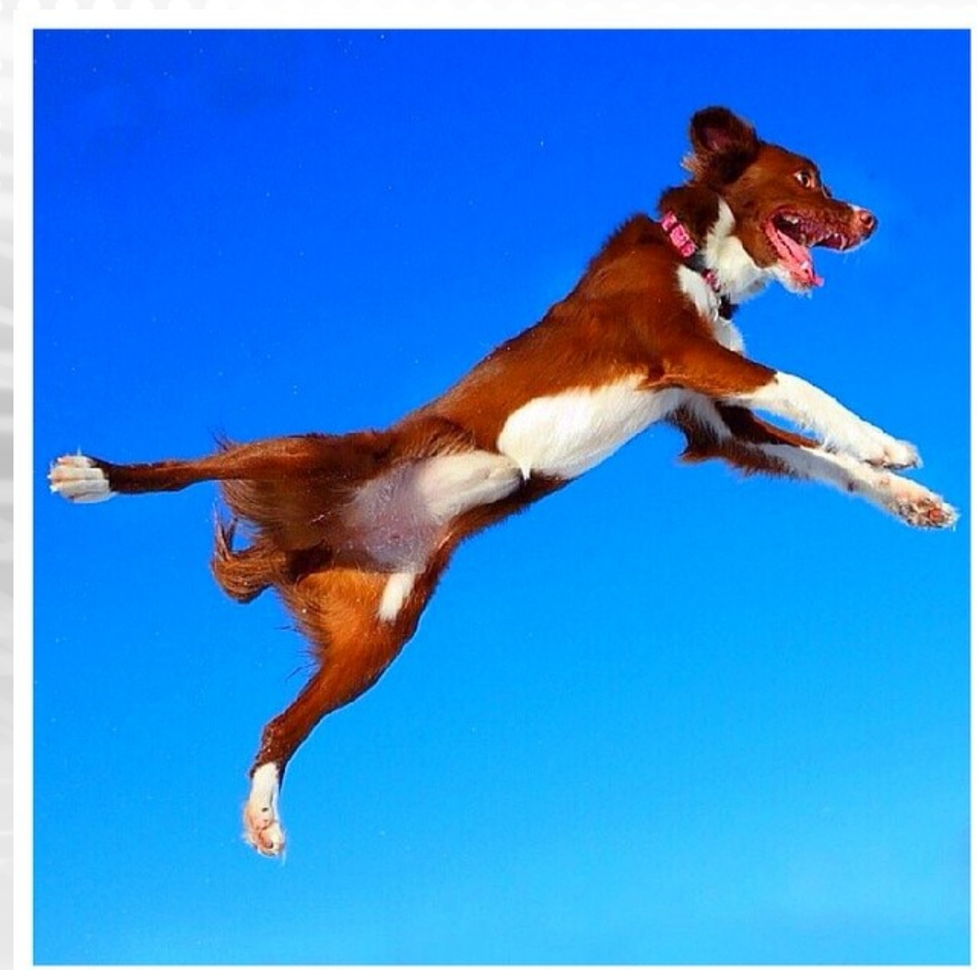
Data Analysis Working Group

Analysis today

Paul Laycock



BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Introduction to the Data Analysis Working Group

- **Aim - make publication of physics results more efficient, eliminate monotonous and laborious tasks from physics analysis**

- 1st priority - capture the requirements of analysis by direct consultation:
  - https://indico.cern.ch/event/782504/

- Second 1st priority - survey work of technology pioneers:
  - https://indico.cern.ch/event/789007/

- ***18 excellent talks*** which the three **DAWG** convenors will try to summarise
  - **What have we learned? What can we improve? How bad is it?**

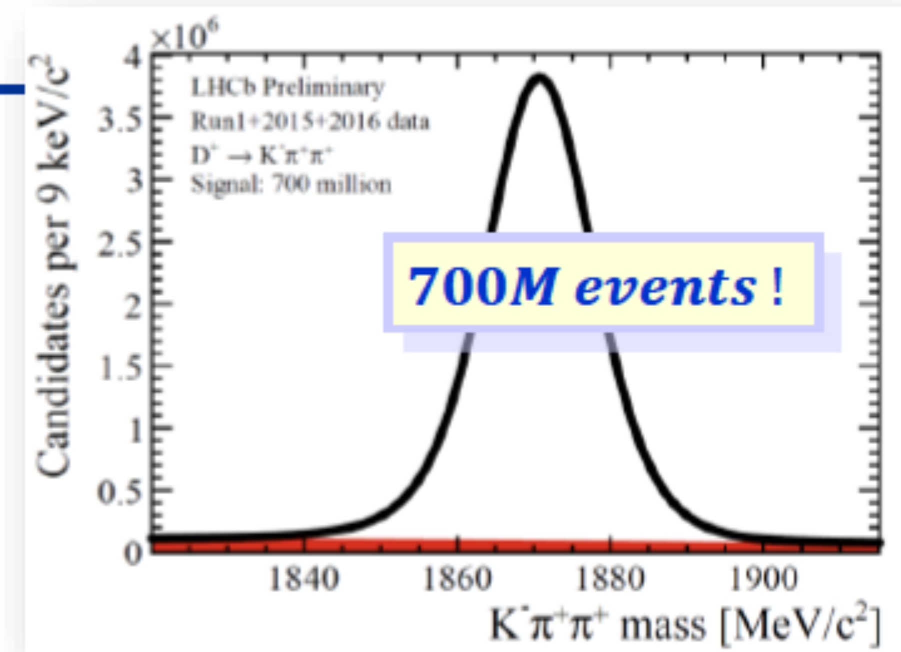
- Many thanks to the speakers, most of the material in this talk originated there
  - Credit goes to the original presentations (not always credited here, sorry!)

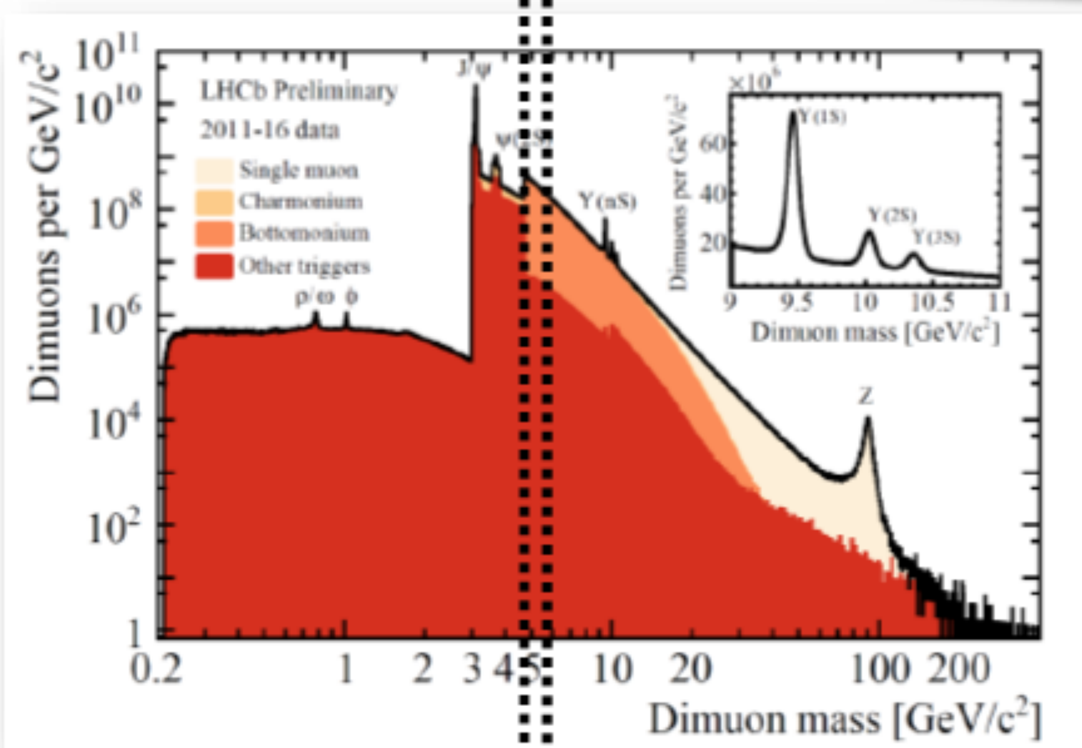
What is analysis?


CSC 2016, Mol Belgium@CERN

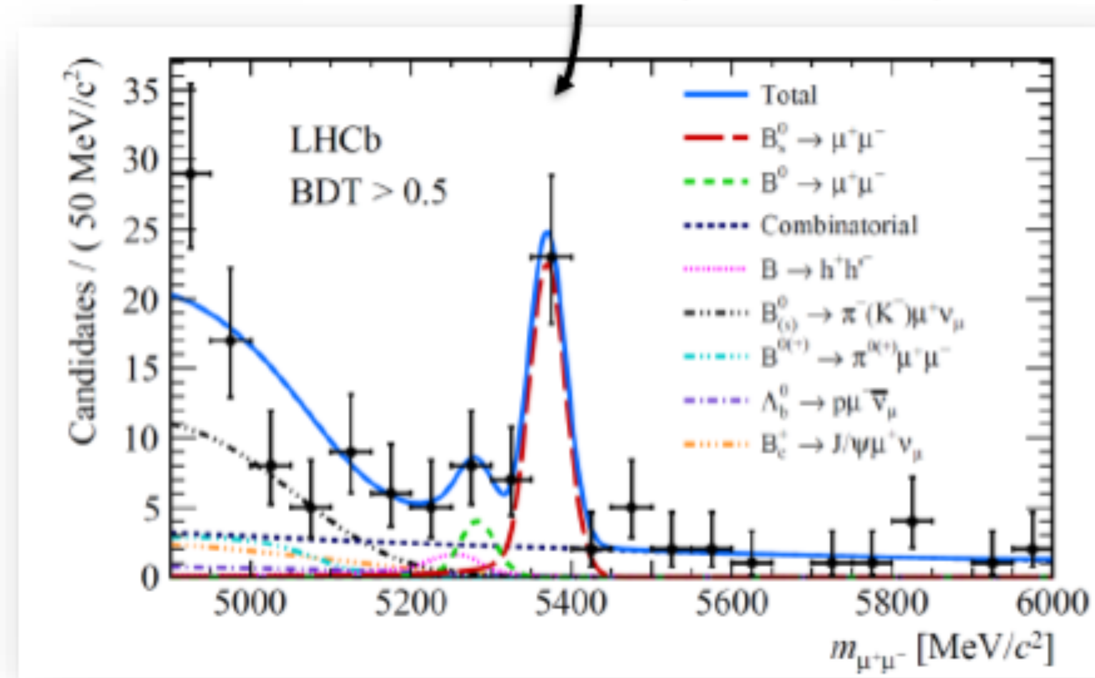
A question of many scales

Wildly different Challenges !



