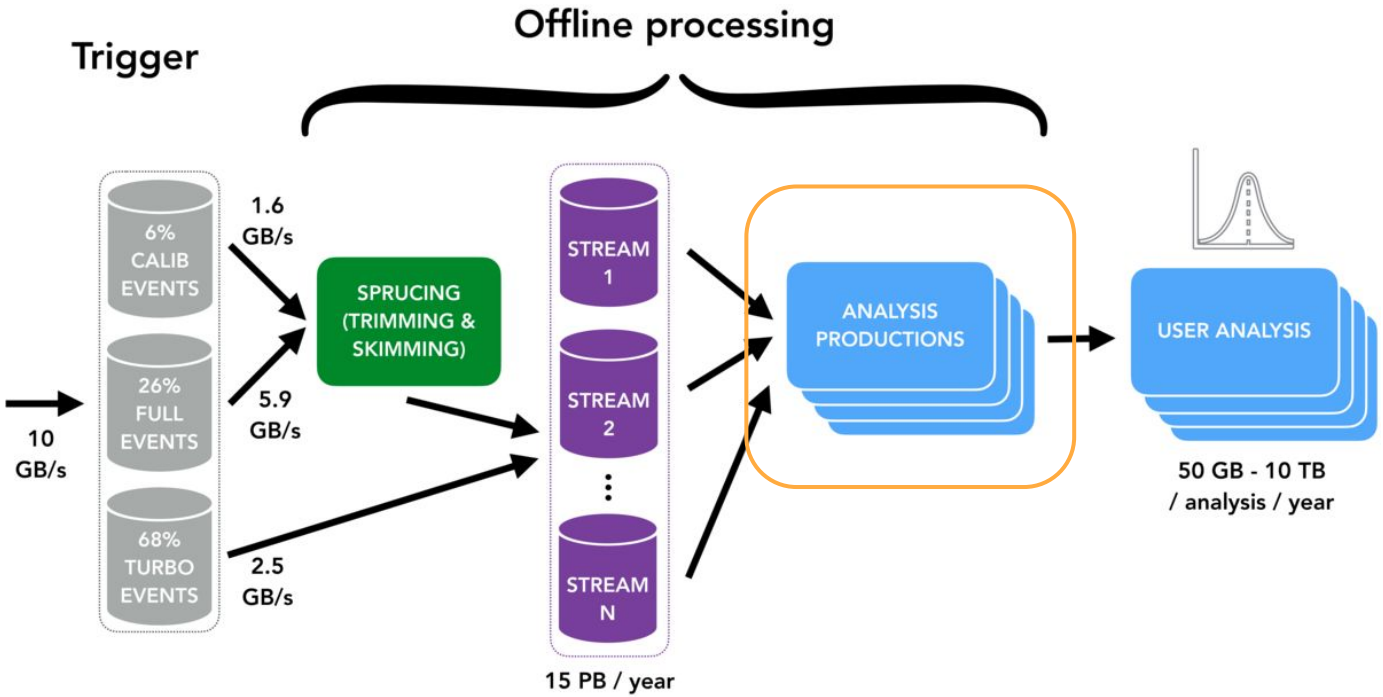
NtupleWizard overview

[Comput Softw Big Sci 7,(2023) 6]

- Can be used to create AP job scripts for LHCb-internal analysts



Analysis productions in LHCb



Goals:
Allow analysts to creatively solve their analysis.

Flexible choice of tools and methods.

Preserve ingredients needed for **interpretation of data**

centrally managed and preserved | managed by proponents / PWG

data preparation | data interpretation

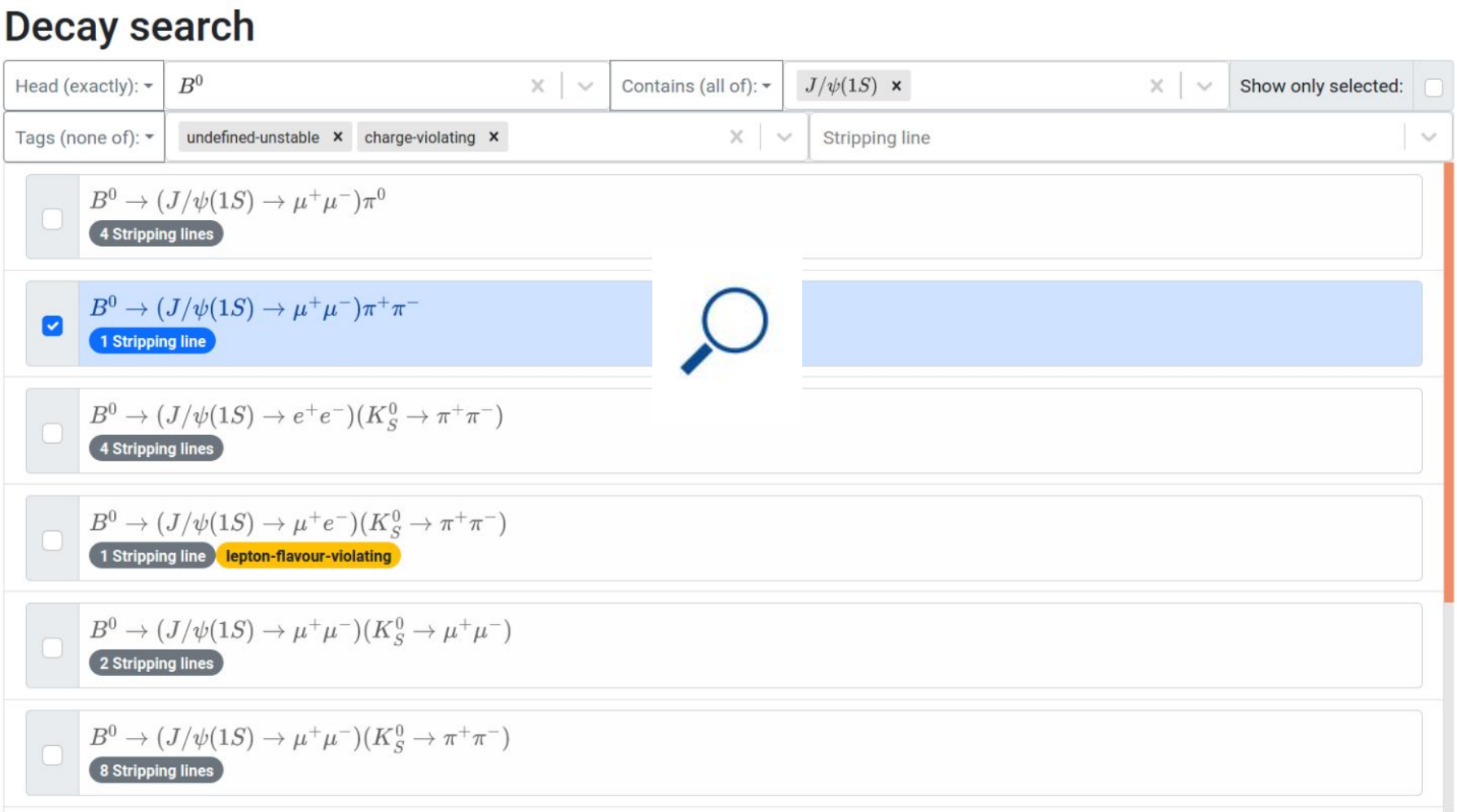


Fig. 3 Example of the decay candidate search function of the Ntuple Wizard.

