

Offline Computing

Emma Torró Pastor

IFIC - Valencia

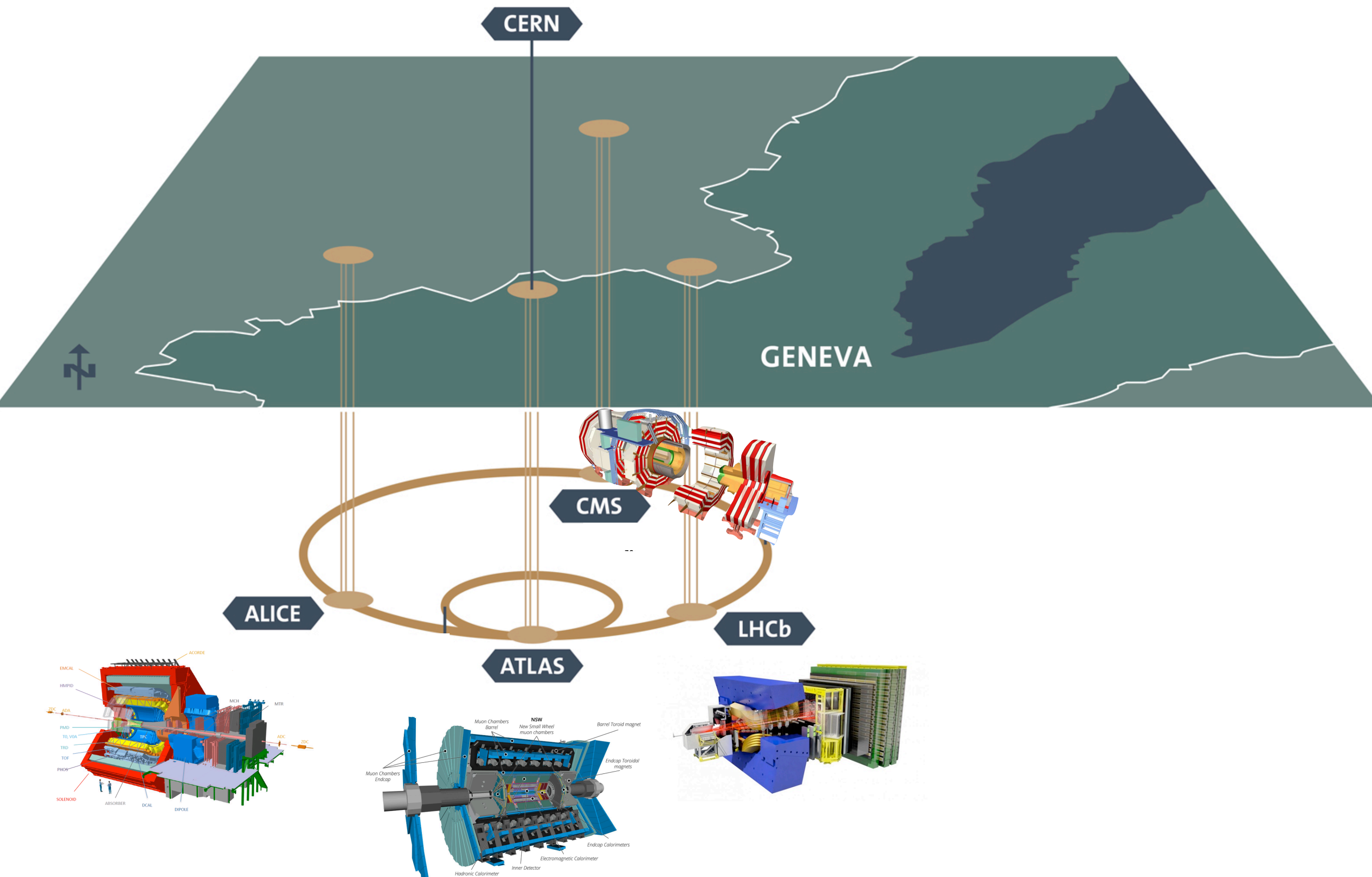
COMCHA School

Zaragoza, April 15th, 2026



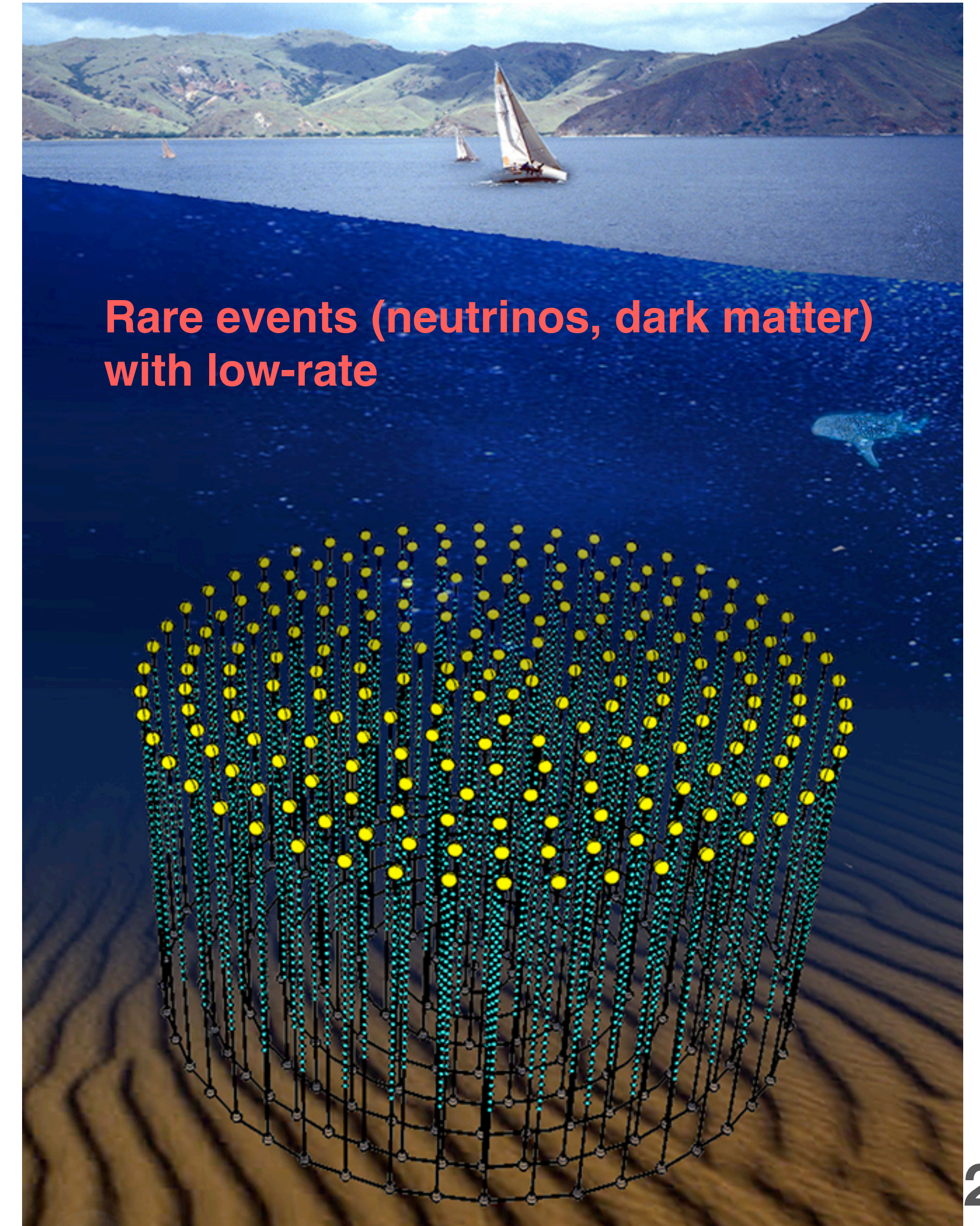
HEP vs. Astrophysics data challenges

LHC experiments



Millions of collisions/sec

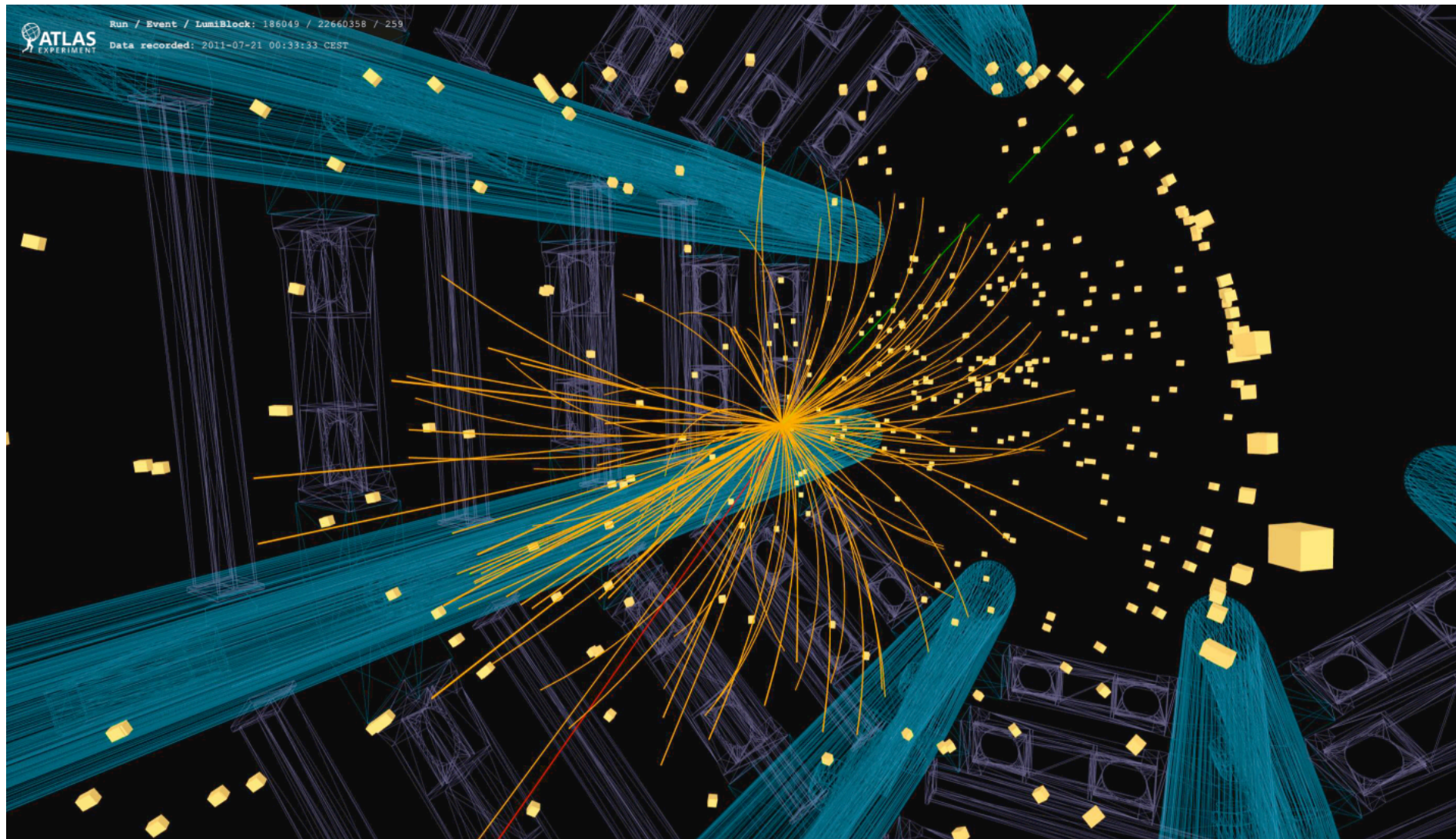
Astrophysics



Two regimes...