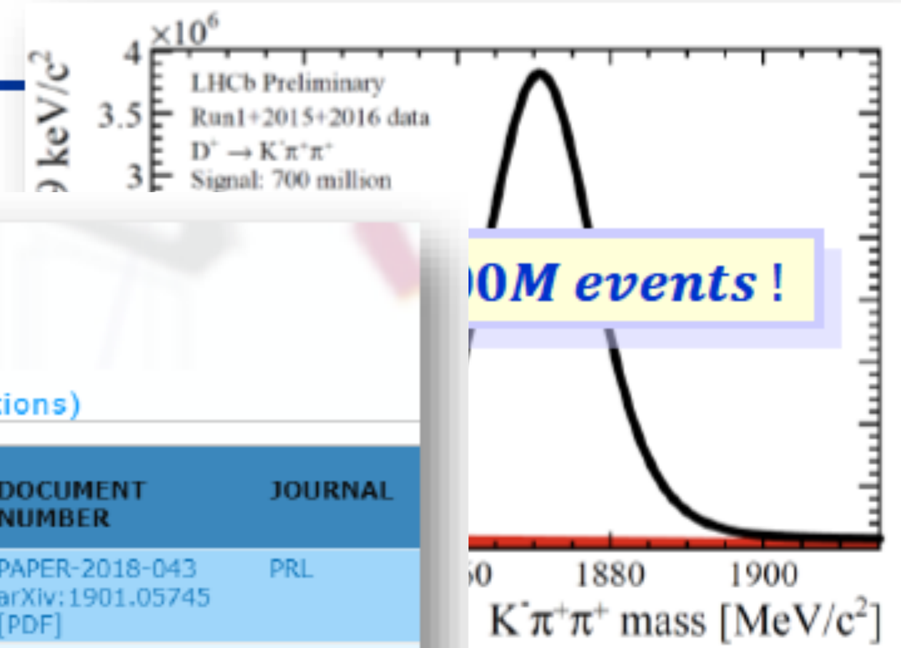


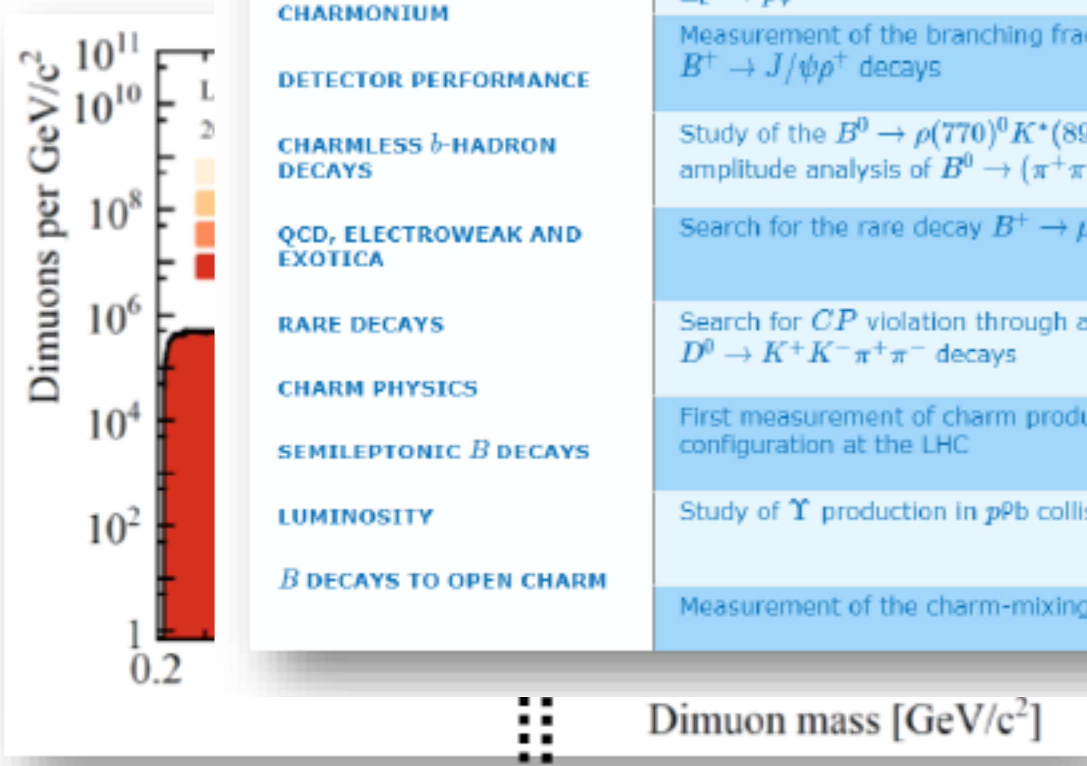
$$\mathcal{B}(B_s^0 \to \mu^+\mu^-) = (3.0 \pm 0.6\,^{+0.3}_{-0.2}) \times 10^{-9}$$

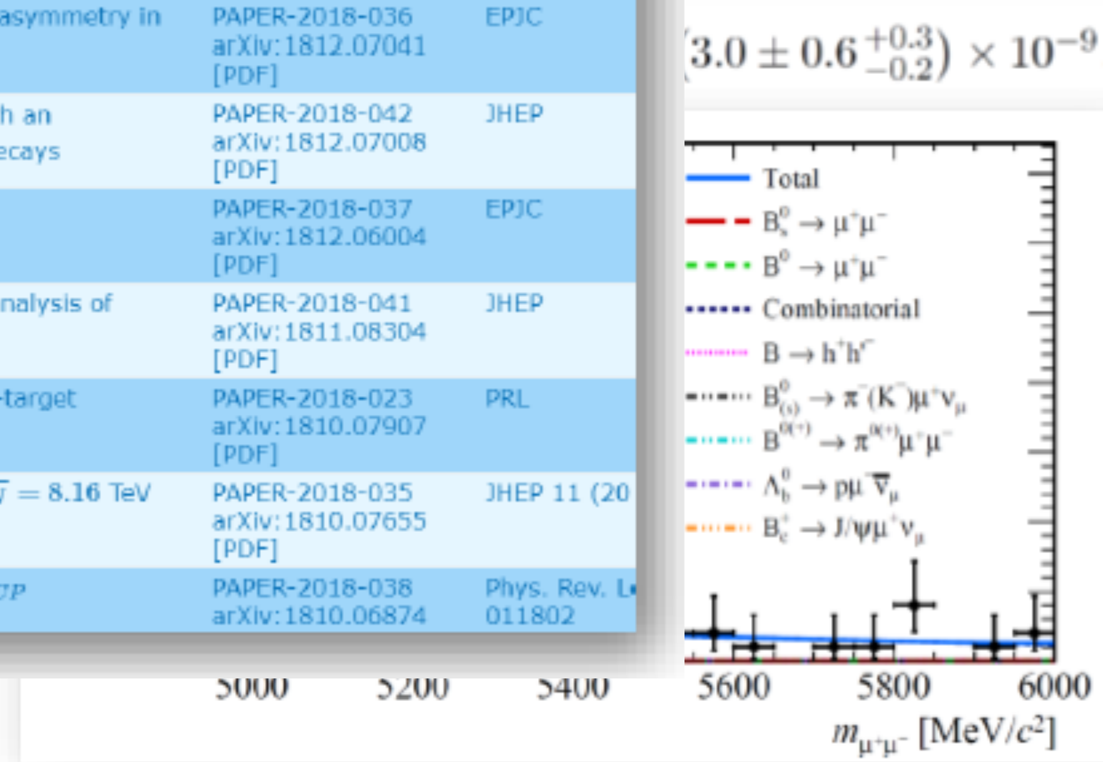


Note : structure given the numerous "trigger lines" with different requirements

A question of many scales







LHCb publications

459 papers

[to restricted-access page]