NTuple Wizard

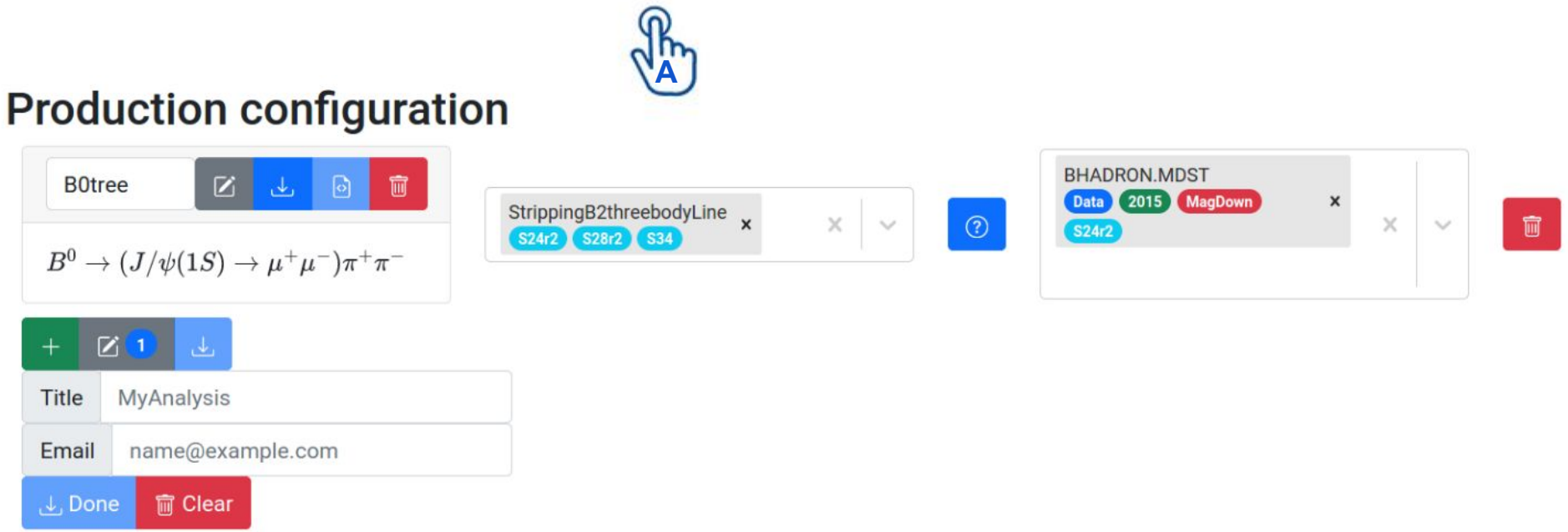
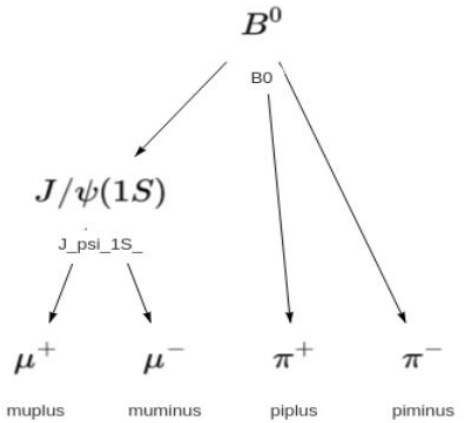


Fig. 4 Example of the data set selection and production configuration step of the Ntuple Wizard.

Configure $B^0 \to (J/\psi(1S) \to \mu^+\mu^-)\pi^+\pi^-$



Select by category

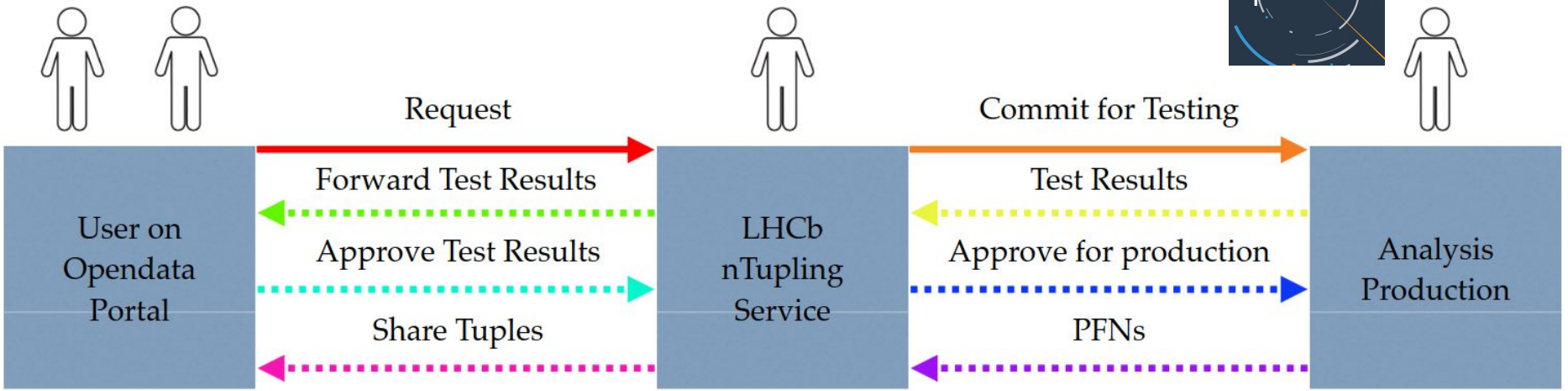
Hadron | Meson | Lepton | X0 | X+ | X- | Down | Beauty | Charm | Up | LongLived | ShortLived | Stable | StableCharged | Scalar | Vector | Spinor

Current selection: $B^0 \to (J/\psi(1S) \to \mu^+\mu^-)\pi^+\pi^-$

5 TupleTools
TupleToolANNPID
TupleToolEventInfo
TupleToolGeometry
TupleToolKinematic
TupleToolPid

Wizard Integration with OD Portal - WIP

Output: Plain Root NTuples



- NTuples will be cleaned up after 30* days
- Can be promoted to permanent records with DOI

- Requests are processed on LHCb internal analysis productions system on the grid

Improving documentation of the NTuples

Problem:

- Unclear what quantities are contained within the NTuple
- Entries in the NTuple can be quite cryptic and variable naming is customizable in many cases
- Relations between quantities are lost, e.g. Decay Mother -> Daughter relation

Interoperable: The (meta)data should be based on standardized vocabularies, ontologies, thesauri etc. so that it integrates with existing applications or workflows.