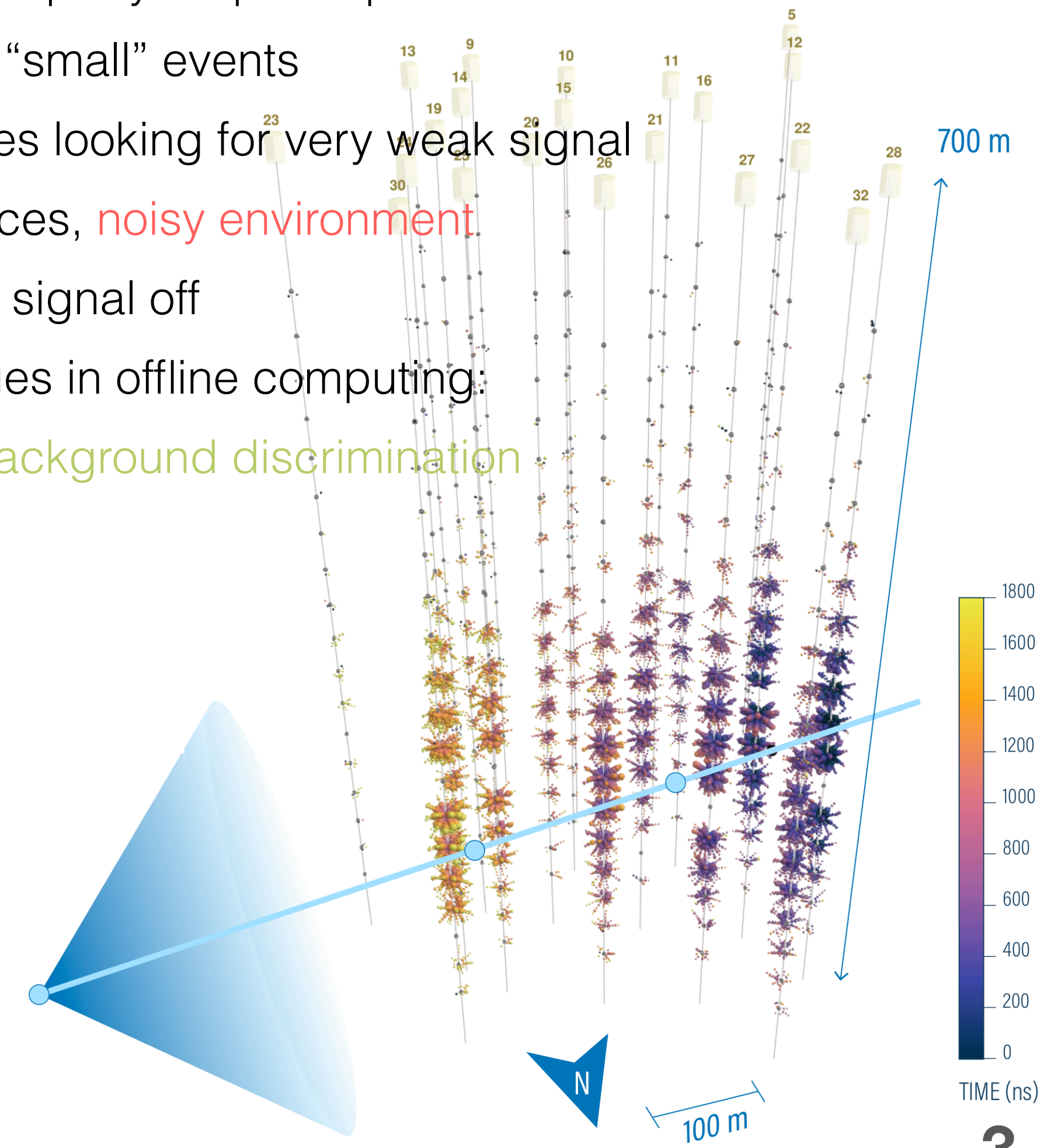
LHC experiments

- Huge datasets collected
 - Millions of collisions/sec
 - High rate, very crowded events, O(PB) per year per experiment
 - High-energy objects
- **Controlled environment**: SM backgrounds + detector backgrounds
 - Can turn signal off to measure detector backgrounds
- Big challenges in offline computing:
 - Signal / background discrimination
 - Storage
 - Fast analysis

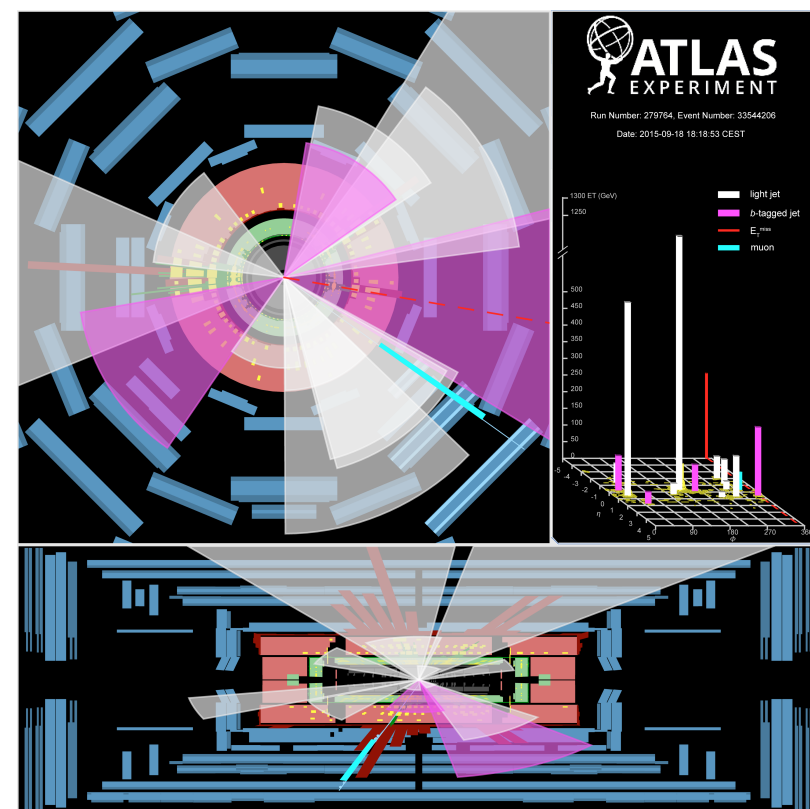


Astrophysics

- Rare events (neutrinos, dark matter)
 - Few events per year per experiment
 - Low-rate, “small” events
 - Some times looking for very weak signal
- Natural sources, **noisy environment**
 - Can't turn signal off
- Big challenges in offline computing:
 - Signal / background discrimination



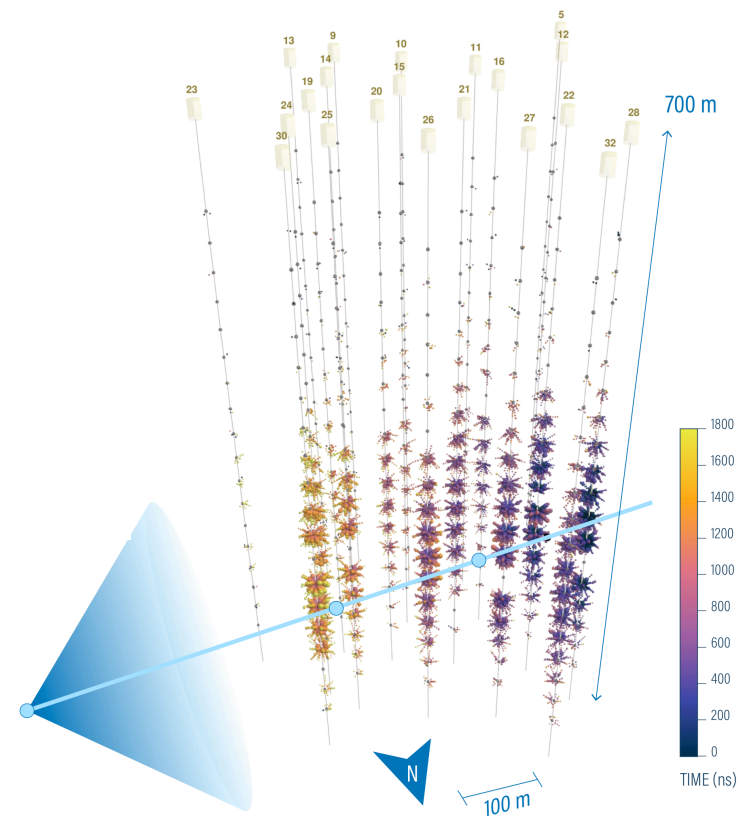
Two regimes, one computing pipeline



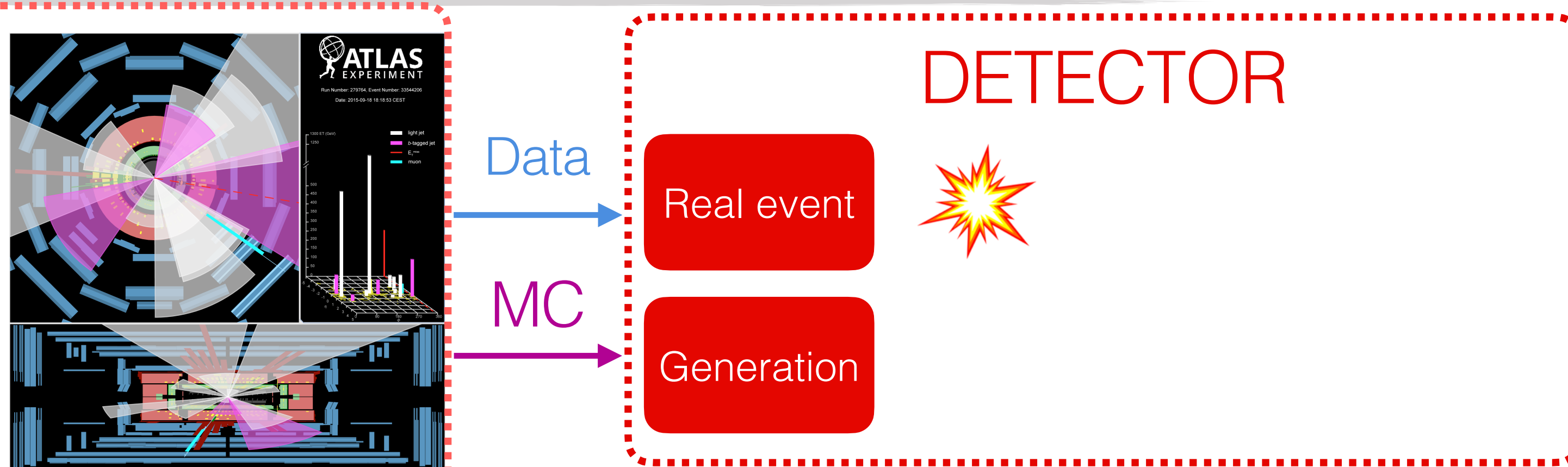
Data

MC

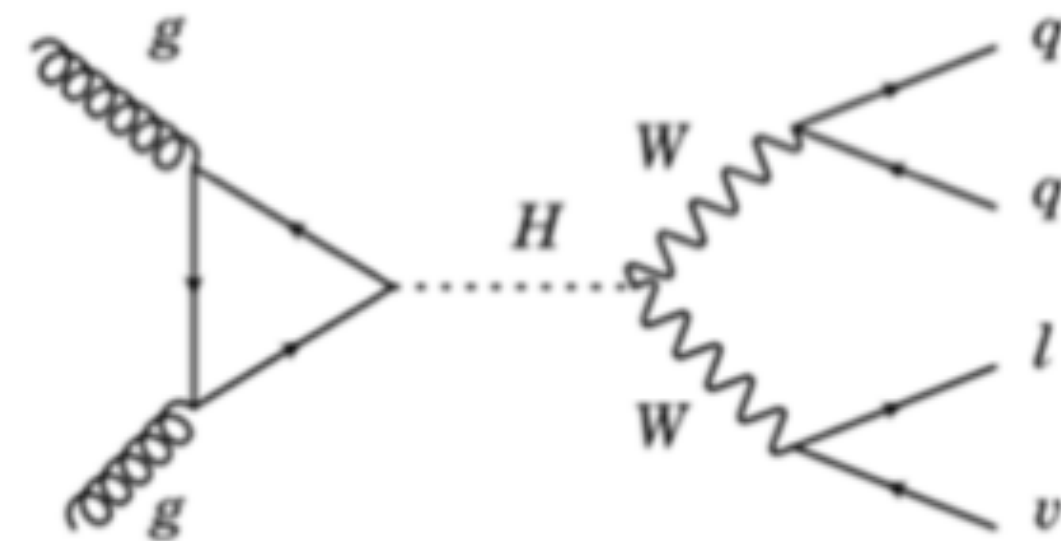
DETECTOR



Two regimes, one computing pipeline

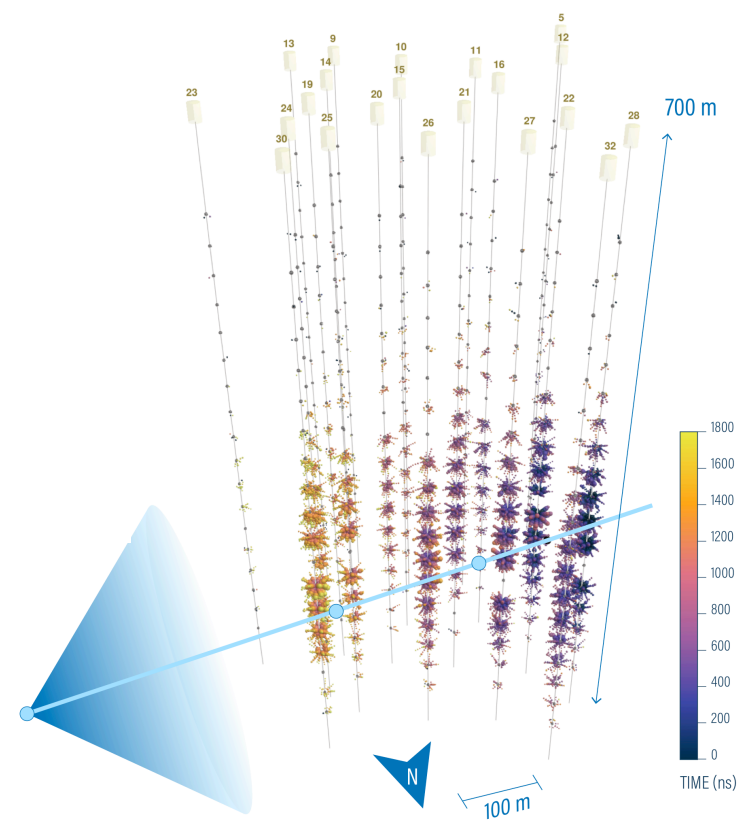
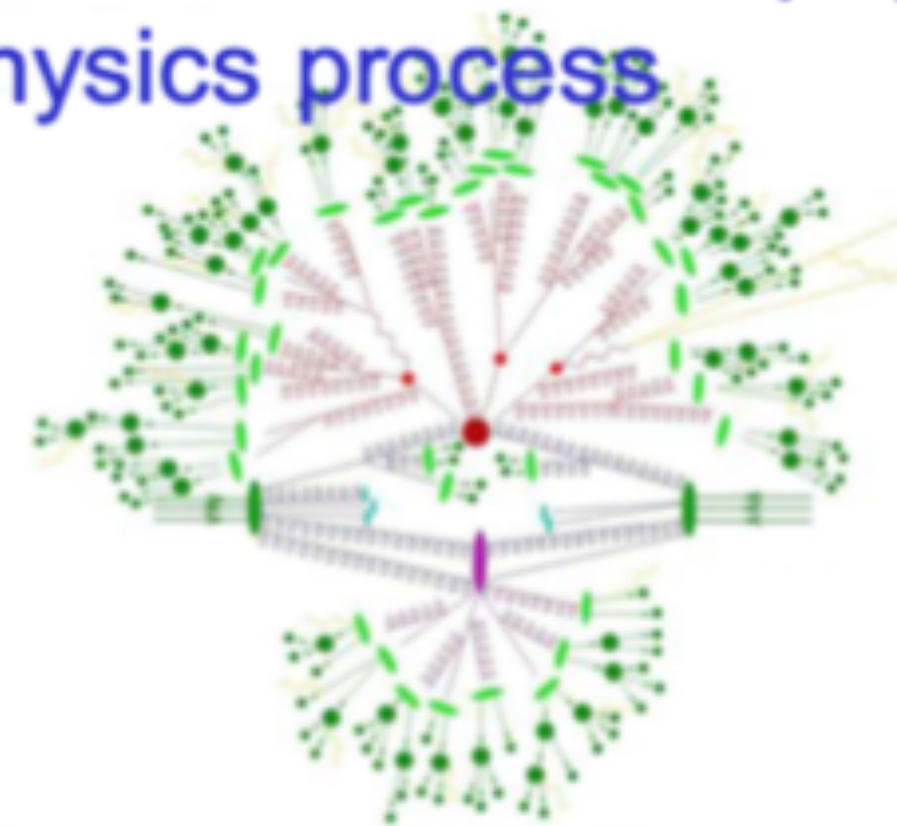