PUBLICATIONS PER WORKING GROUP

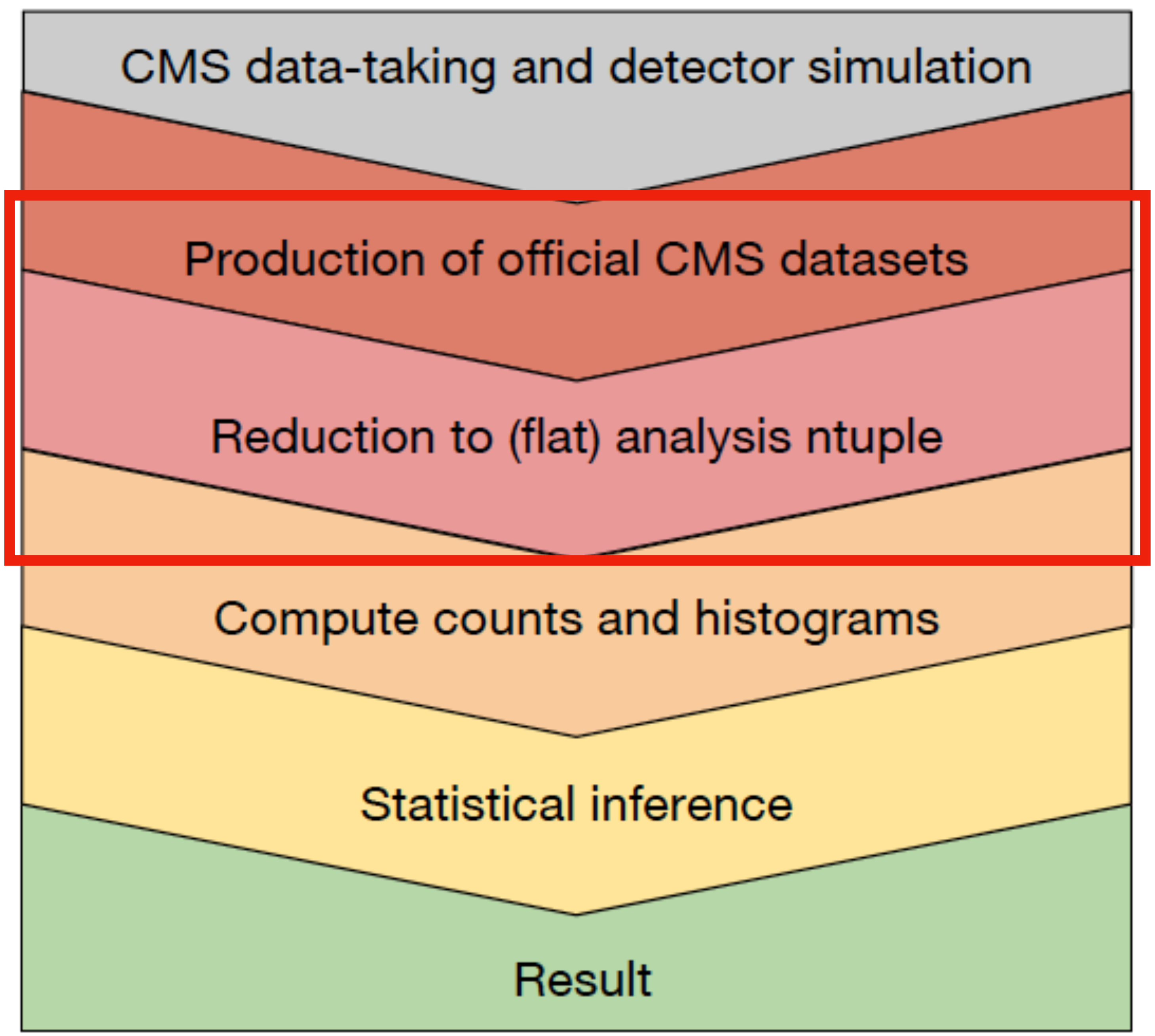
- IONS AND FIXED TARGET
- FLAVOUR TAGGING
- b-HADRONS AND QUARKONIA
- B DECAYS TO CHARMONIUM
- DETECTOR PERFORMANCE
- CHARMLESS b-HADRON DECAYS
- QCD, ELECTROWEAK AND EXOTICA
- RARE DECAYS
- CHARM PHYSICS
- SEMILEPTONIC B DECAYS
- LUMINOSITY
- B DECAYS TO OPEN CHARM

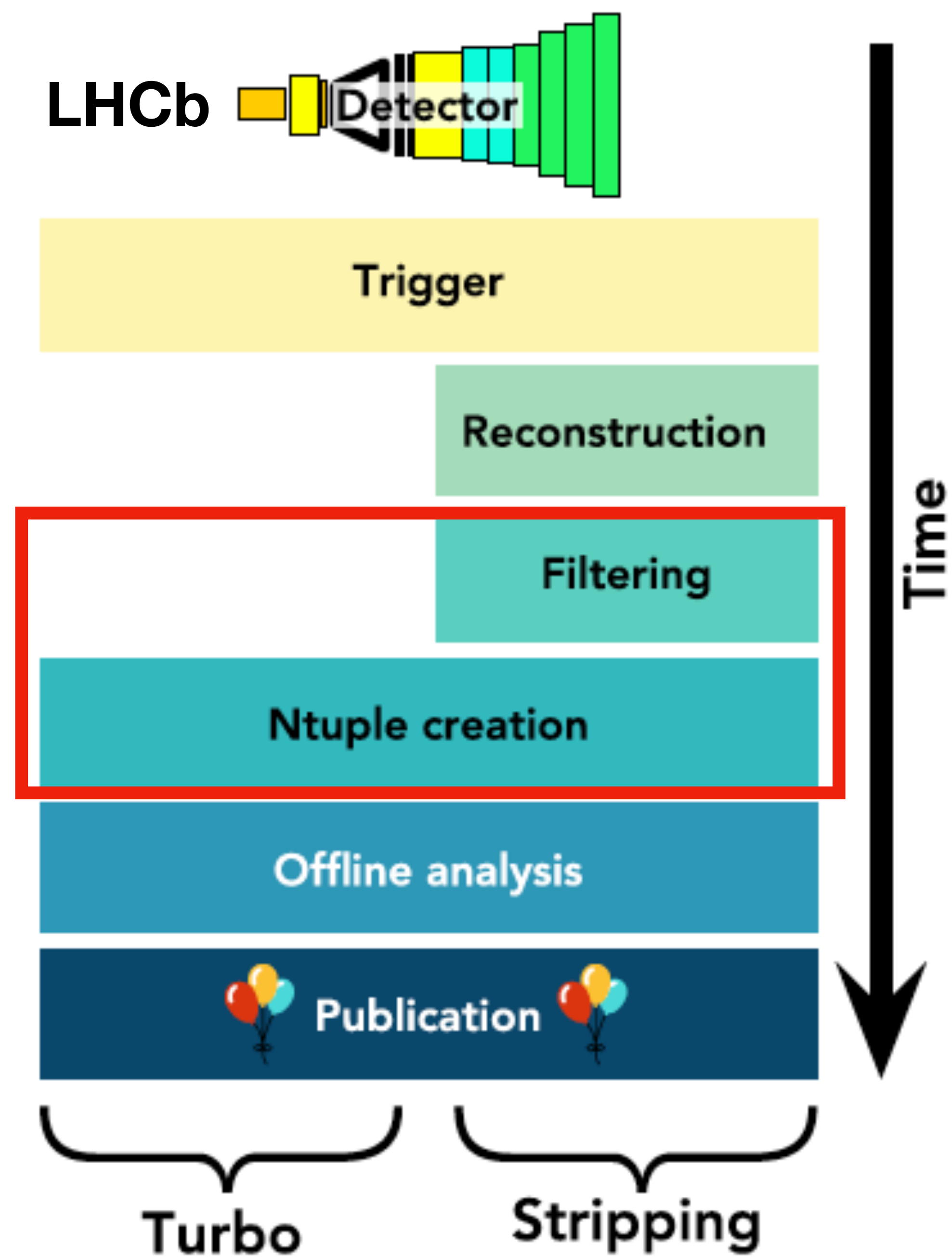
List of papers (Total of 459 papers and 25745 citations)

TITLE	DOCUMENT NUMBER	JOURNAL
Model-independent observation of exotic contributions to $B^0 \to J/\psi K^+\pi^-$ decays	PAPER-2018-043 arXiv:1901.05745 [PDF]	PRL
Observation of the doubly Cabibbo-suppressed decay $\Xi_c^+ \to p\phi$	PAPER-2018-040	JHEP
Measurement of the branching fraction and CP asymmetry in $B^+ \to J/\psi\rho^+$ decays	PAPER-2018-036 arXiv:1812.07041 [PDF]	EPJC
Study of the $B^0 \to \rho(770)^0 K^*(892)^0$ decay with an amplitude analysis of $B^0 \to (\pi^+\pi^-)(K^+\pi^-)$ decays	PAPER-2018-042 arXiv:1812.07008 [PDF]	JHEP
Search for the rare decay $B^+ \to \mu^+\mu^-\mu^+\nu_\mu$	PAPER-2018-037 arXiv:1812.06004 [PDF]	EPJC
Search for CP violation through an amplitude analysis of $D^0 \to K^+K^-\pi^+\pi^-$ decays	PAPER-2018-041 arXiv:1811.08304 [PDF]	JHEP
First measurement of charm production in fixed-target configuration at the LHC	PAPER-2018-023 arXiv:1810.07907 [PDF]	PRL
Study of Υ production in pPb collisions at $\sqrt{s_{NN}} = 8.16$ TeV	PAPER-2018-035 arXiv:1810.07655 [PDF]	JHEP 11 (20
Measurement of the charm-mixing parameter y_{CP}	PAPER-2018-038 arXiv:1810.06874	Phys. Rev. L 011802