Reusable: Metadata and data should be well-described so that they can be replicated and/or combined in different research settings.

Well known problems, also in other fields of data science!

"Semantic Web"

Knowledge Graphs

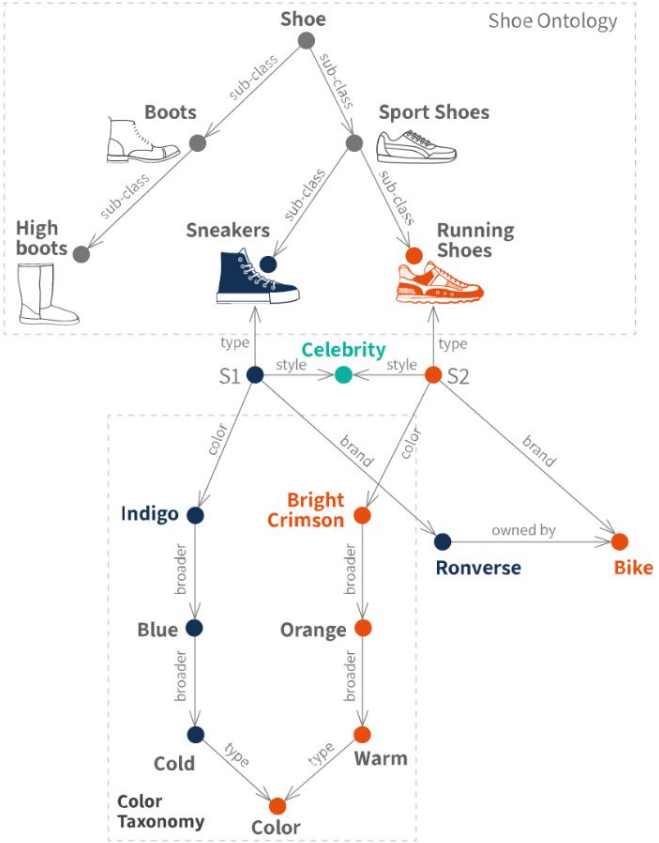
[illustration from ontotext.com]

Plain Graph

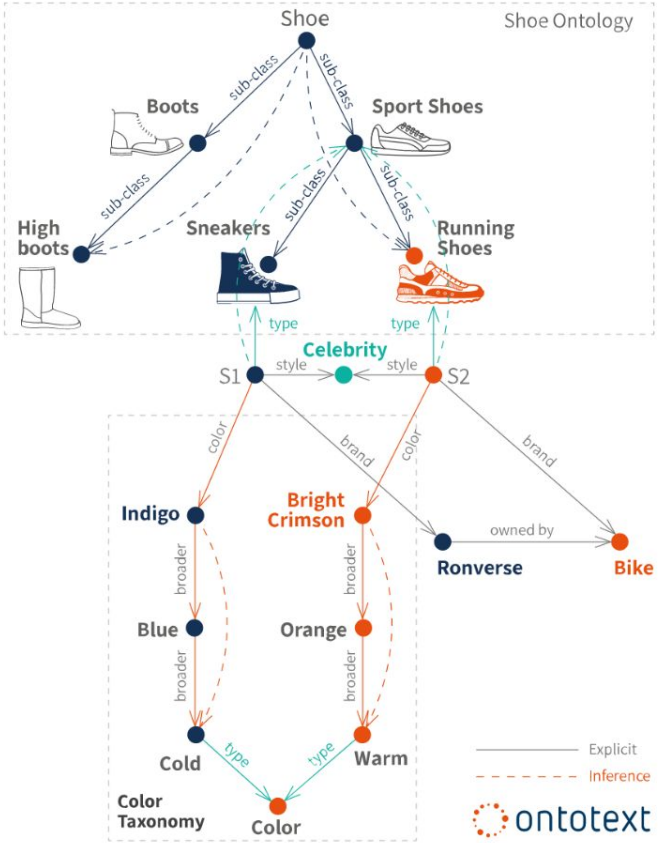
Standards for building a knowledge graph are available, which are universally digestible

Domain knowledge (Particle Physics, LHCb, Spectroscopy...) represented in Ontology

Knowledge Graph



Knowledge Graph with Inference



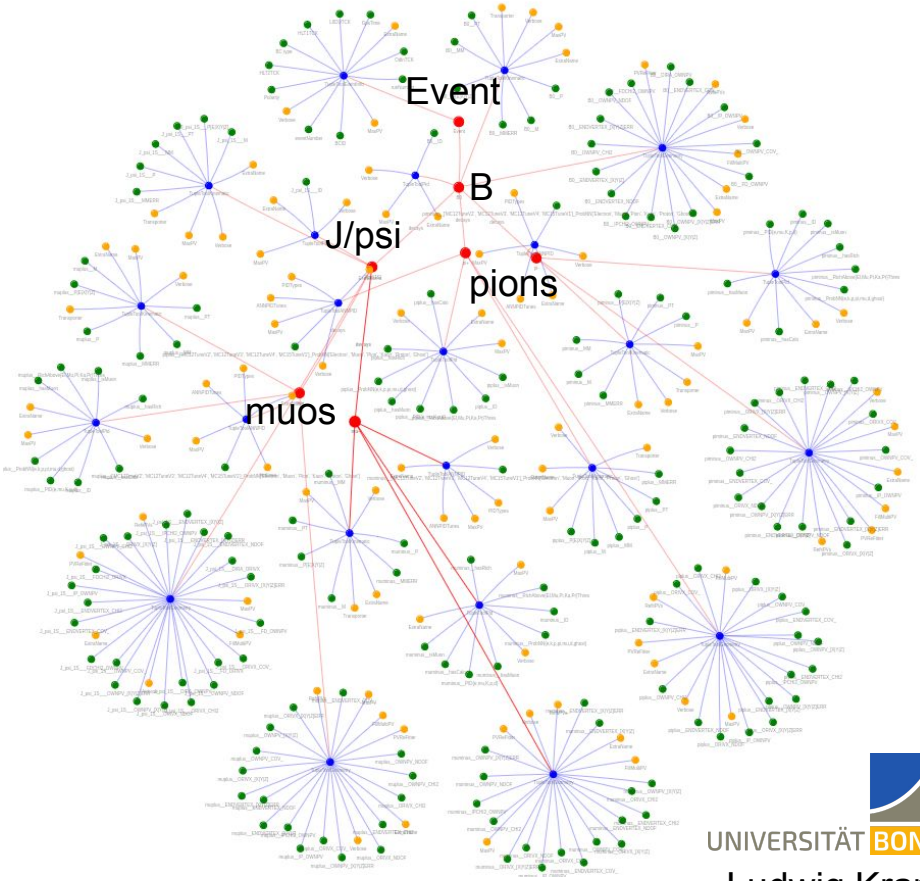
Frist Graph representation of an LHCb NTuple