Simulation of high-energy
physics process

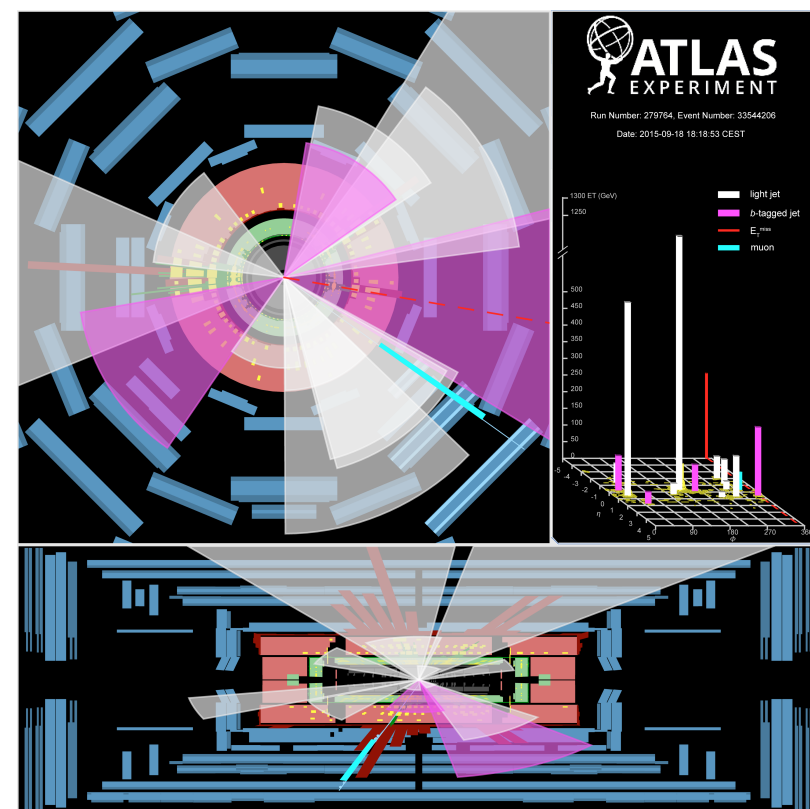


Pythia, Sherpa,
MadGraph, ...

Simulation of 'soft physics'
physics process



Two regimes, one computing pipeline



Data

MC

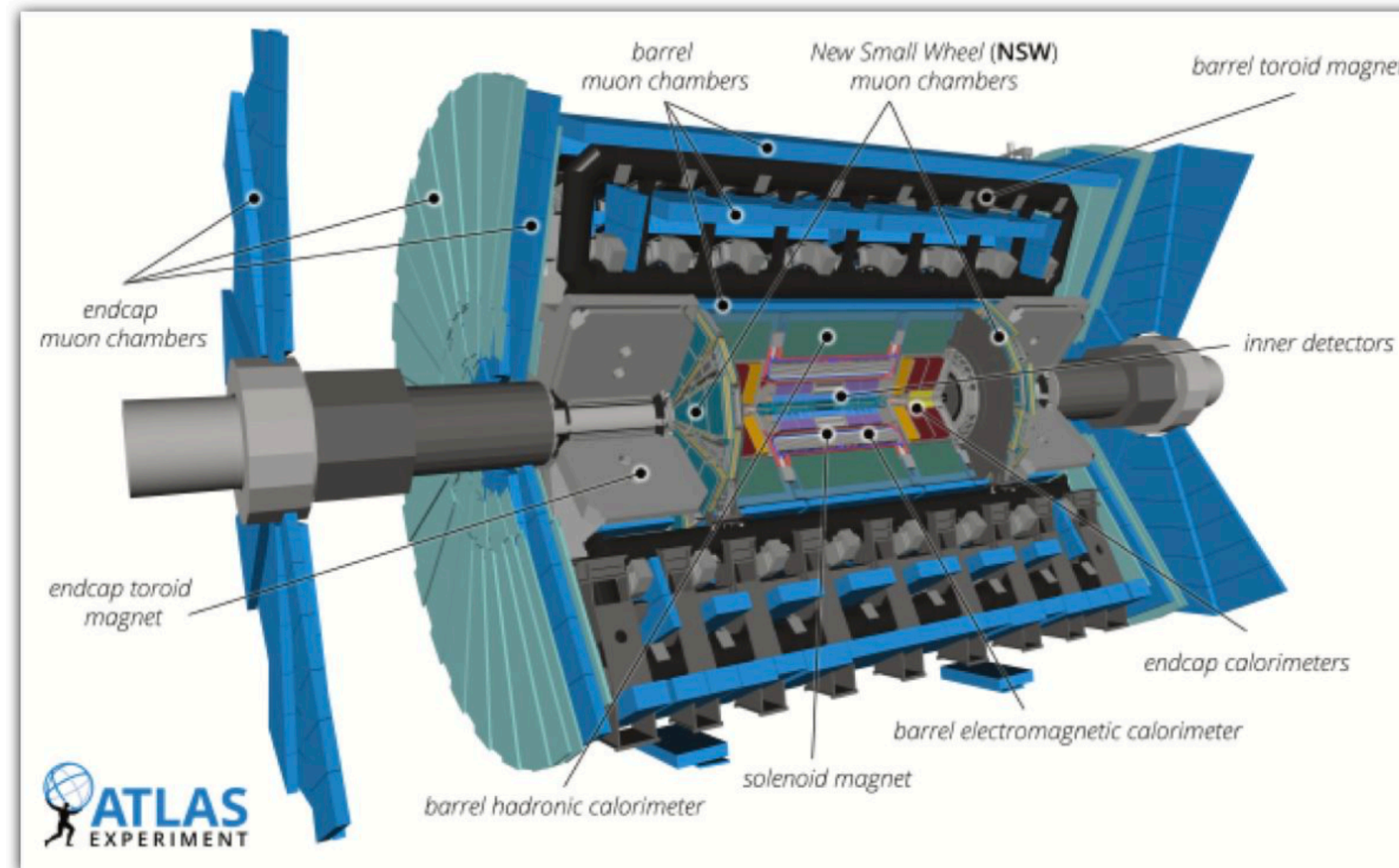
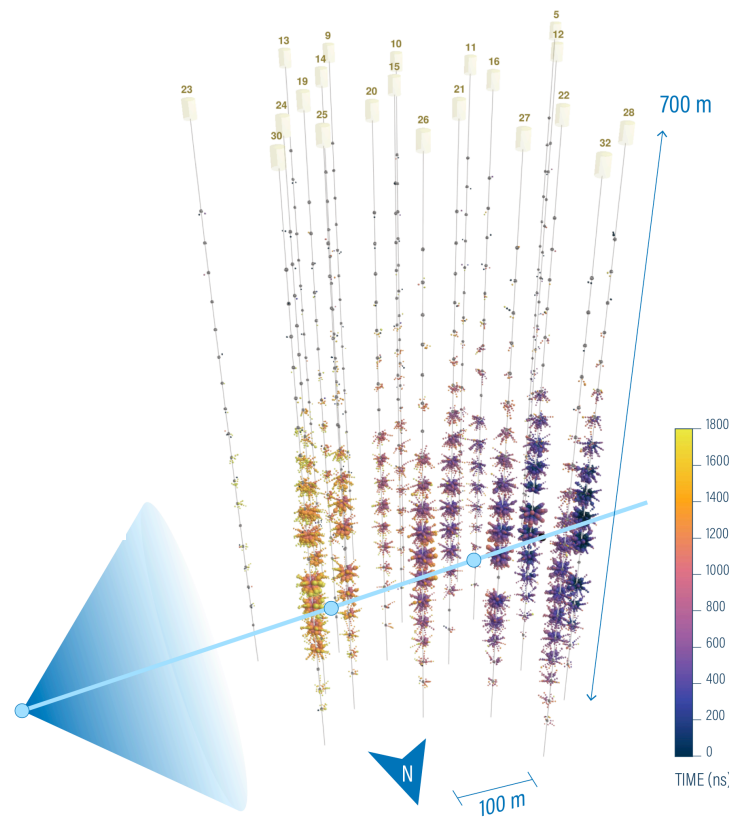
DETECTOR