Note : structure given the numerous “trigger lines” with different requirements

Analysis workflows

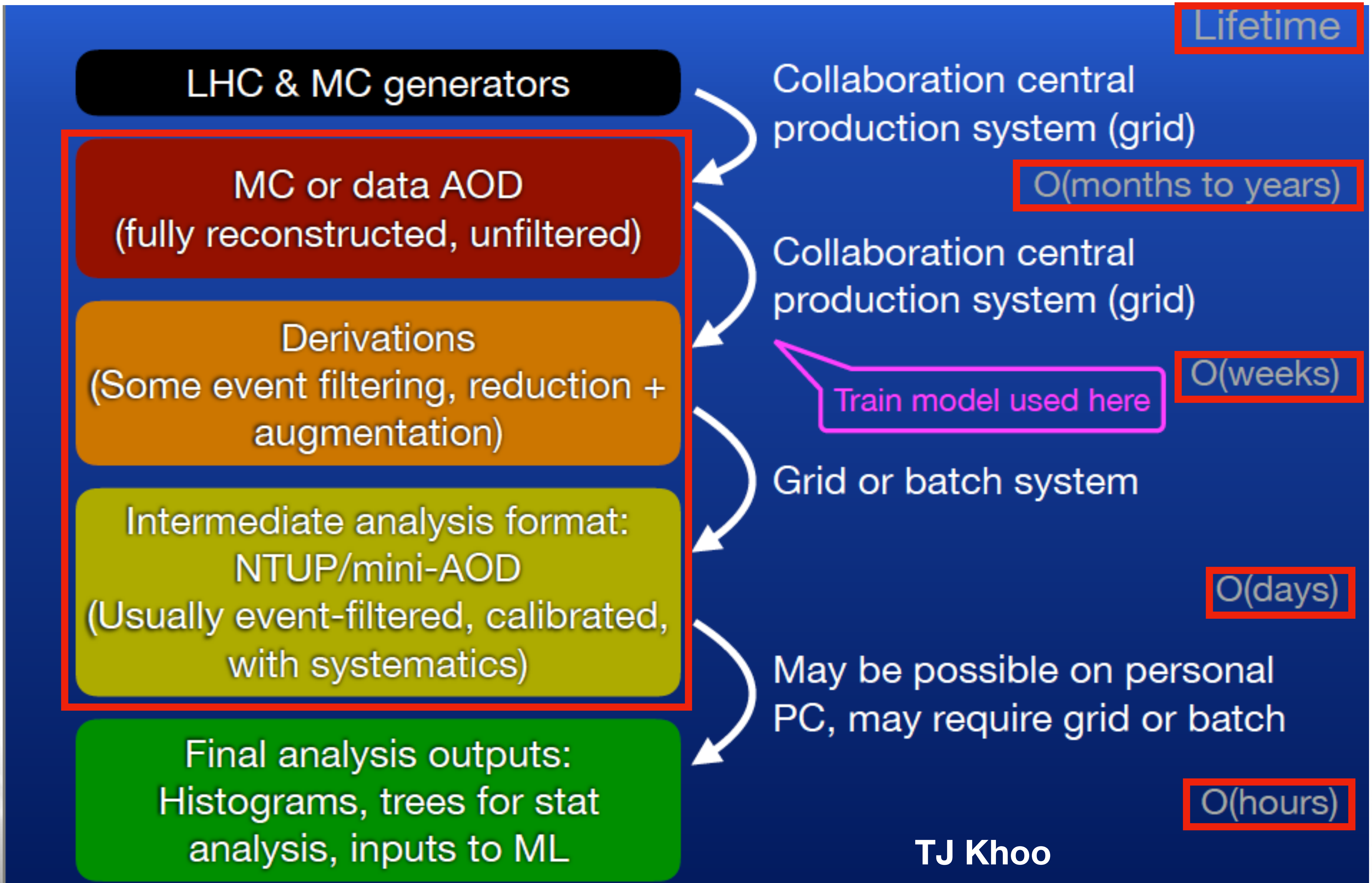




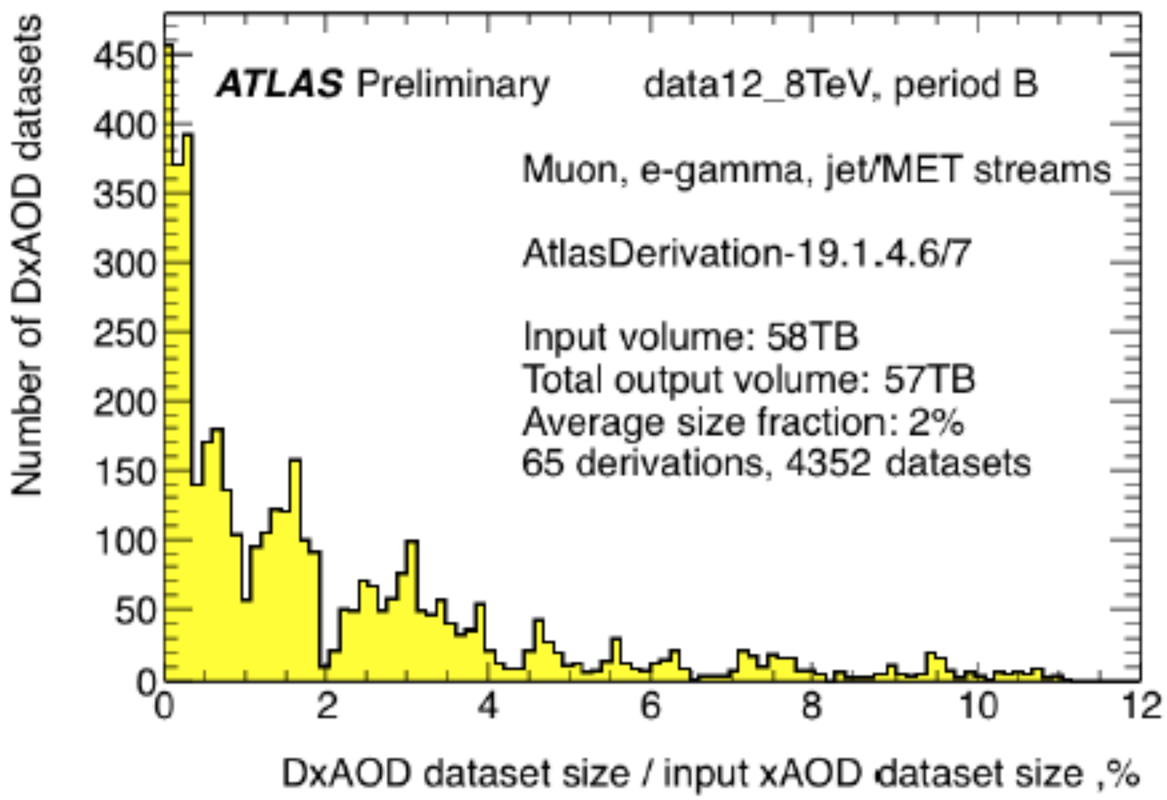
- Tend to focus on **heavy lifting** here rather than the final stages
  - More on statistical inference et al in **Andrea's** talk