Beispiel hier B-> Jpsi pi+ pi-

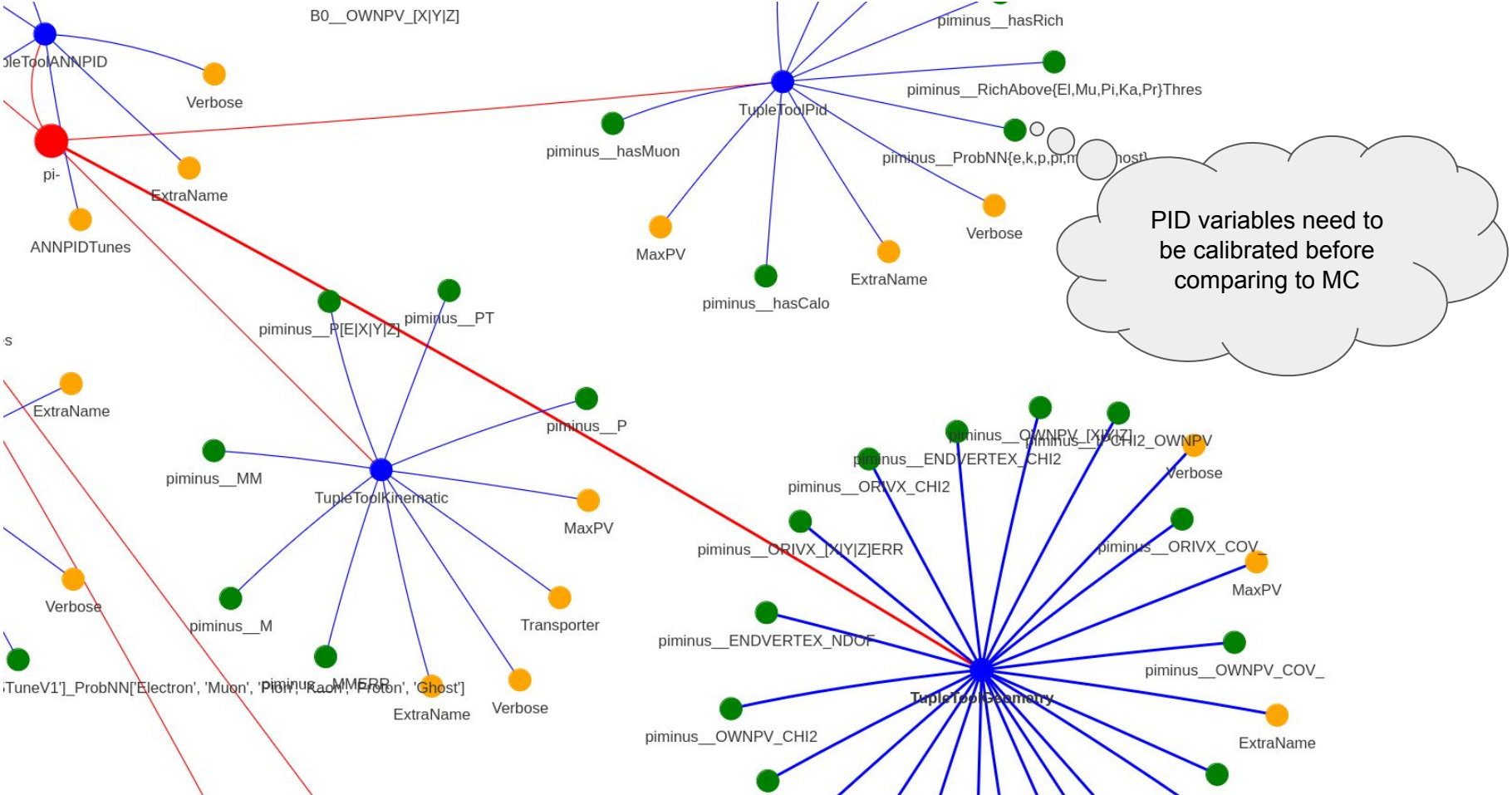
Preserving the links between the variables and the tools will allow to produce documentation for each one.

Nontrivial applications:

- Assistant to discover useful tools within Wizard
- Discover potential postprocessing/calibration steps implied by some variables