Real event

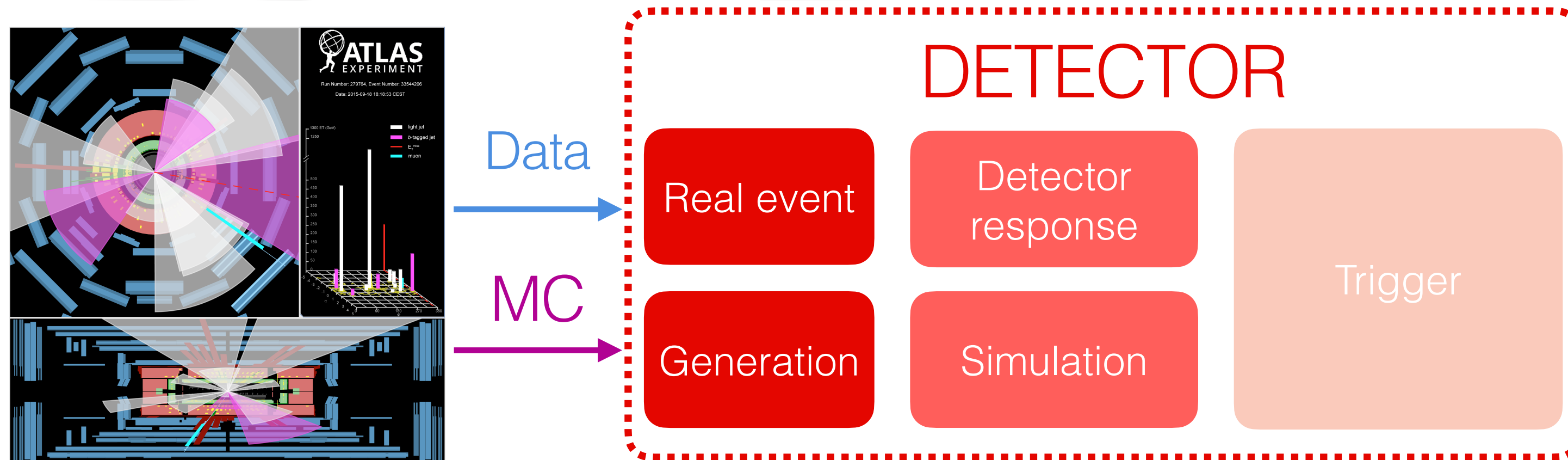
Detector response

Generation

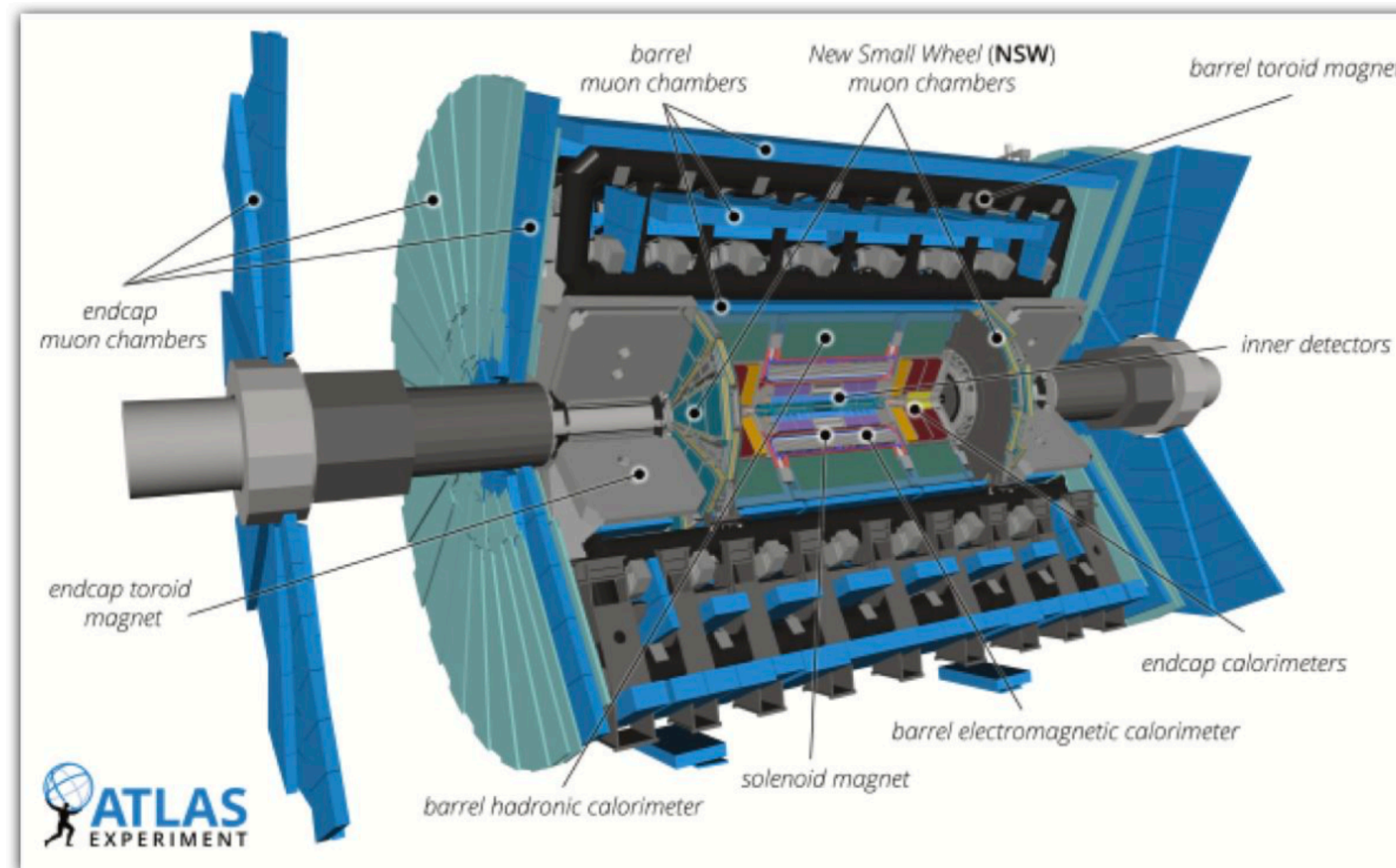
Simulation



Two regimes, one computing pipeline



- Data AcQuisition
 - Getting data from the detector
- Trigger
 - Filtering which of those data are useful and keeping only that



ATLAS detector is big!

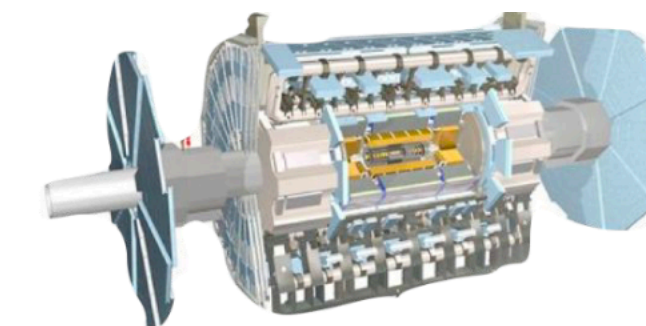
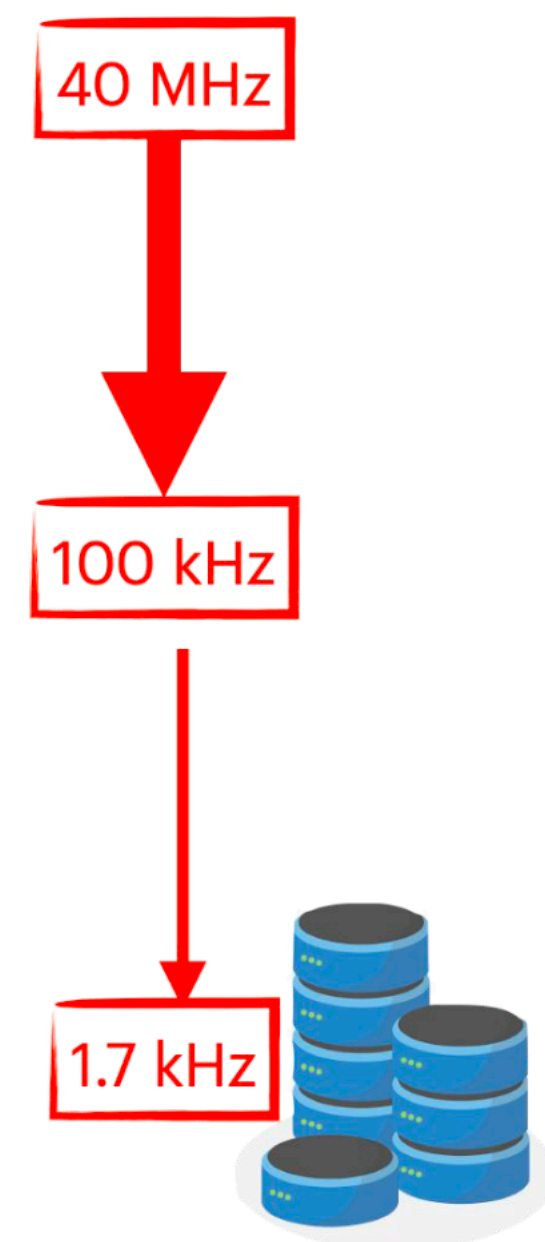
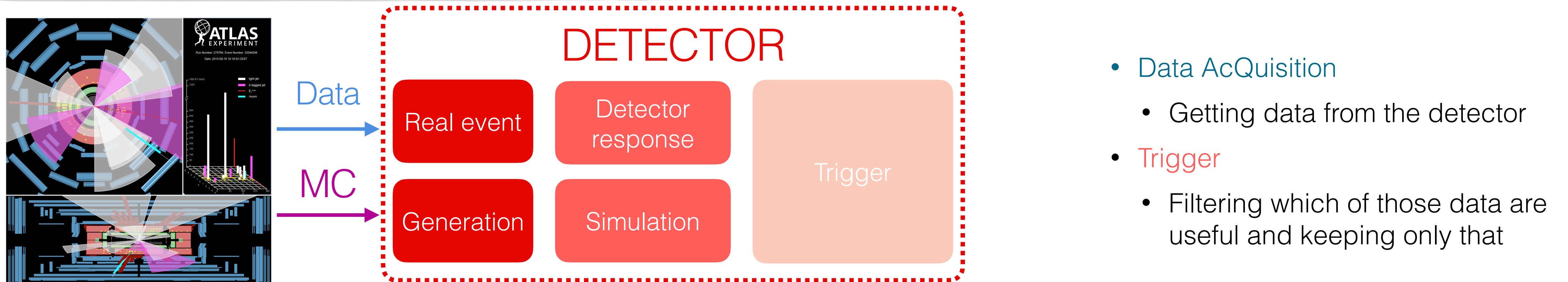
Millions of channels in Inner Detector, Calorimeter, Muon Spectrometer add up to 2 MB/event digital data

LHC collides at 40 MHz

80 TB/s to record everything
3h to fill 1 exabyte ~Google data centre?

<https://what-if.xkcd.com/63/>

Two regimes, one computing pipeline



Level-1 Trigger (L1)

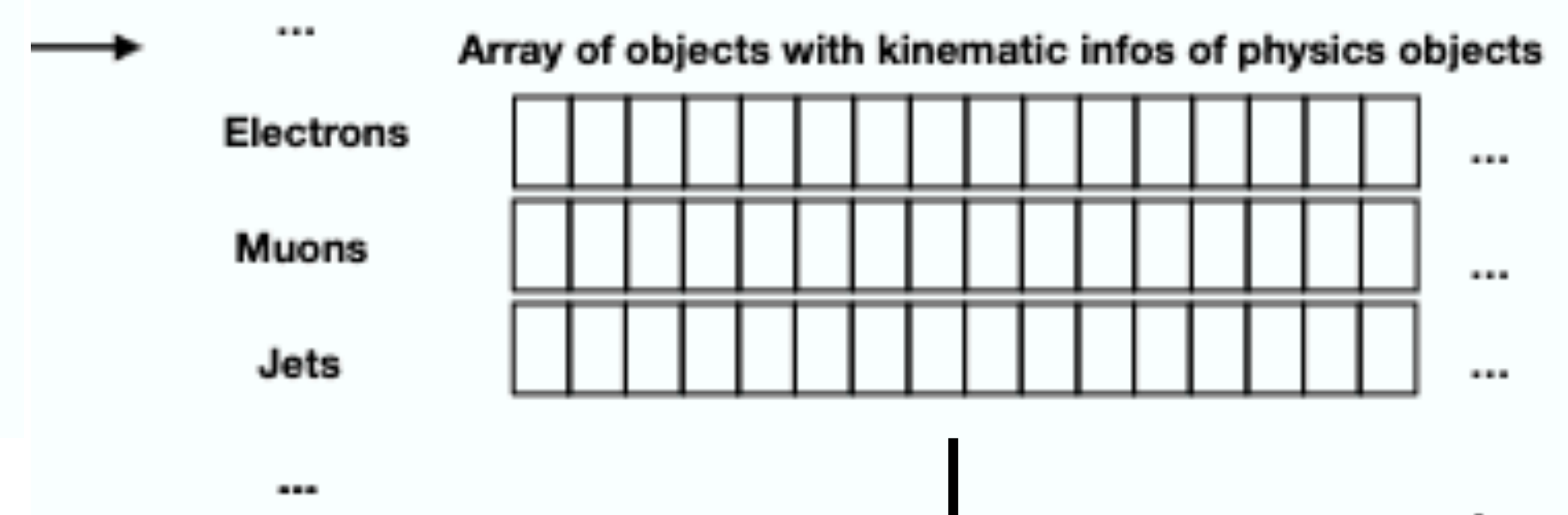
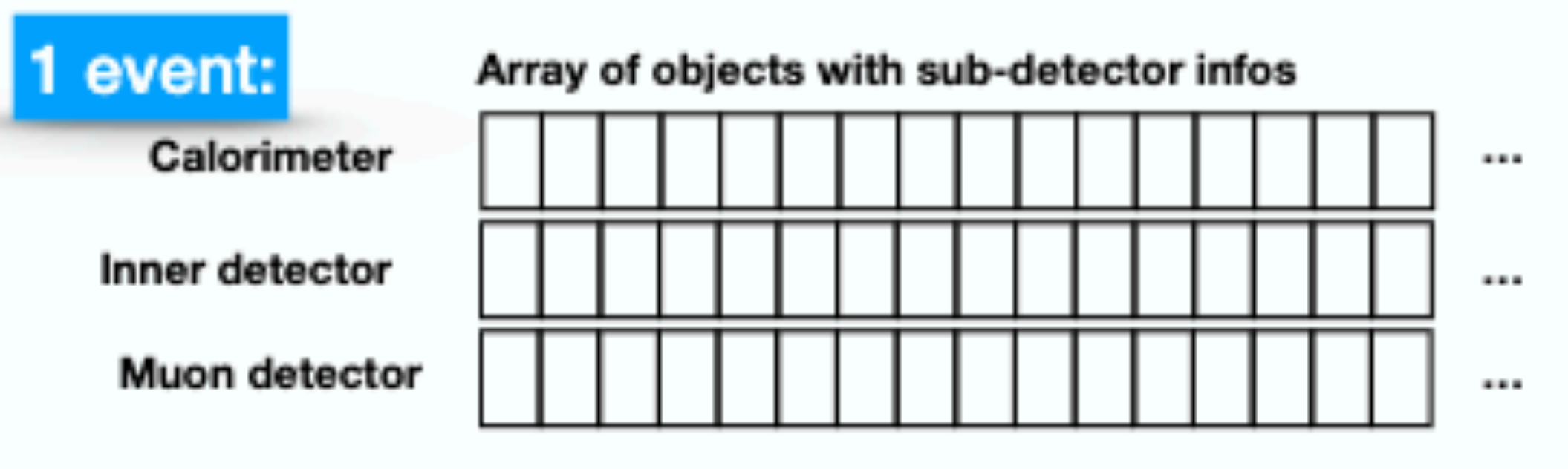
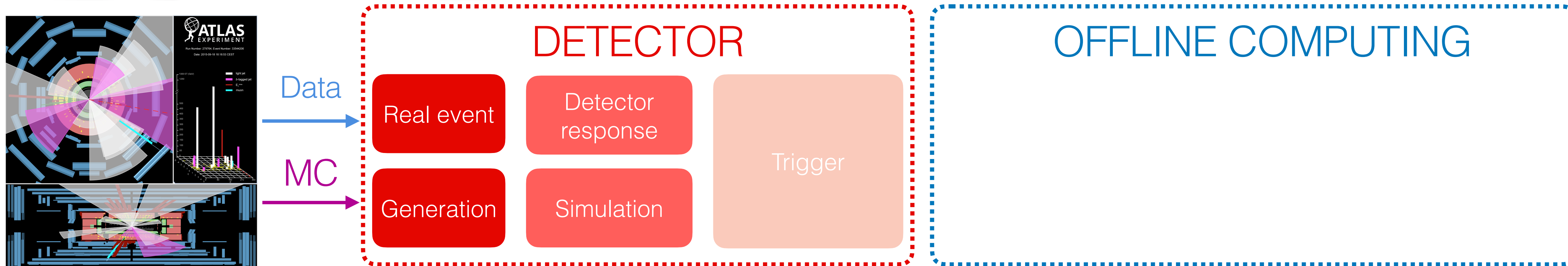
- Hardware-based, fast, coarse, no tracking