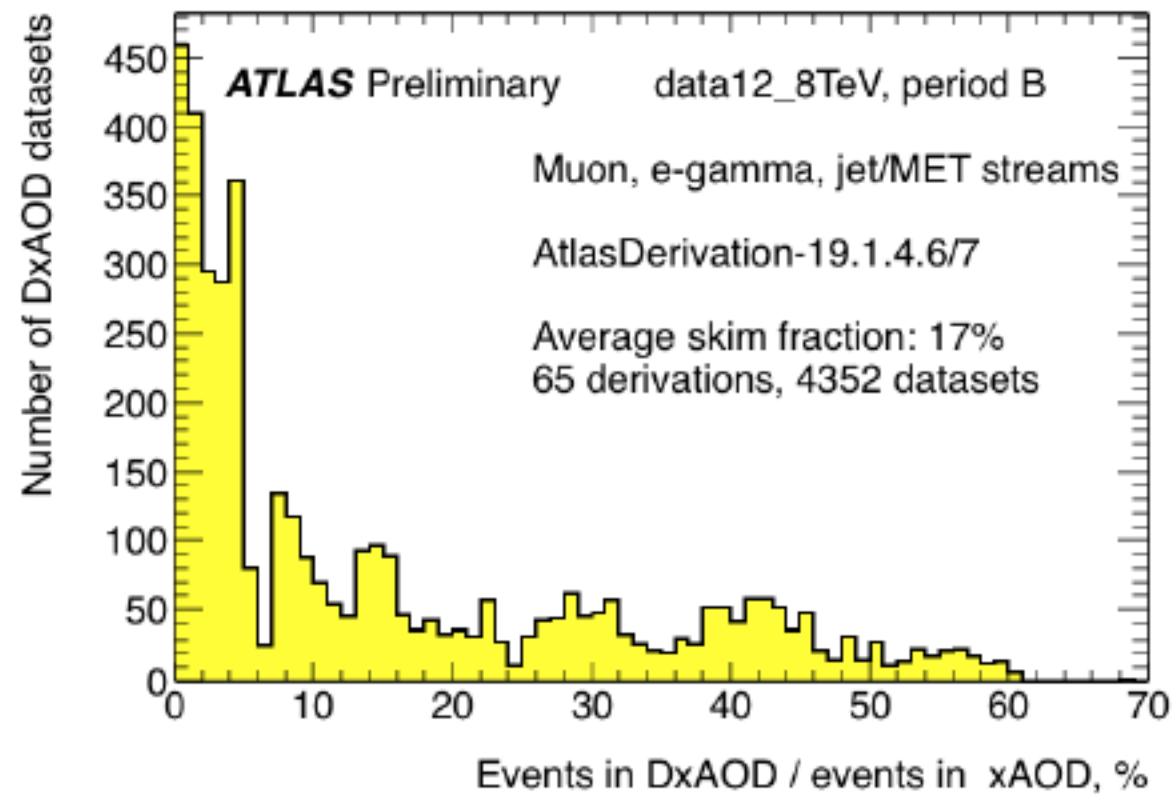

SYAHRUL RAMADAN photography

Repeated heavy lifting



Data reduction trains - Fact and fiction



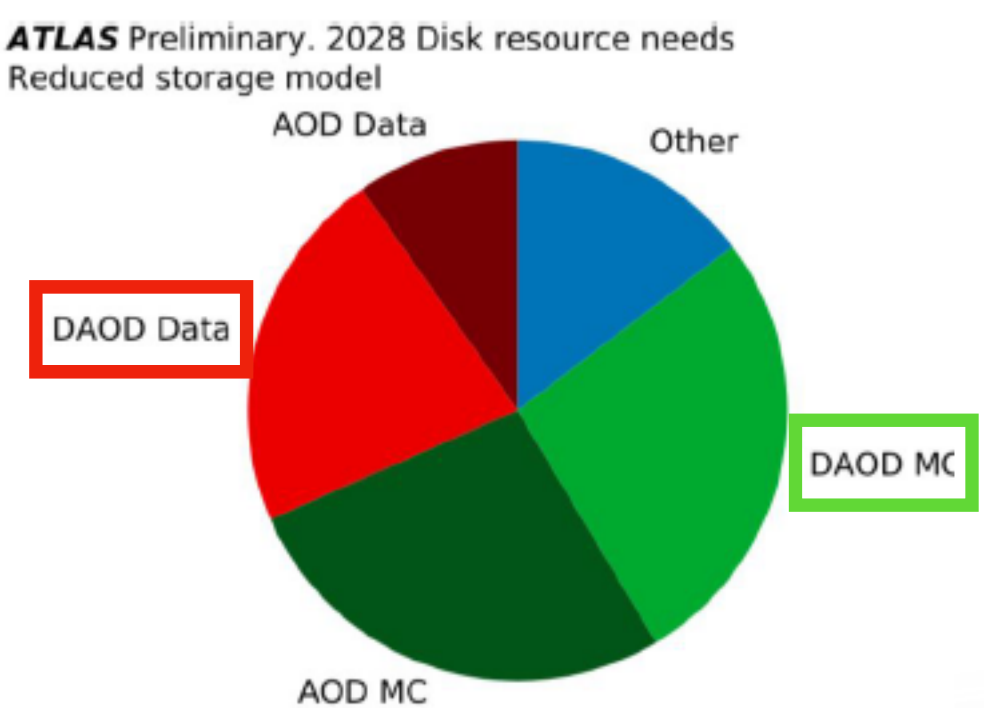


CMS nano-AOD ~1kB/event

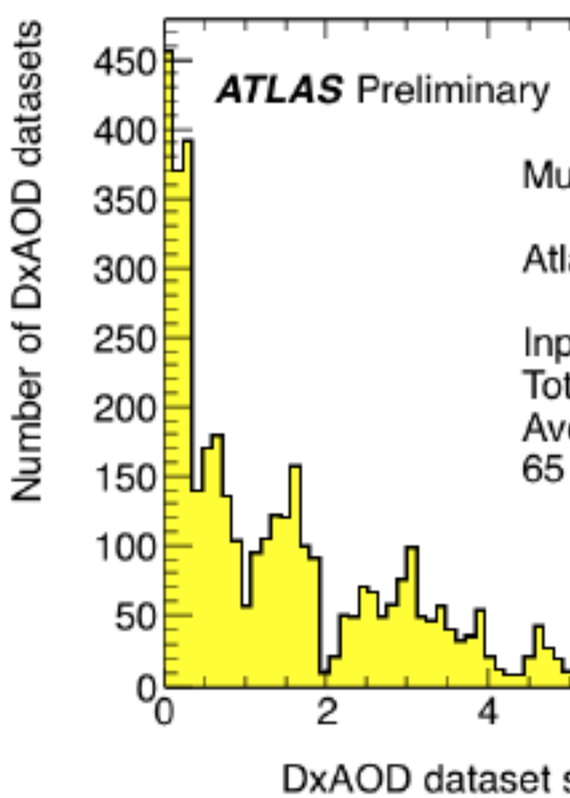
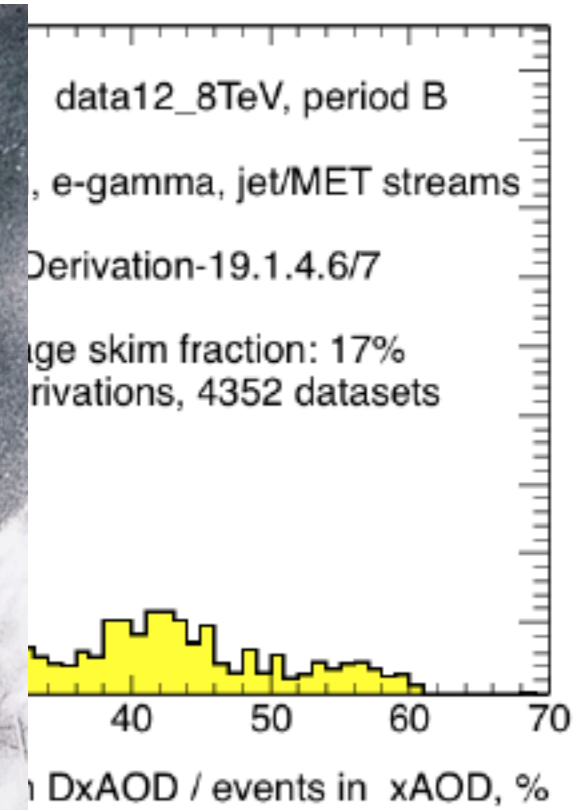
- **expected to cover > 50% of analyses**

ATLAS derivations

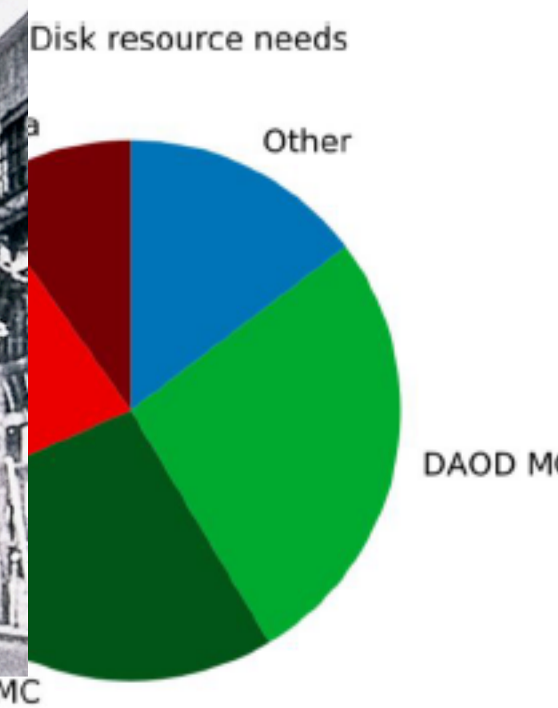
- **2014** (*top*): small efficient data format
- **2028** (*right*): more than half the storage



Data reduction trains - Fact and fiction

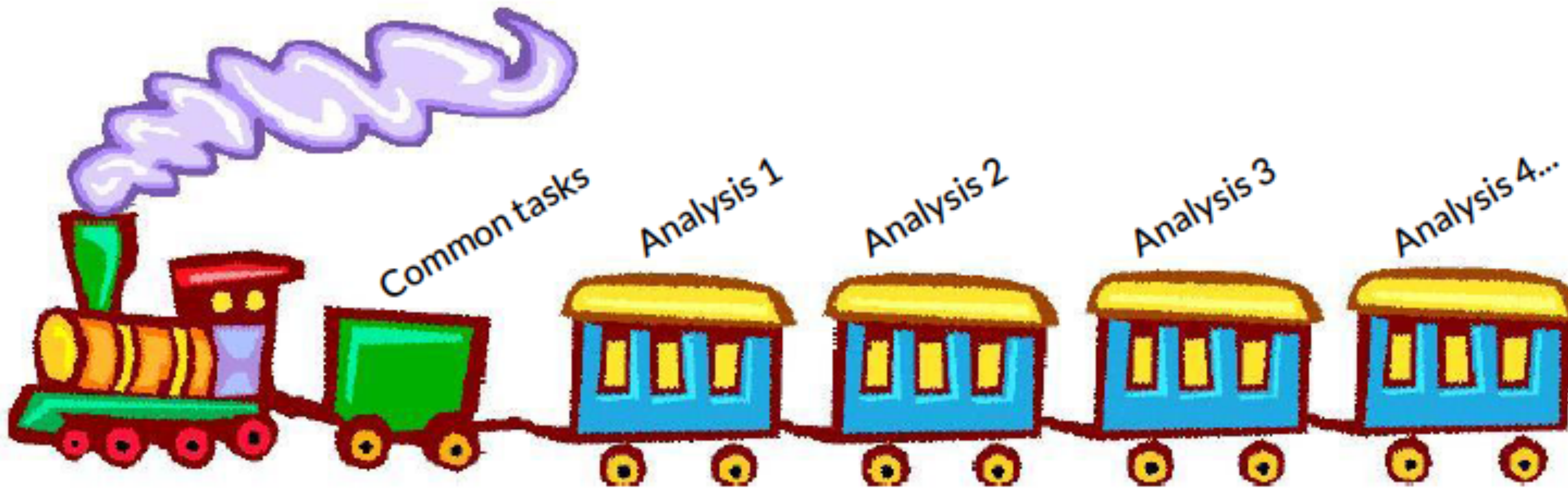



ATLAS deriva



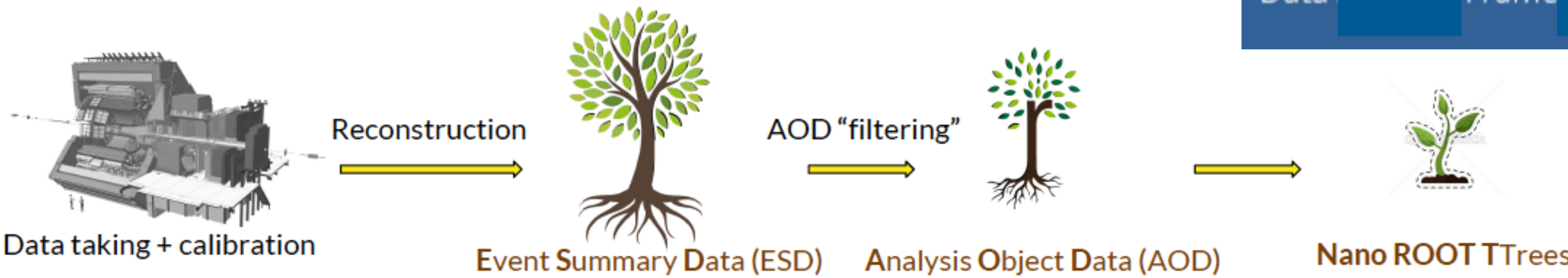
- **2014** (*top*):
  - small efficie
- **2028** (*right*):
  - more than h