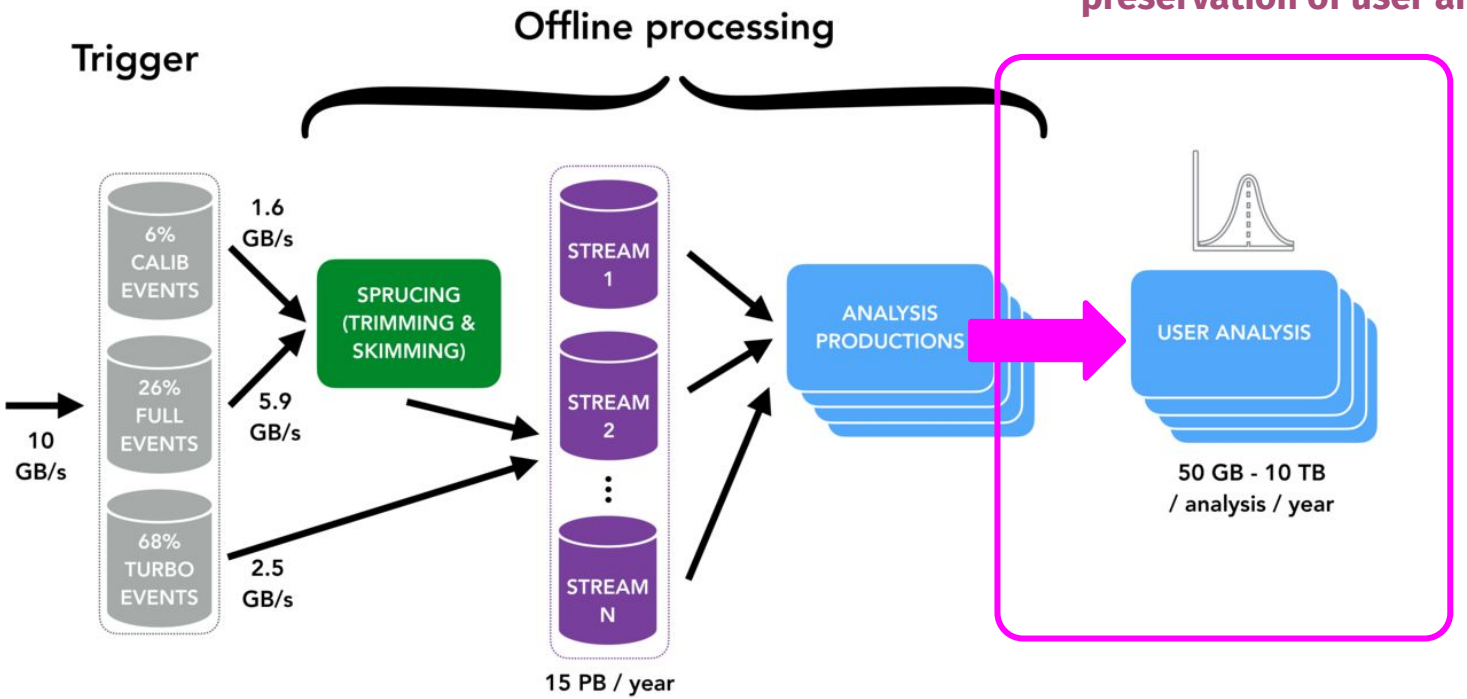


Frist Graph representation of LHCb NTuple



Preservation of User Analyses

This talk:
preservation of user analyses



Goals:
Allow analysts to creatively solve their analysis.

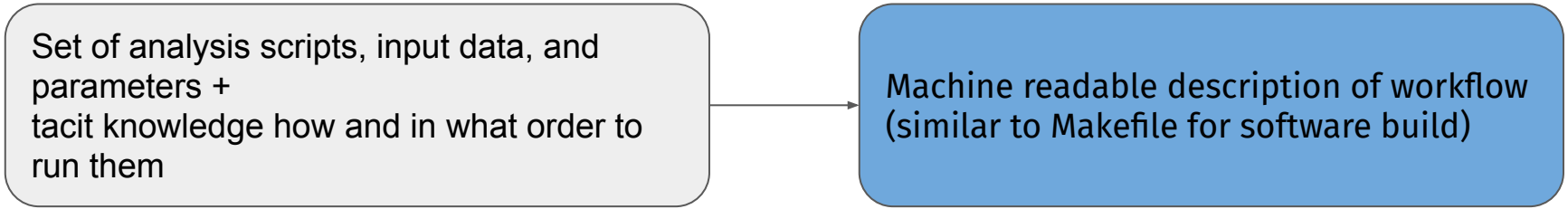
Flexible choice of tools and methods.

Preserve ingredients needed for **interpretation of data**

centrally managed and preserved | managed by proponents / PWG

data preparation | data interpretation

Snakemake workflow description



- Snakemake selected as top recommendation after comparative review in 2017
  (see LHCb-INT-2017-021)

- Wide use inside collaboration
- Feature complete
- Easy to get started
- Supported by CERN REANA

Snakemake is very well documented

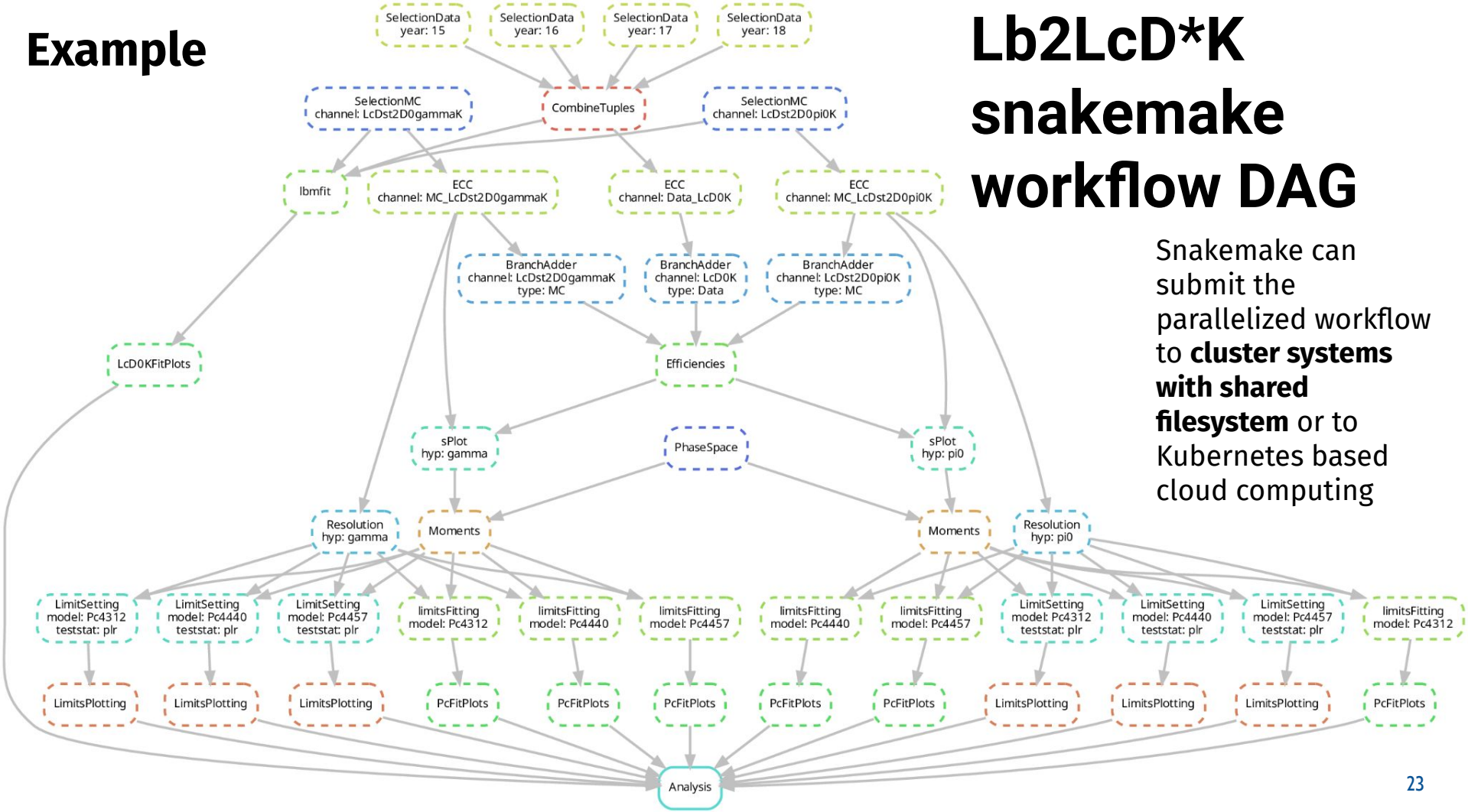
https://snakemake.readthedocs.io/en/stable/

https://snakemake.readthedocs.io/en/stable/snakefiles/best_practices.html

https://hsf-training.github.io/analysis-essentials/snakemake/README.html

Example

Lb2LcD*K snakemake workflow DAG



Snakemake can submit the parallelized workflow to **cluster systems with shared filesystem** or to Kubernetes based cloud computing