Can provide “Regions-Of-Interest” (RoI) for reconstruction & selection on only limited areas of detector by HLT

High Level Trigger (HLT)

- Software-based, sophisticated, full granularity, (limited) tracking

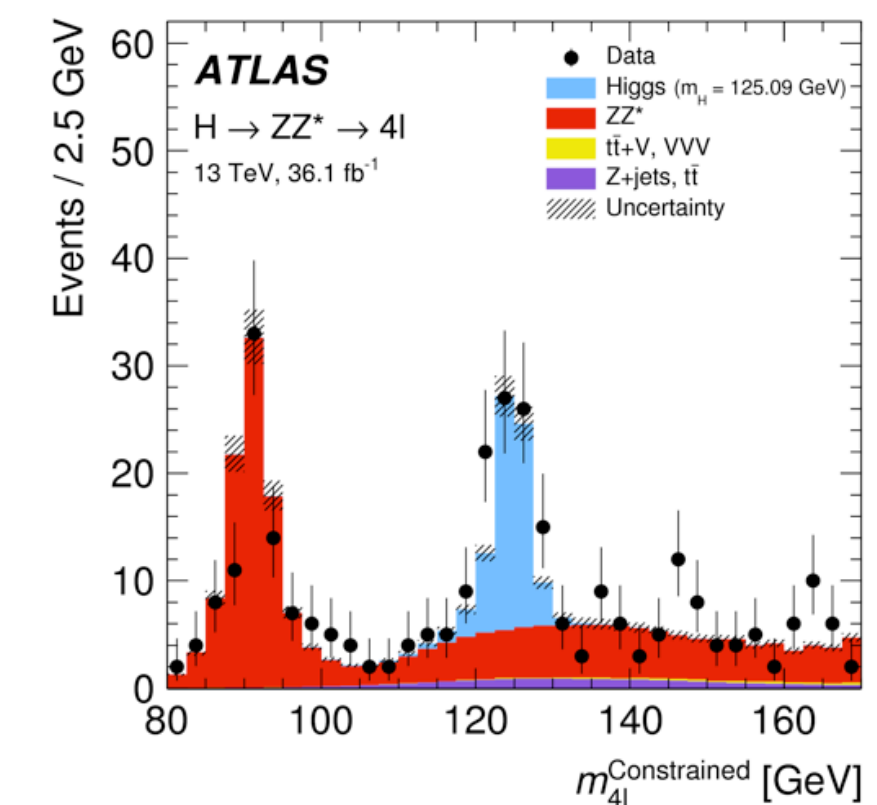
Offline computing



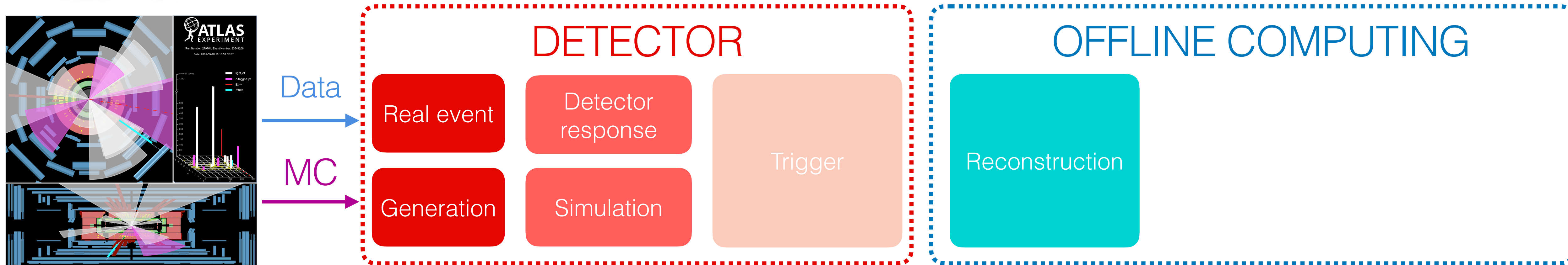
Offline Computing

Where raw detector signals become physics objects like **electrons, muons, jets**

1. Reconstruction (tracks, jets, missing energy)
2. Calibration
3. Data reduction
4. Analysis



Offline computing

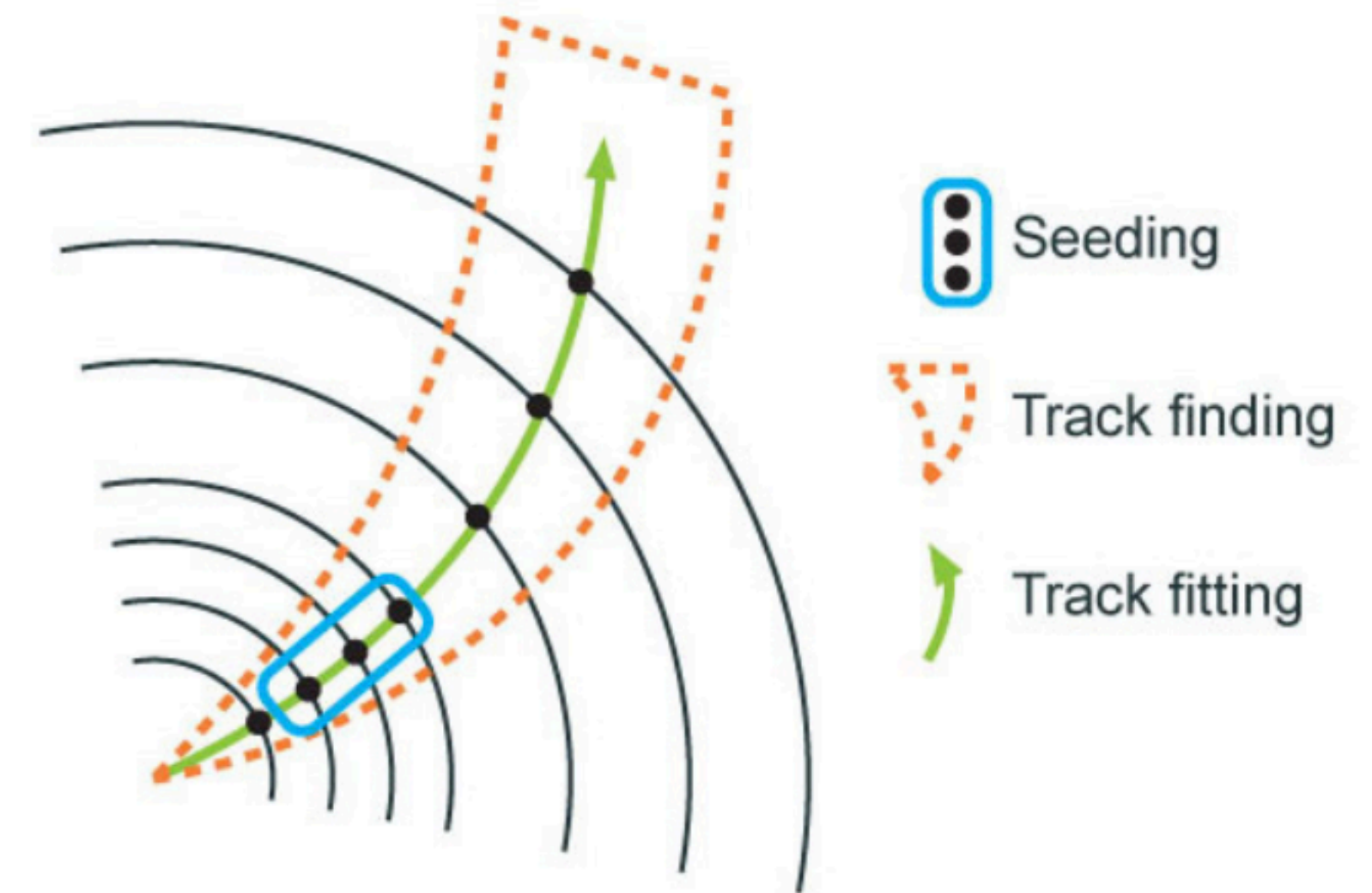


Convert raw detector signals into high-level physical objects (tracks, vertices, electrons, muons, photons, jets)

Inner Detector Tracking: Reconstructs charged particle paths using silicon pixels, strips, and straw tubes.

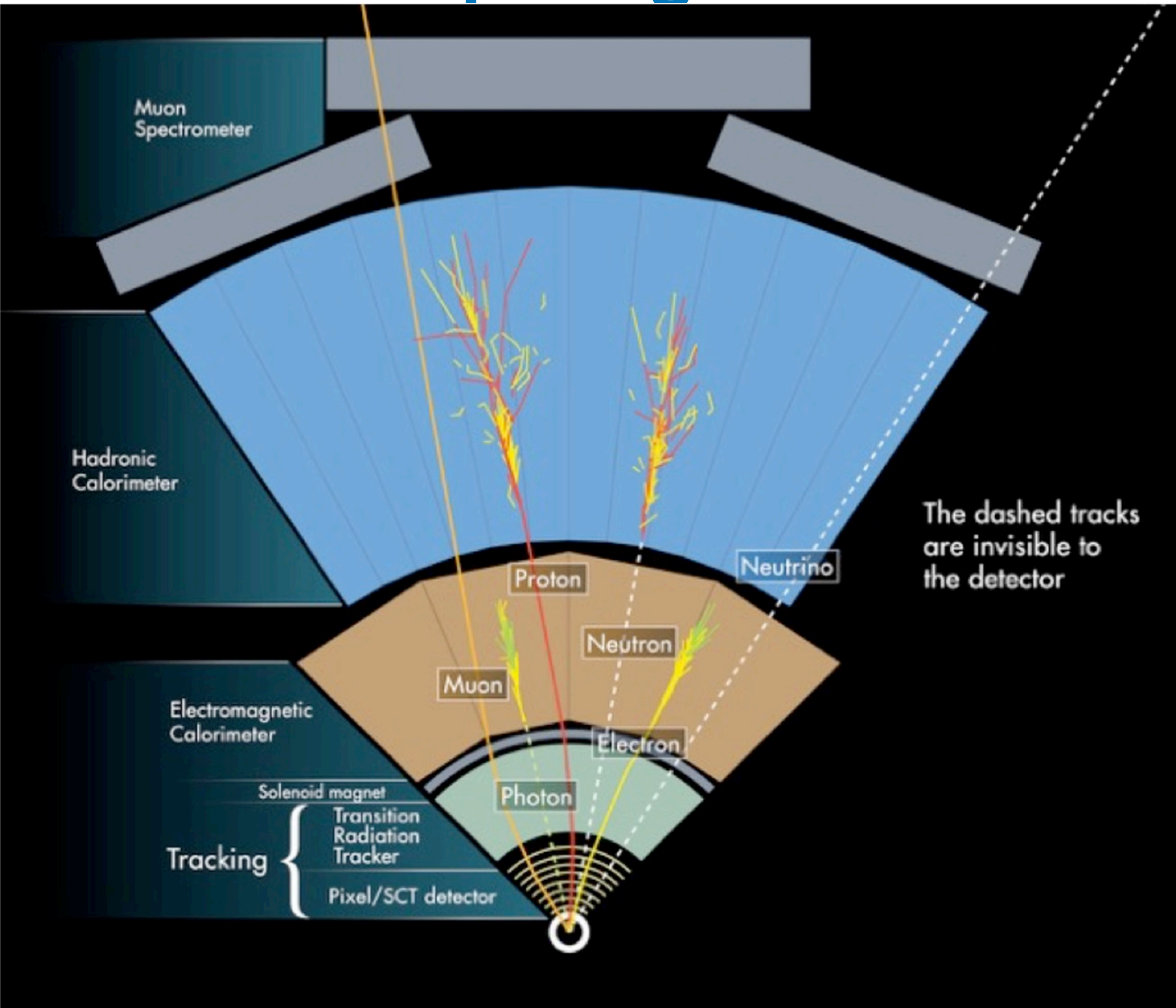
Calorimeter Clustering: group 3D calorimeter cells based on energy significance, helping to reconstruct particles like photons and jets.

Muon Reconstruction: Combines information from the Inner Detector and the Muon Spectrometer to identify high-purity muon candidates.



Using complex software algorithms.