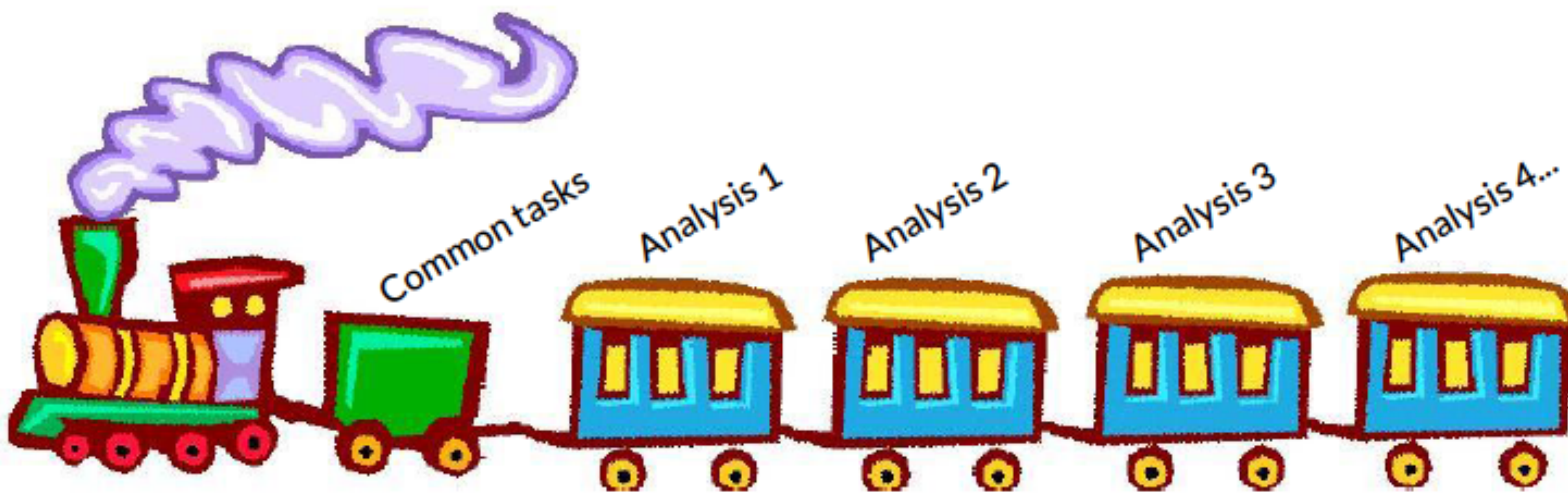
ALICE analysis trains



ALICE successfully driving analysis trains over **AOD** input

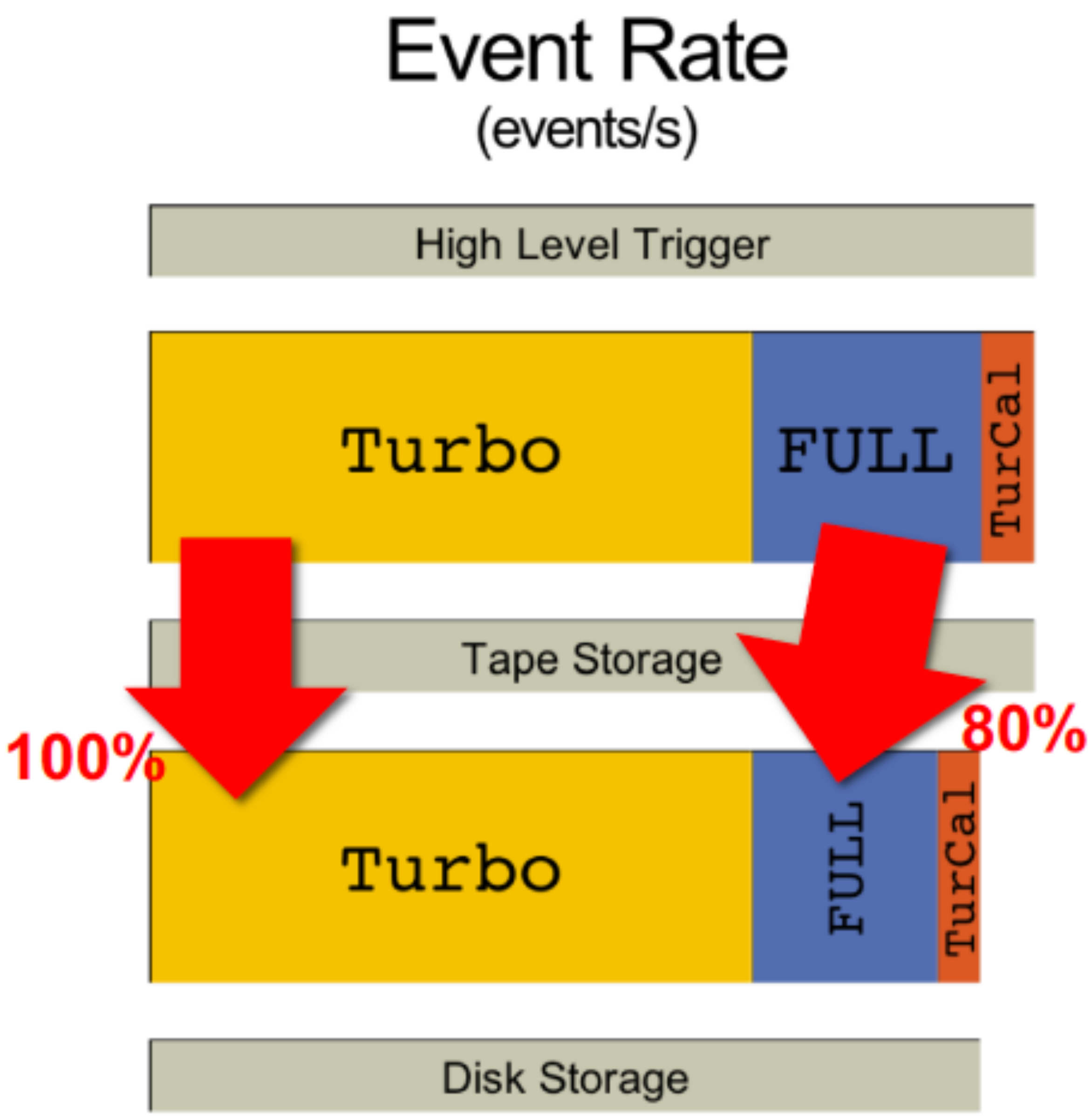
Going nano





ALICE plans to go nano like **CMS**, while **ATLAS** aiming for ***10-50 kB/event***

Analysis workflows - best practice?



- ***Turbo stream analysis (data scouting):*** ***reco and calibration done once in HLT*** ***<u>no more reprocessing!</u>***
  - in reality for LHCb, two HLT passes
    - TDAQ to record events to a big buffer
      - then prompt calibration
    - then second pass for data reduction
- ***Q. How much physics bandwidth can go this route for other experiments?***

- ***Centrally produced nano-format:*** ***no more reinventing the wheel for producing a data analysis format***
  - in reality, cannot accommodate all analyses, BUT important to use where it is possible (maybe even for ATLAS)
- ***Q. How much physics bandwidth can go this route for all experiments?***

Where: Power vs control

- Grid
  - Portability
  - Dataset sharing
  - Dataset access
  - Reliability
- Local cluster
  - Dataset access
  - Reliability
  - Portability
  - Dataset sharing

How easy for collaborator B to use collaborator A's submission scripts?

How easy for collaborator B to use collaborator A's job outputs?

How easy for collaborator A to use own job outputs?

How long before job outputs 100% available?