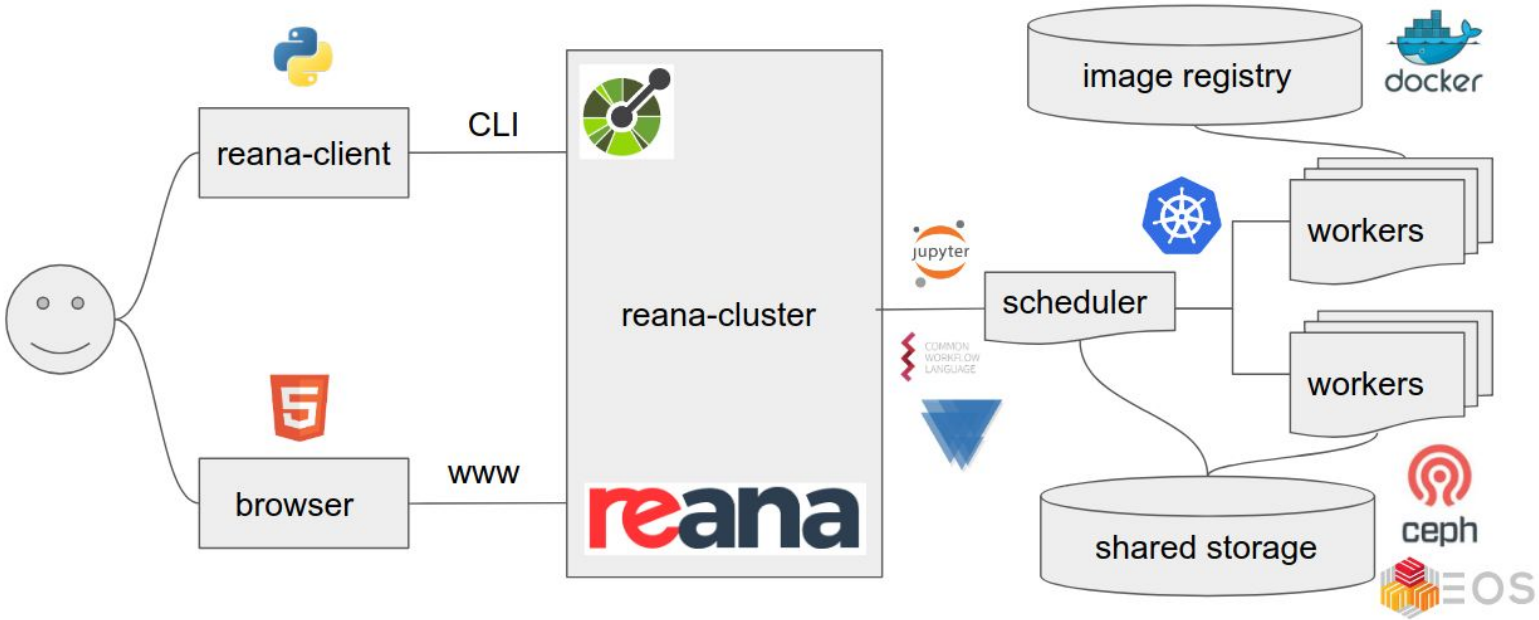
REANA: https://docs.reana.io/

Analysis facility dedicated to developing reproducible analyses

Supported:
Snakemake, Common Workflow Language, Yadage

Kubernetes, HTCondor, Slurm

REANA Mattermost



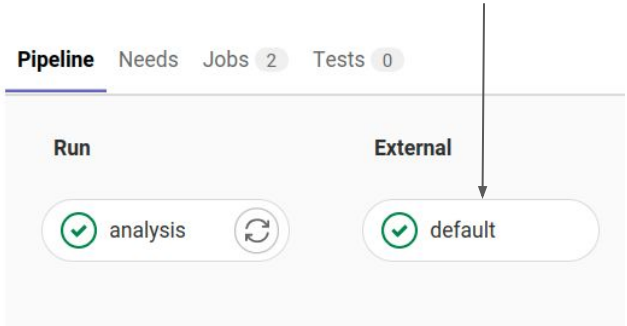
Deploy analysis to REANA via gitlab

https://reana.cern.ch/profile

reana.yaml 415 Bytes

```yaml
# reana-snakemake.yaml
version: 0.1.0
inputs:
  files:
    - scripts/filter_tree.C
    - scripts/helloworld.C
  directories:
    - workflow/snakemake
    - output/logs
  parameters:
    input: workflow/snakemake/config.yaml
workflow:
  type: snakemake
  file: workflow/snakemake/Snakefile
  resources:
    cvmfs:
      - lhcb.cern.ch
      - lhcbdev.cern.ch
      - lhcb-condb.cern.ch
outputs:
  files:
    - x.pdf
```

Switch on to deploy to REANA with next gitlab-ci job



REANA Webinterface https://reana.cern.ch

Interactive session possible via Jupyter

Summary

- LHCb is completing its Run I Open data release
  - Data taken in 2011/12
  - ~ 7500 Preselections
  - ~ 800TB of data
  - MC Samples on demand

- Custom (M)DST format
- Available data ≠ accessible, useable data

- NTuple Wizard will provide
  - Better access to the data, easier to use output format
  - Avoid dedicated open-data replicas of the large data sets
  - Machine readable documentation - knowledge graphs

- Preserving functional objects - such as unbinned Likelihoods - requires preserving runnable code - REANA

Backup

Policies the CERN experiments have given themselves

CERN Open Data Policy 2020

Initiated beginning 2020 by the chair of the European Commission

CERN director of research: Mandate for a working group to draft a common policy for all LHC experiments

Endorsed by the Collaboration Boards of ALICE, ATLAS, CMS and LHCb

CERN Open Science Policy 2022

Includes all experiments at CERN

https://openscience.cern/

Includes a wider scope of topics:

- Open access, open data, open source, open hardware
- Research integrity, research assessment
- Open infrastructure
- Training and outreach, citizen science

New Open Science Steering Board to be instantiated at CERN (S.N. LHCb delegate)

Open data policy: Level 3 data releases

Reconstructed Data (Level 3) Policy: The LHC experiments will release calibrated reconstructed data with the level of detail useful for algorithmic, performance and physics studies. The release of these data will be accompanied by provenance metadata, and by a concurrent release of appropriate simulated data samples, software, reproducible example analysis workflows, and documentation. Virtual computing environments that are compatible with the data and software will be made available. The information provided will be sufficient to allow high-quality analysis of the data including, where practical, application of the main correction factors and corresponding systematic uncertainties related to calibrations, detector reconstruction and identification. A limited level of support for users of the Level 3 Open Data will be provided on a best-effort basis by the collaborations.