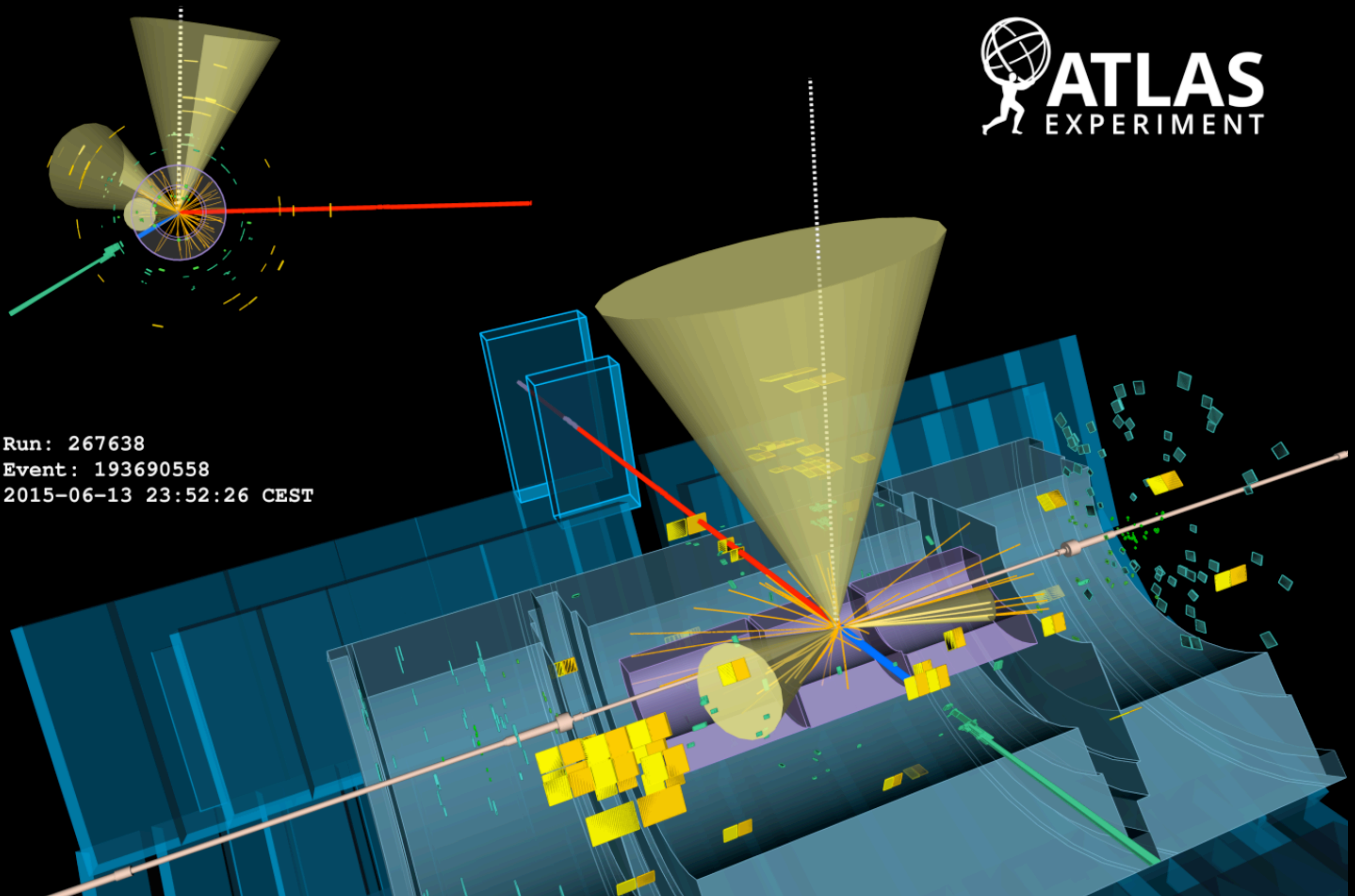
Offline computing



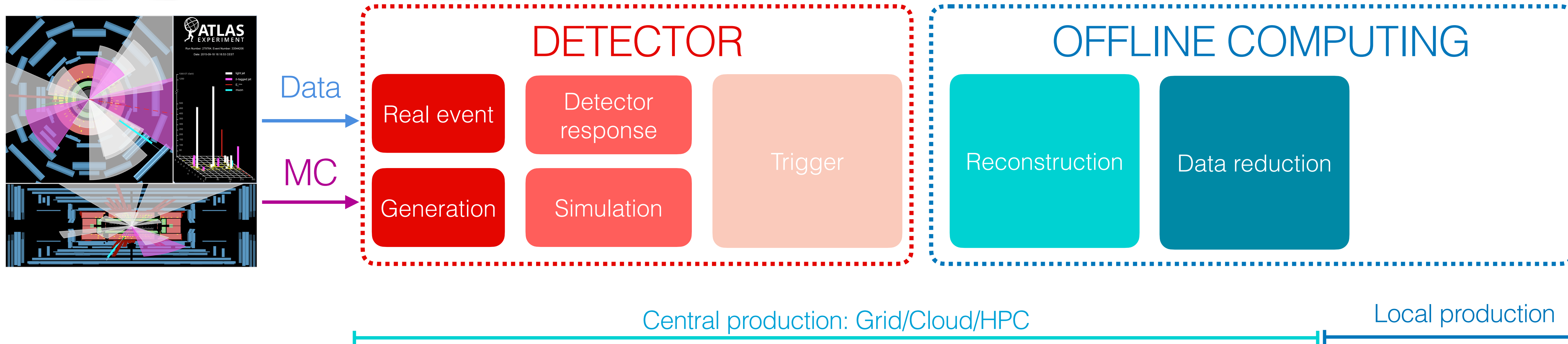
OFFLINE COMPUTING

Reconstruction

Run: 267638
Event: 193690558
2015-06-13 23:52:26 CEST

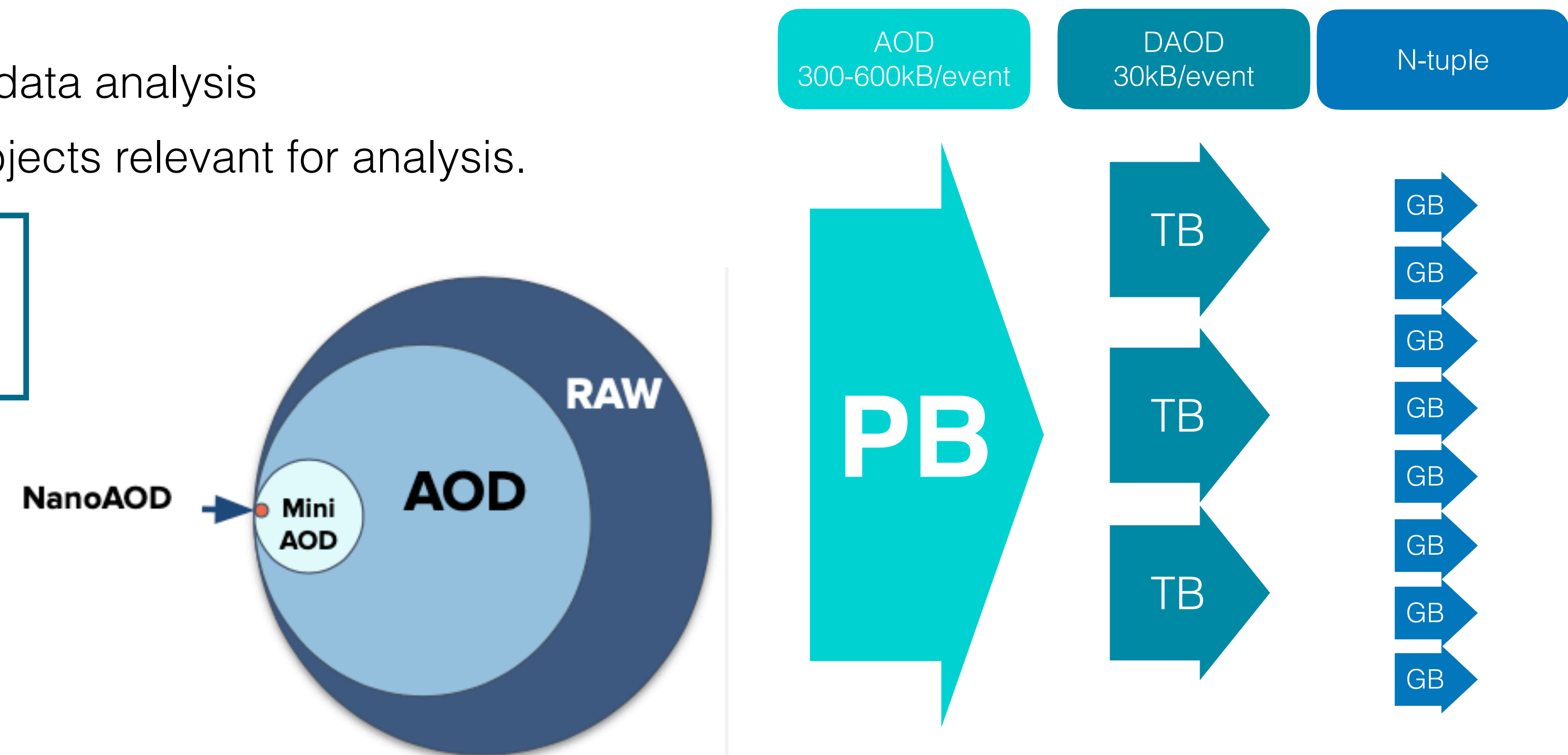


Data reduction



- Reducing datasets sizes to facilitate data analysis
- Selecting subsets of events/tracks/objects relevant for analysis.

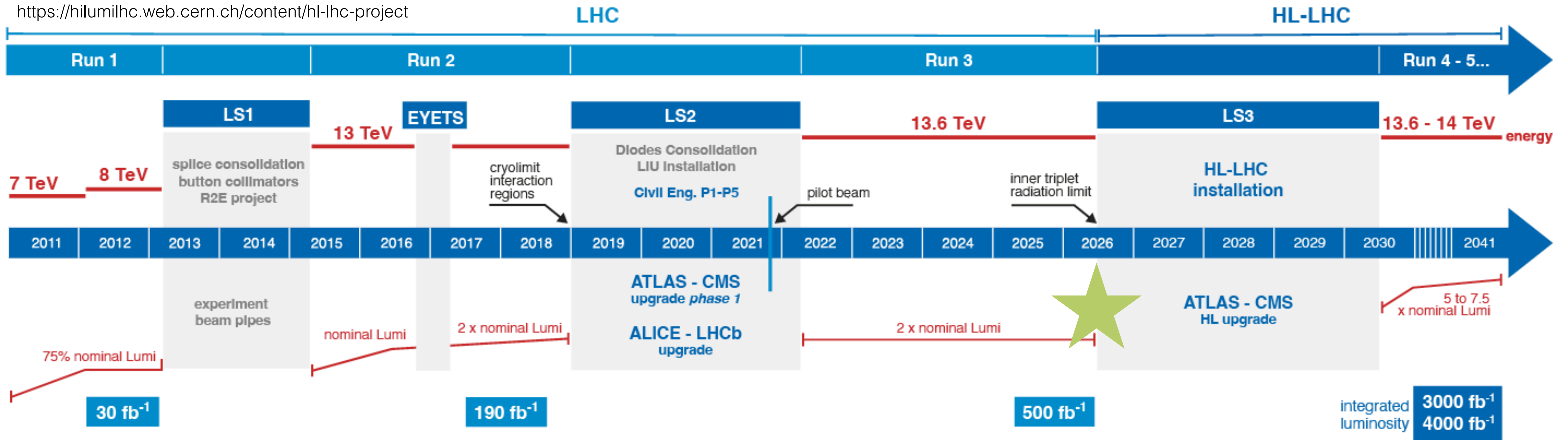
• **Challenge:** keeping all interesting physics, scalable, possibly in real-time or quasi-real-time.