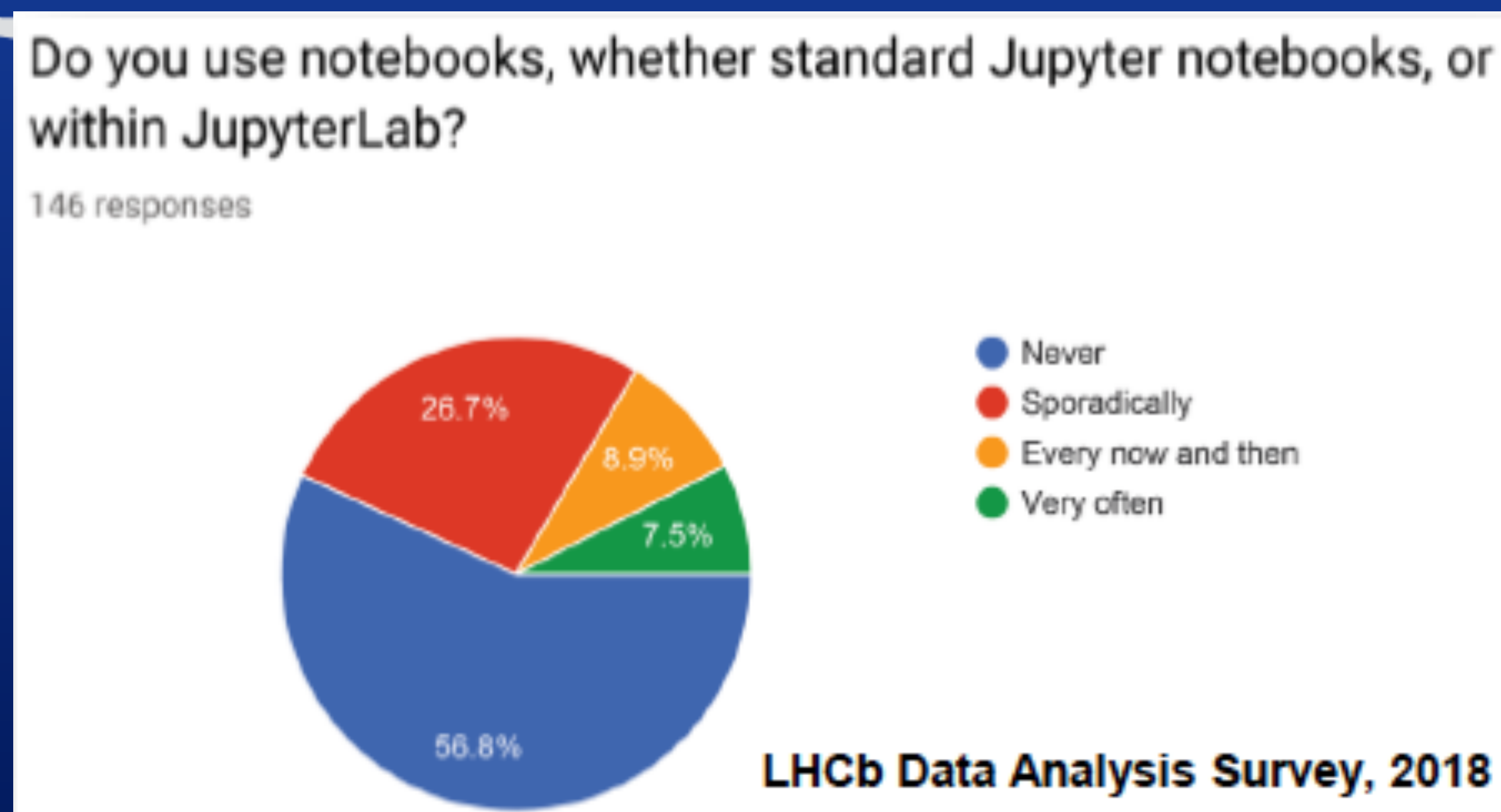
TJ Khoo

Wherever: Power and control?

- Grid
  - Portability
  - Dataset sharing
  - Dataset access
  - Reliability
- Local cluster
  - Dataset access
  - Reliability
  - Portability
  - Dataset sharing

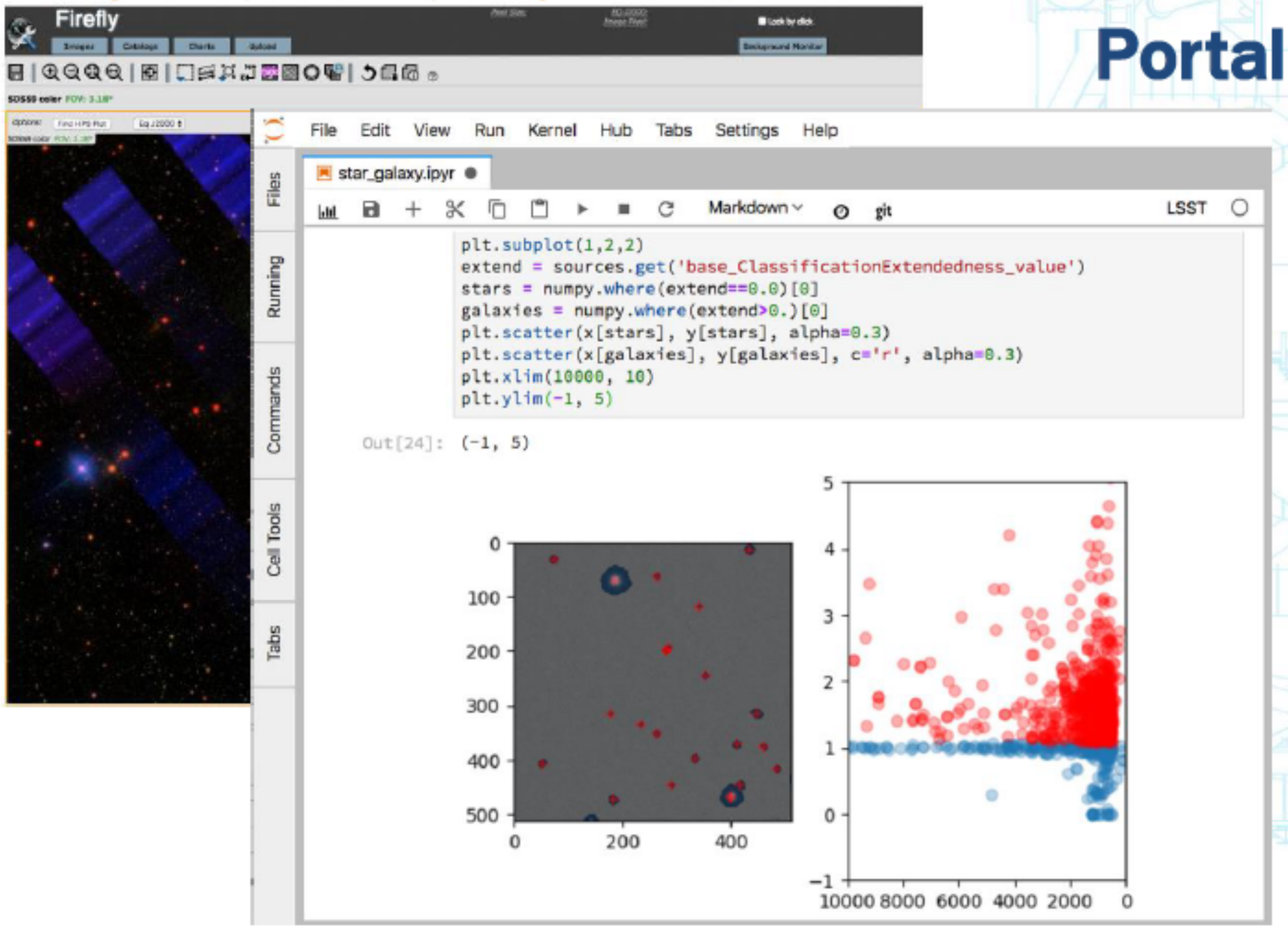
Hiding the "how" is a common theme

see declarative analysis in Andrea's talk



Analysis platforms

LSST Science Platform



Portal

Jupyter Notebooks

Web APIs

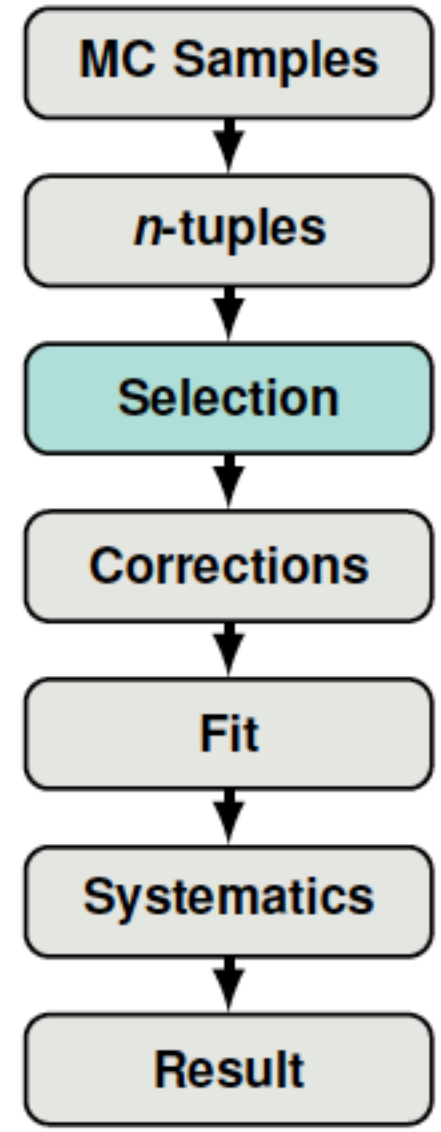
- Data access via IVOA-standard protocols
- Same interfaces that support other aspects

• See also Lukas's talk from Tuesday

Belle II analysis software stack

Analysis of the *n*-tuples is done with Python:

- *Pandas* and *numpy*
- *root_pandas* or *uproot* to load ROOT files
- *scikit-learn* or basf2 MVA package for MVA methods
- *matplotlib* for plots
- convert *n*-tuples to hdf5 files (these are loaded ∼10 times faster)
- data analysis in *jupyter notebooks*



Why Python?

- Well documented!
- Easy to integrate into the rest of the analysis
- Modern and nice interface. . .