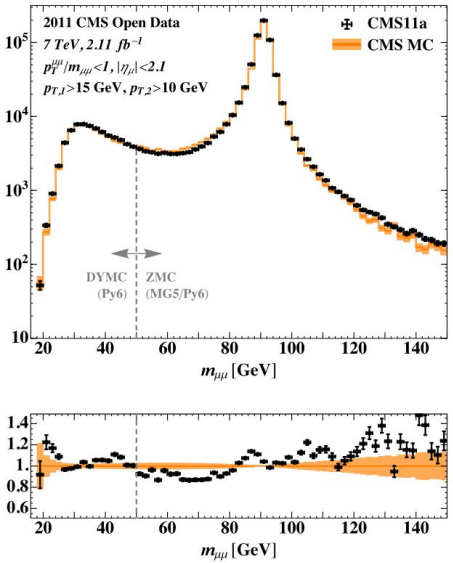
Level 3 data is addressed at professional researchers

How is LHC Open Data going to be used?

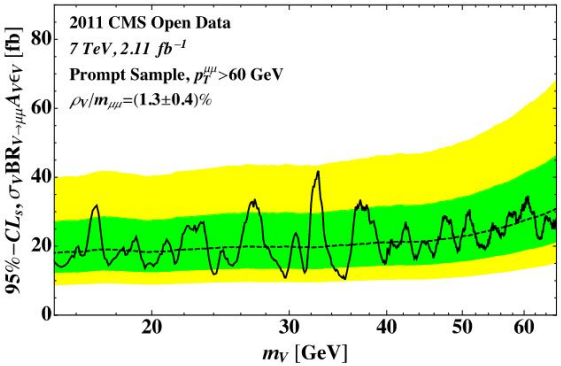
No experience for LHCb data, yet.

But various studies done on CMS open data.
Overview: arXiv:2106.05726

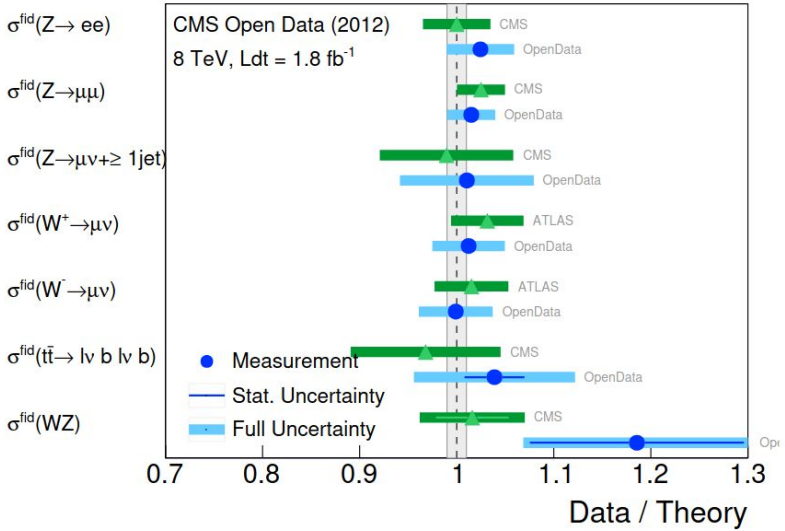
Dimuon spectrum [PRD100(2019)015021]:



Search for narrow dimuon resonances



SM cross section measurements on CMS open data [1907.08197]



Going beyond level 3 data

Open science and Open data policies:

We should publish ntuples and statistical models to make our results more impactful

5. Research integrity, reuse and reproducibility

CERN is committed to ensuring the integrity of research. In order to facilitate the reuse of its research products, CERN provides infrastructures to accommodate the scale and complexity of its research outputs. Reuse and reproducibility are facilitated by practising comprehensive analysis preservation to capture relevant research objects, such as research data releases with supporting metadata, auxiliary data, linked software, reproducible analysis workflows, documentation, etc.

what to publish depends on individual analysis

Published Results (Level 1) Policy: Peer-reviewed publications represent the primary scientific output from the experiments. In compliance with the CERN Open Access Policy, all such publications are available with Open Access, and so are available to the public. To maximise the scientific value of their publications, the experiments will make public additional information and data at the time of publication, stored in collaboration with portals such as HEPData,[4] with selection routines stored in specialised tools. The data made available may include simplified or full binned likelihoods, as well as unbinned likelihoods based on datasets of event-level observables extracted by the analyses. Reinterpretation of published results is also made possible through analysis preservation and direct collaboration with external researchers.

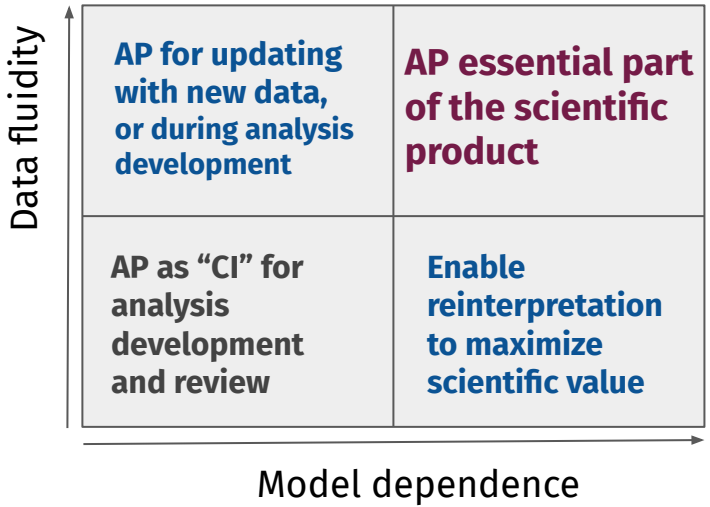
Who should use full analysis preservation (AP)?