Challenges for the HL-LHC

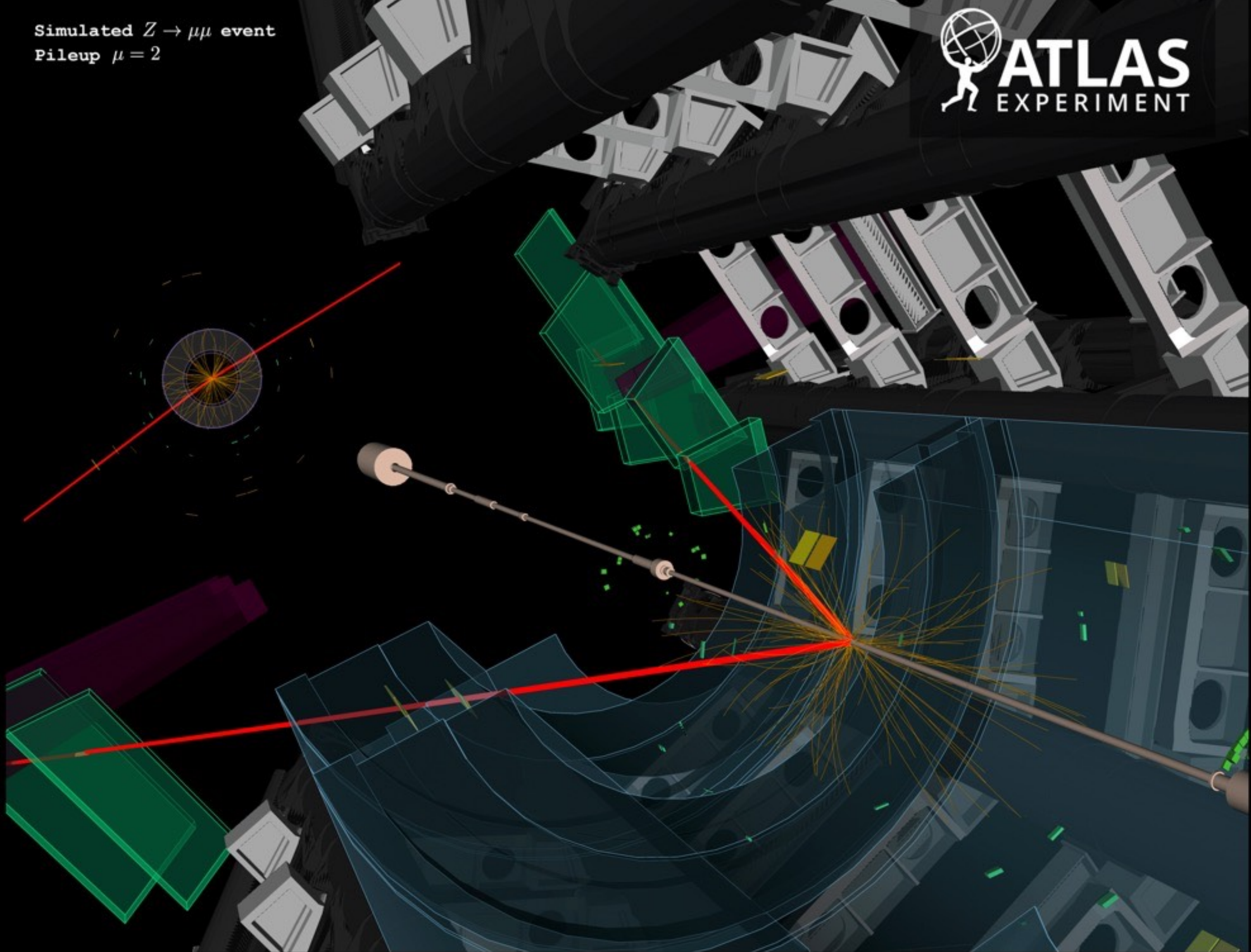


<https://hilumilhc.web.cern.ch/content/hl-lhc-project>

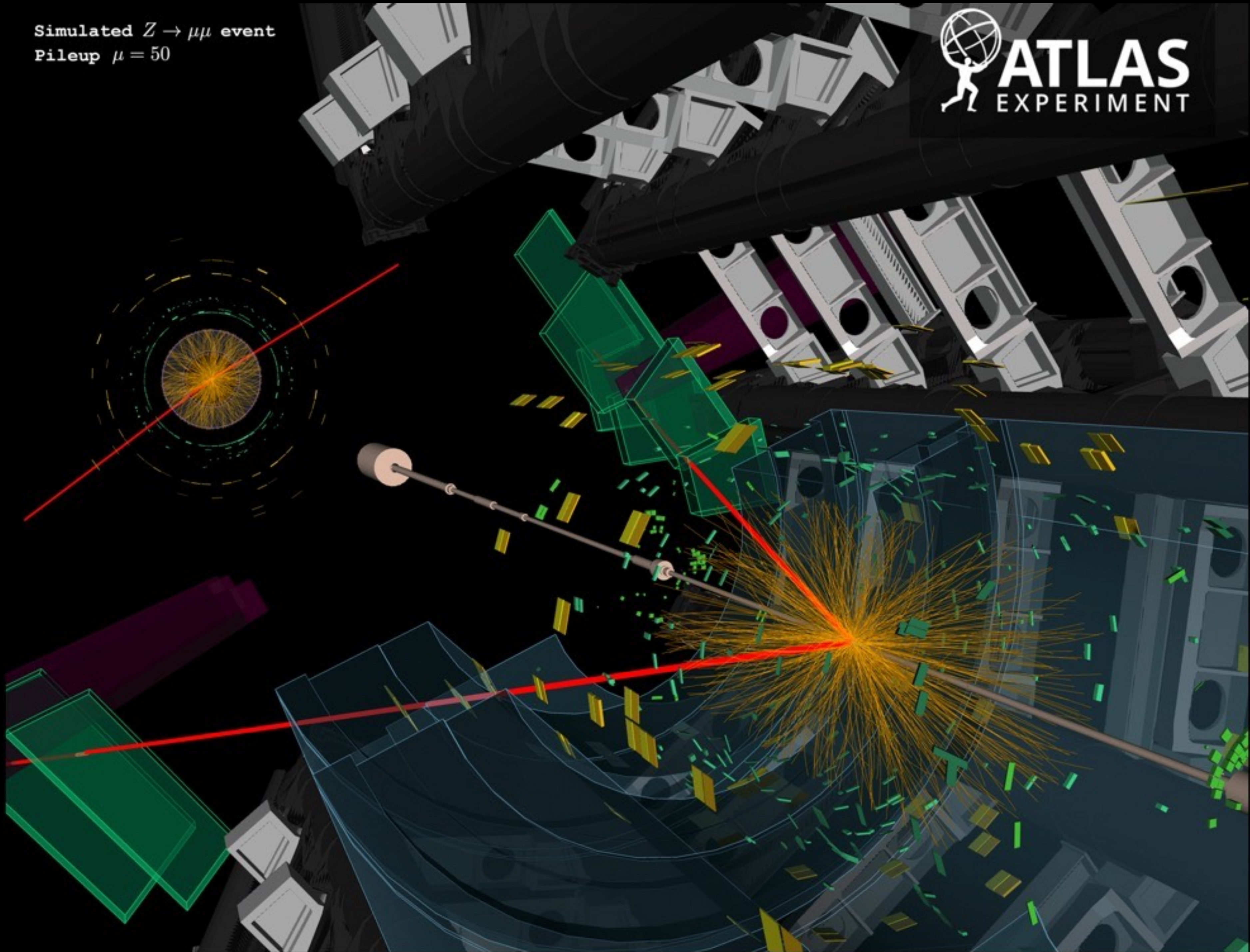


- HL-LHC will start in 2030. Planning to collect 3000/fb data
 - Massive increase in data rates (up to 10x Run 2 levels)
 - Improved detector granularity and upgrades
 - High pileup (ATLAS and CMS): 132–200 interactions per crossing
- much larger event sizes

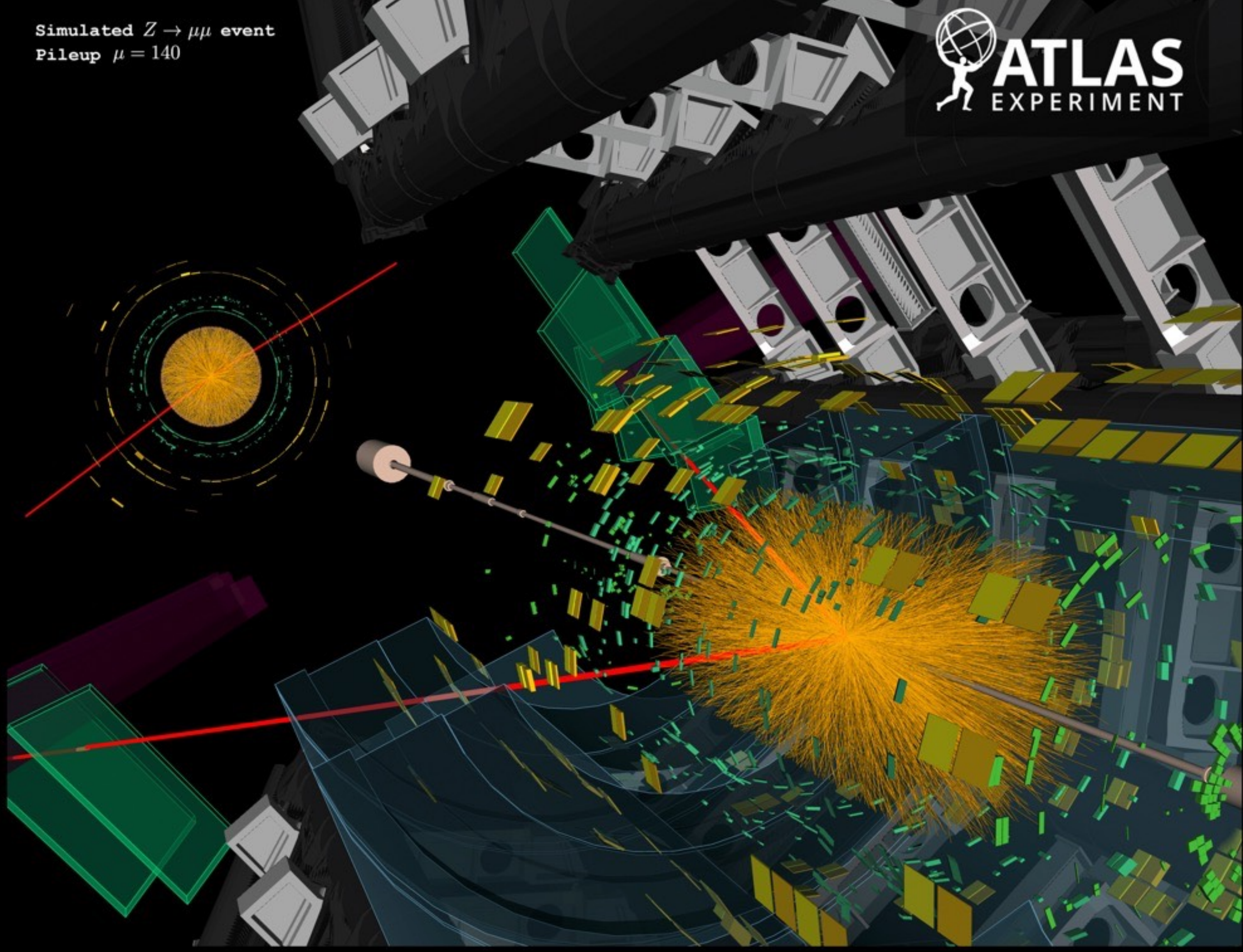
Simulated $Z \rightarrow \mu\mu$ event
Pileup $\mu = 2$