Belle II - best practice analysis code

UI

A simple example

```
import basf2
from modularAnalysis import inputMdst, reconstructDecay, fitVertex, variablesToNtuple
from stdCharged import stdPi
from stdPhotons import stdPhotons

mypath = basf2.Path()

# configure modules
inputMdst("default", basf2.find_file('analysis/tests/mdst.root'), path=mypath)
stdPi("good", path=mypath)
stdPhotons("good", path=mypath)
reconstructDecay('rho0:myrhos -> pi+:good pi-:good', '0.5 < M < 1.0', path=mypath)
fitVertex('rho0:myrhos', path=mypath)
reconstructDecay('B0:myBs -> rho0:myrhos gamma:good', '5.0 < M < 6.0', path=mypath)

# output modules
momenta = ['px', 'py', 'pz']
variablesToNtuple('B0:myBs', momenta, path=mypath)

basf2.process(mypath)
```

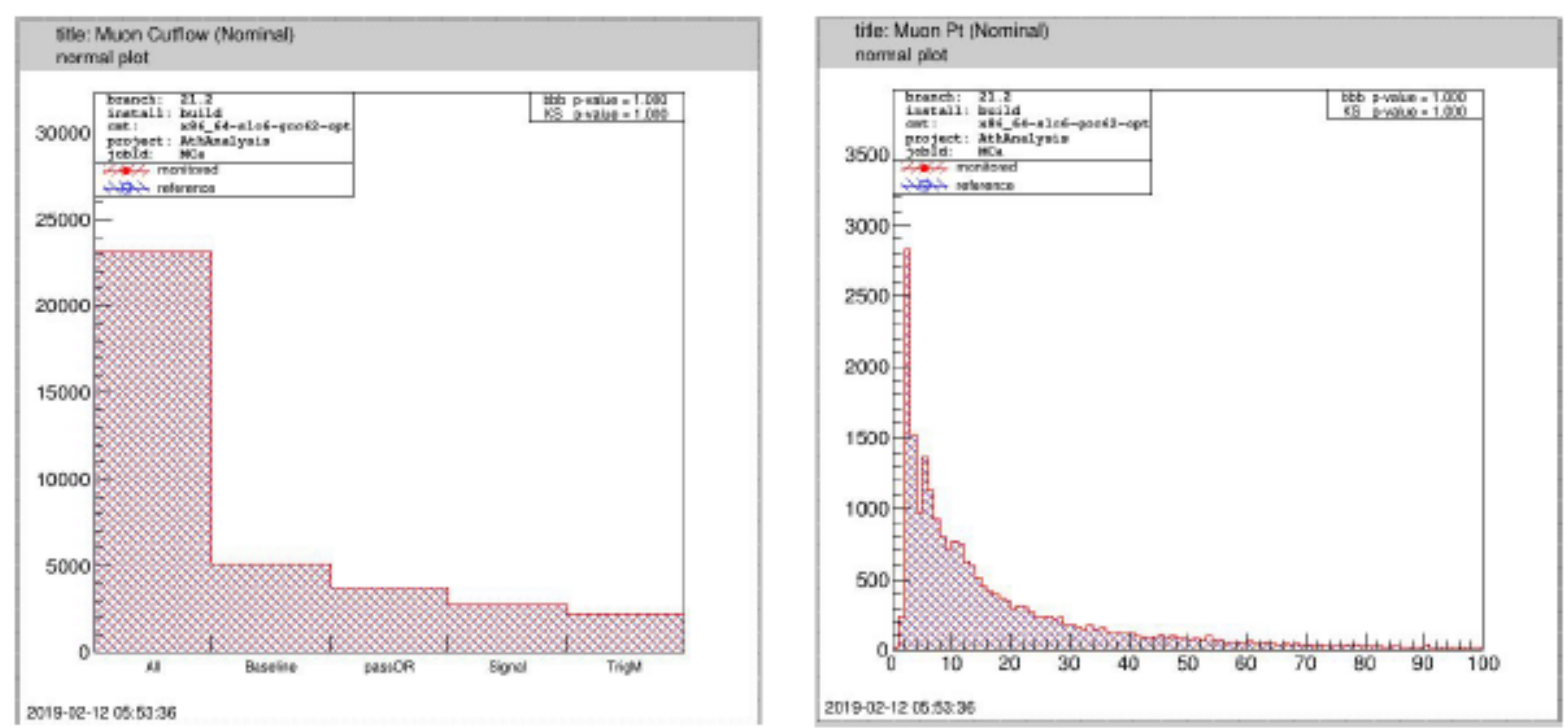
Beyond analysis functionality

SUSYTools @ Hass AbouZeid

<u>Configuration diff</u>
High-level analysis comparison

	Default	Analysis 1	Analysis 2
EleBaseline.Pt	10000.	10000.	7000.
EleBaseline.Eta	2.47	2.47	2.47
EleBaseline.Id	LooseAndBLayerLLH	LooseAndBLayerLLH	VeryLooseLLH
EleBaseline.CrackVeto	false	true	false
EleBaseline.z0	0.5	0.5	
Ele.Et	25000.	10000.	20000.
Ele.Eta	2.47	2.47	2.47
Ele.CrackVeto	false	true	false
Ele.Iso	Gradient	FCTight	Gradient

<u>CI pipeline</u>
Carefully control code

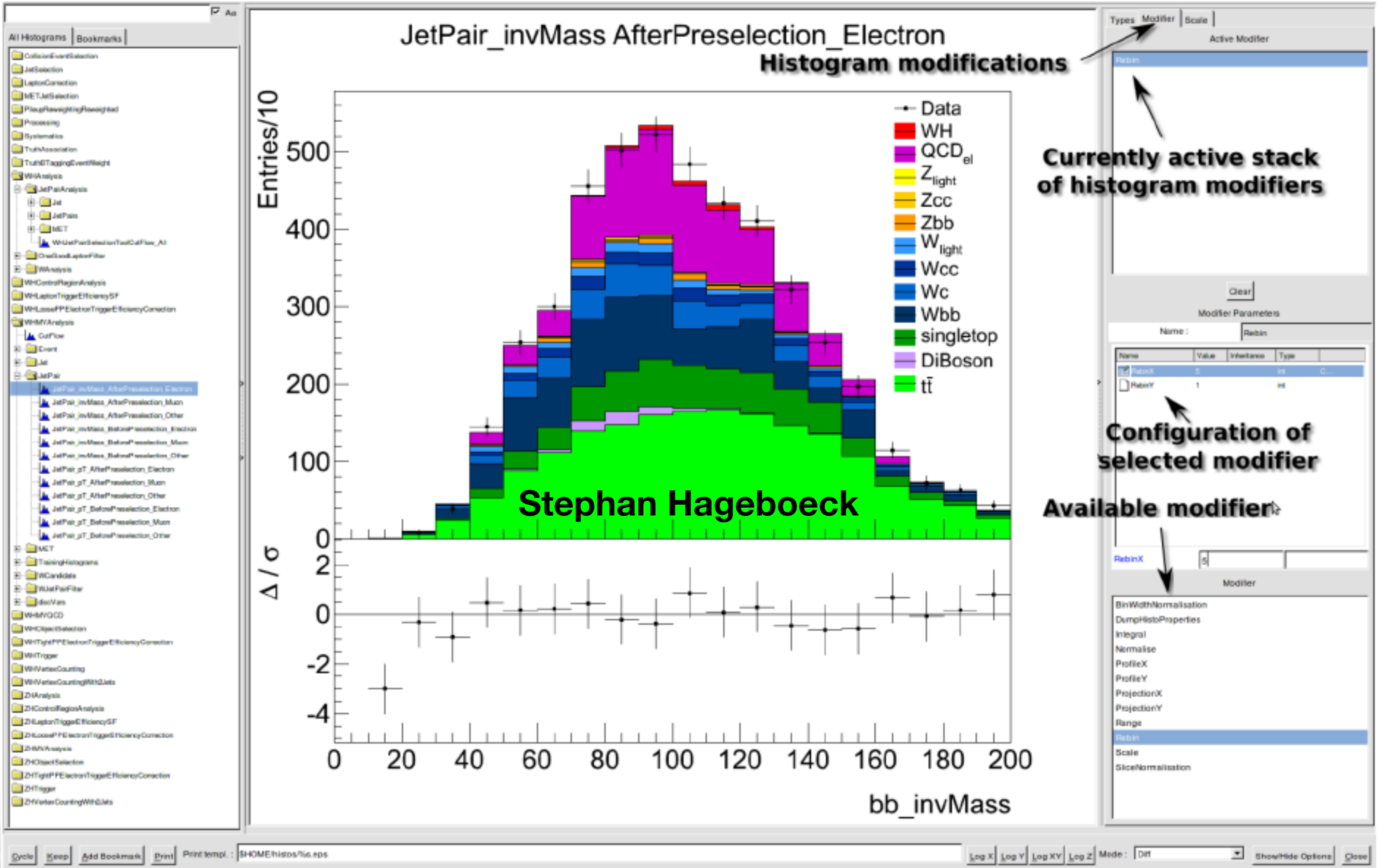


Reference histogram updated automatically every night

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

GUI Overkill



U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

Observations / Questions

- Analysis is diverse, but we see recurring themes and solutions

- **Reducing I/O for heavy lifting:**
- ***Trains*** an accepted solution, can more workflows use this concept?
- ***Common nano-AOD*** centrally produced, less reinventing the wheel on format
  - **Q**. How much bandwidth can go this route?
- ***Turbo stream*** calibrate once
  - **Q.** How much bandwidth can go this route? How strong is the physics case to justify not doing that?
- Convergence on ***Jupyter*** notebooks as analysis platform, ***hiding the how is good***

- Trend towards ***declarative analysis***, especially for **LHCb/Belle II**
  - **Does anything prevent other experiments?**

- Addressing ***systematics*** is still a challenge, see **Andrea's** talk
  - Can we attack (some of) this as a community? What is best practice?
  - Not covered - ***Monte Carlo***

But first, the next talk

Winter is comin'