Data fluidity

- updating analysis with new data
  - **e.g. early measurements**
- control channels and their analysis for calibrations and efficiencies
  - during commissioning
  - precision measurements

- **combining measurements**

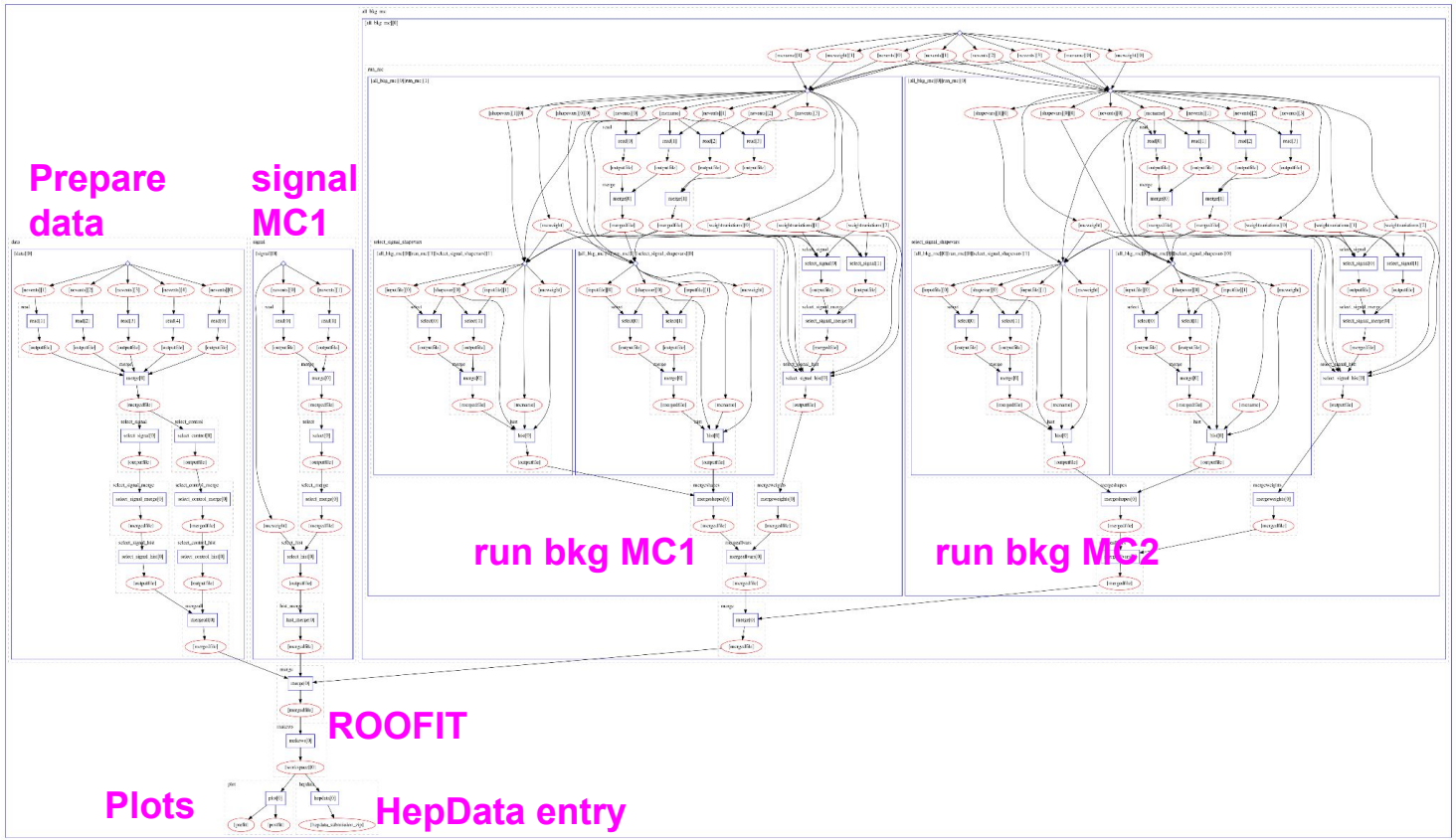
The level of detail of analysis preservation and published research products need to be decided case-by-case



Model dependence

- significant phenomenology input
  - **amplitude analyses!**
- choice of observables based on theory input
- **auxiliary inputs**: e.g. Formfactors
- MC generators / samples
- statistical methodology

https://github.com/reanahub/reana-demo-bsm-search



Analysis Productions

Starterkit Lesson

- Ntuple production metadata preserved automatically!
- Not yet supporting Run 3 DaVinci
  - Conversion will be simple from

```
lb-run DaVinci/vXrY \
      gaudirun.py my_options.py
```

- Need to maintain link between ntuple production and analysis
- Will be able to use apd (Analysis Production Data)
- Provides PFN(s) for datasets
  - Designed to allow analyses to be rerunnable long-term

```
import apd
datasets = apd.AnalysisData("MyWG", "MyAnalysis")

rule train_bdt:
    input:
        data = datasets(datatype="2022", mc=False),
        mc = datasets(datatype="2022", mc=True)
    output:
        fn = "classifier.pkl"
    shell:
        "scripts/train_bdt.py --data {' '.join(input.data)} --mc {' '.join(input.mc)}"
```

Concluding remarks: Curating Research Products

- Different scientific questions require
  different levels of detail in the empirical evidence.
  - **Level of model-dependence** will influence how much the experimental data can be "compressed" into a few numbers.
  - Techniques that allow reinterpretation of the data
    are the same as those needed to adapt to a fluid dataset

- Decisions on the level of detail of analysis preservation have to be tuned to the individual study - IMHO: avoid one-fits-all solutions
  - It is possible to support this with a small number of generic tools,practices, and standards

- This **data curation** requires dedicated resources.
  - Maximizing scientific value is not for free

- The technologies used to support the effort are very useful beyond fundamental science. Come join us!

The Open Science Philosophy (at CERN)

Recognize the universal importance of the fundamental scientific knowledge produced at CERN and the key role of openness in the pursuit of CERN organisational mission.

Commits to the advancement of science and wide dissemination of knowledge by adopting practices to make scientific research more open, global, collaborative and responsive to societal changes.

In fulfilment of the collective moral and fiduciary responsibility to member states and the broader global scientific community

Data collected at the LHC is a heritage to humanity.

It has been obtained through collaborative work using public funds.

Therefore, CERN is committed to preserve, curate, steward and share the data with the public.

Goals of Open Data - Maximizing Scientific Value

- Validation / reproduction of published results
- Reinterpretation of the data
  - test future theories
  - refine phenomenological models
  - use different statistical tools
- Reuse of data sets
  - Combined analyses
  - Use collected data as input for future studies
  - Algorithm development (e.g. machine learning community)
- Data mining
  - search for interesting physics in unexplored parts of the data
  - use new techniques to (re-)select data

We cannot anticipate the questions future generations might ask of this data.

require different levels of data complexity

Open Science Landscape - Recent Trends

- Funding agencies: requests for data management plans
- Publishers: requests for data products allowing to
  - validate / reproduce results
  - reuse data for further studies

Science Community: “**Data is not enough**”:

- Papers with code https://paperswithcode.com/
- Interactive publications
- Federated infrastructures and computing/science portals (e.g. NFDI)
- Not a new realization (see e.g. DPHEP study group 2013 status report) but technology (esp cloud computing, containerization) has made progress!
- Development driven especially through bioinformatics and machine learning / AI community