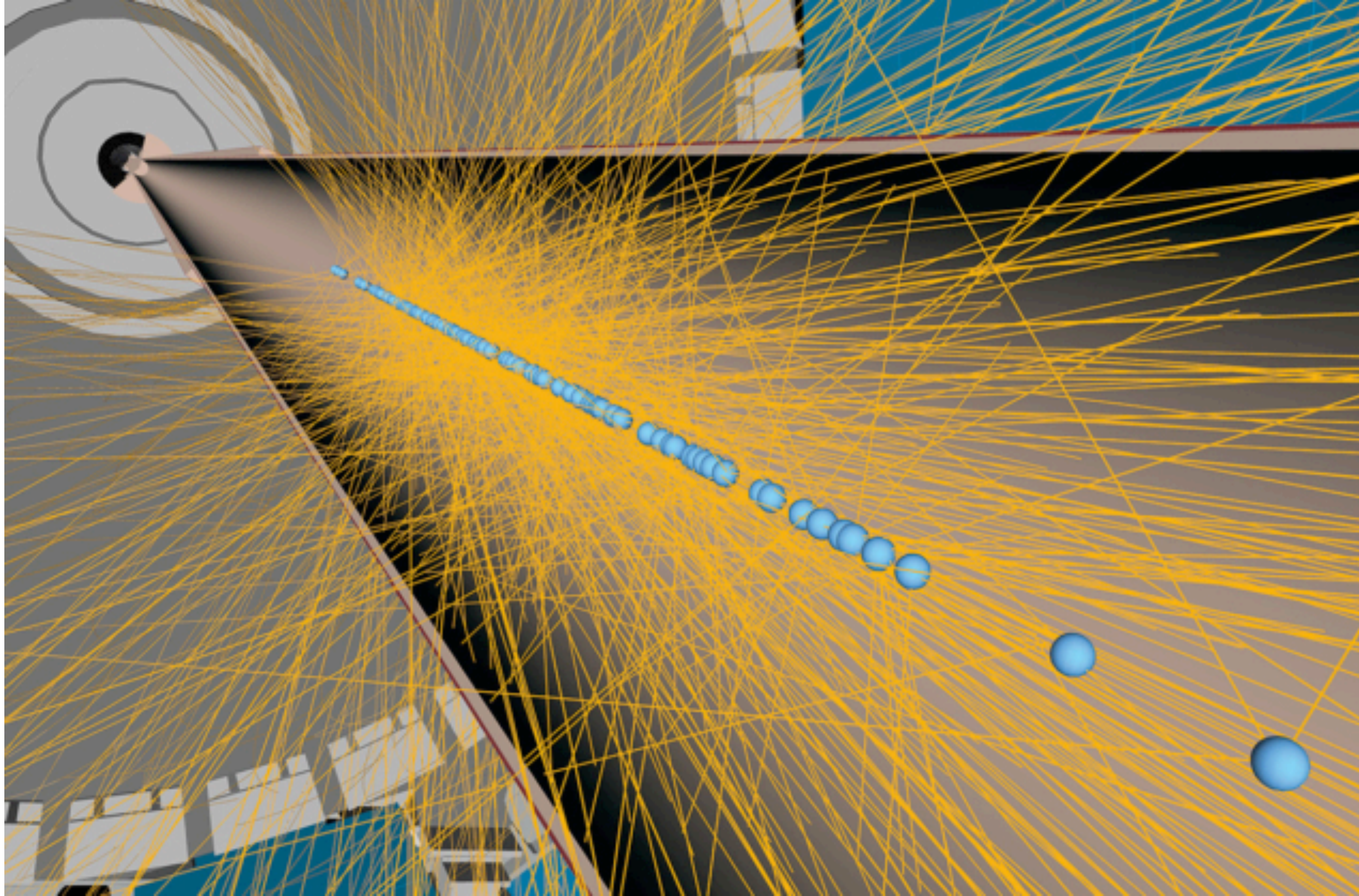


Simulated $Z \rightarrow \mu\mu$ event
Pileup $\mu = 50$



Simulated $Z \rightarrow \mu\mu$ event
Pileup $\mu = 140$

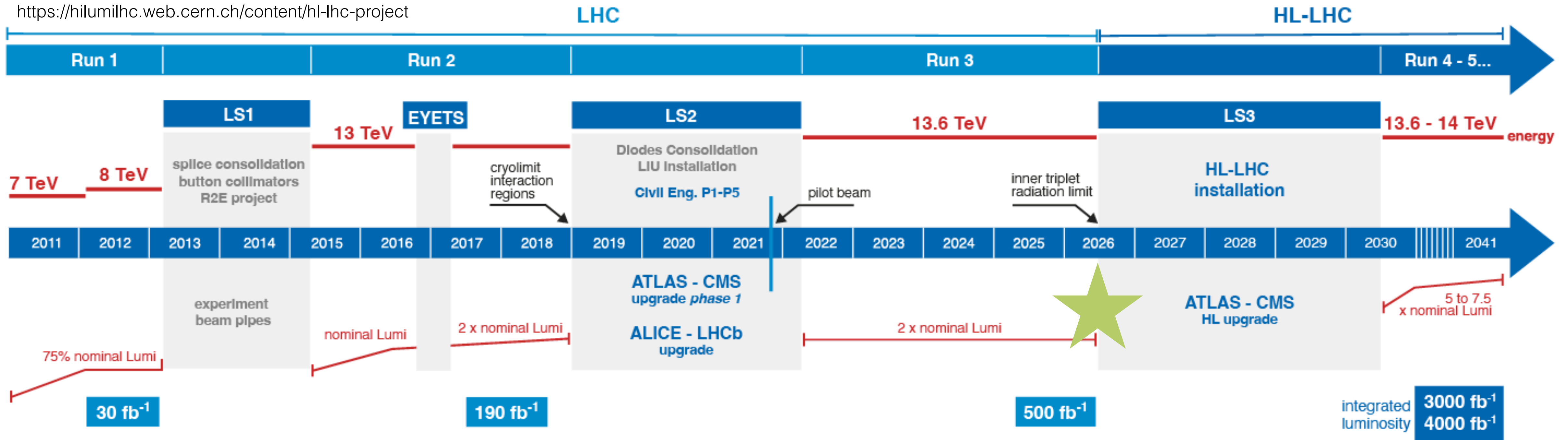




Challenges for the HL-LHC



<https://hilumilhc.web.cern.ch/content/hl-lhc-project>

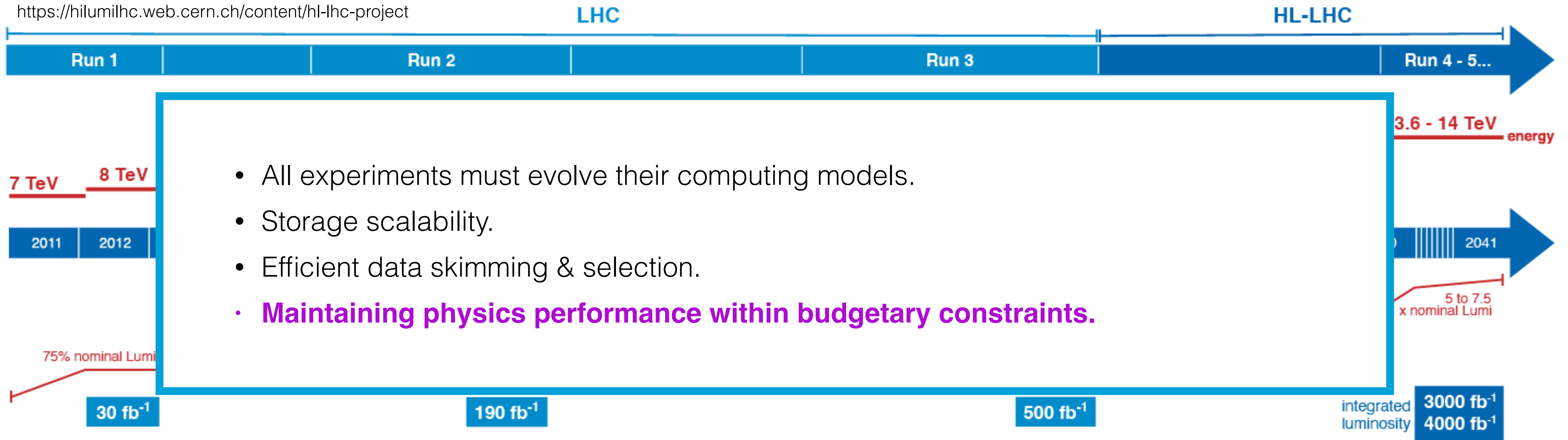


- HL-LHC will start in 2030. Planning to collect 3000/fb data
 - Massive increase in data rates (up to 10x Run 2 levels)
 - Improved detector granularity and upgrades
 - High pileup (ATLAS and CMS): 132–200 interactions per crossing
 - Advanced trigger systems or trigger-less readout mean higher event recording rates.
 - Long operational period: ~10 years of data taking.
 - Massive Monte Carlo simulations required for physics analyses and systematics studies.
-] much larger event sizes

Challenges for the HL-LHC



<https://hilumihc.web.cern.ch/content/hl-lhc-project>



- HL-LHC will start in 2030. Planning to collect 3000/fb data
 - Massive increase in data rates (up to 10x Run 2 levels)
 - Improved detector granularity and upgrades
 - High pileup (ATLAS and CMS): 132–200 interactions per crossing
 - Advanced trigger systems or trigger-less readout mean higher event recording rates.
 - Long operational period: ~10 years of data taking.
 - Massive Monte Carlo simulations required for physics analyses and systematics studies.
- much larger event sizes

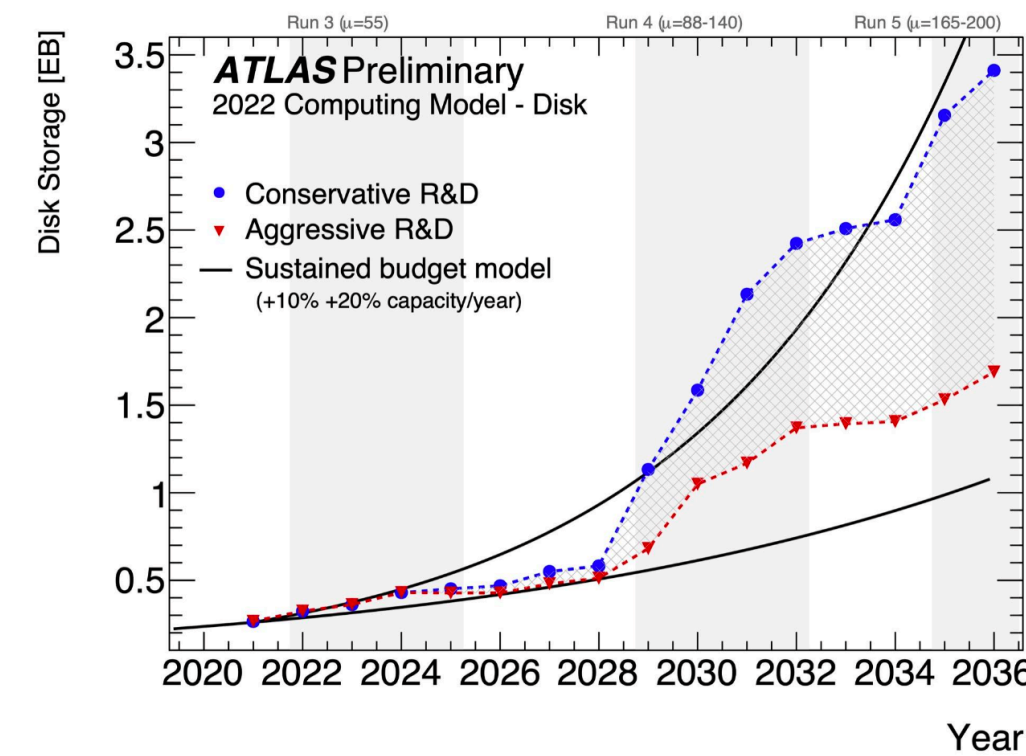
HL-LHC Data Volume Projections

- Different data formats to be considered:

- Raw detector data
- Simulated data (MC)
- Reconstructed data (AOD)
- Derived analysis datasets

- **ATLAS, CMS:** ~7(~10) EB expected including disk and tape by the end of Run 5, depending on compression, trigger thresholds, and data formats.
- **LHCb:** O(1) EB by the end of Run 5. 10x pileup in Run 5 + high rate full software trigger
- **ALICE:** Continuous readout mode, 170 GB/s to the disk buffer. ~2 EB by the end of Run 5

Disk



Tape

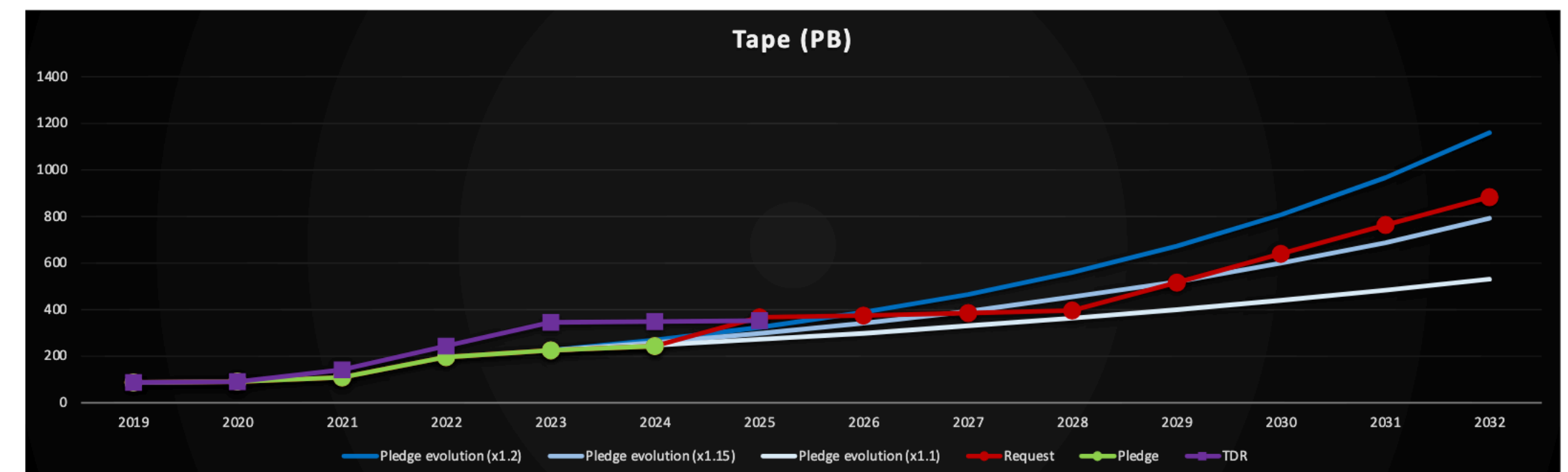
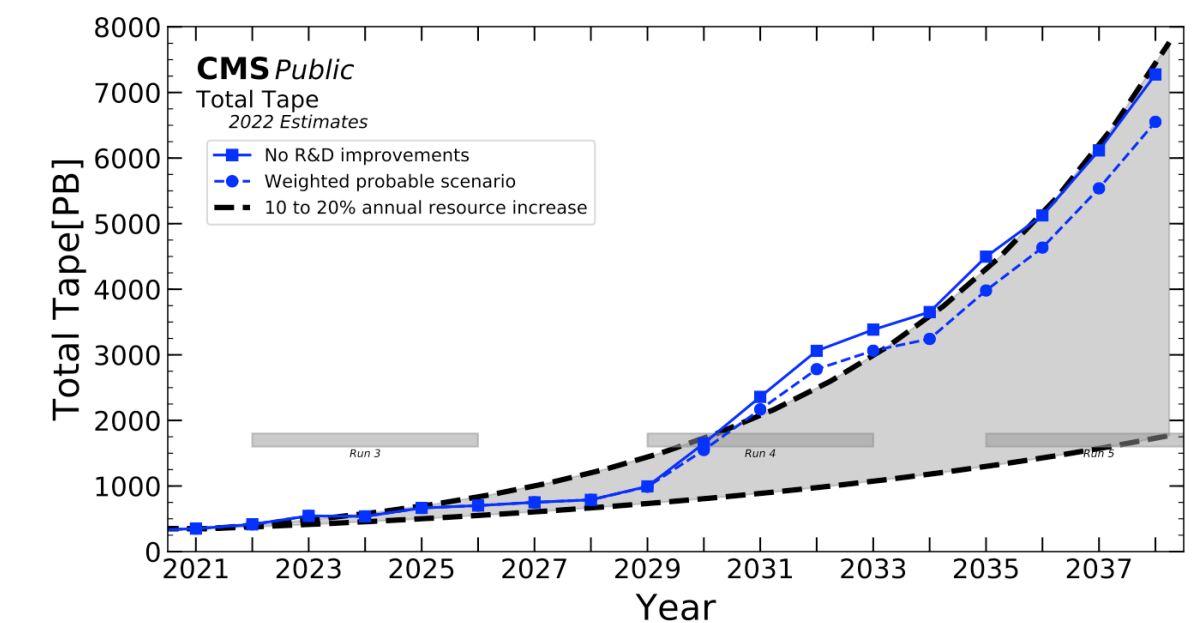
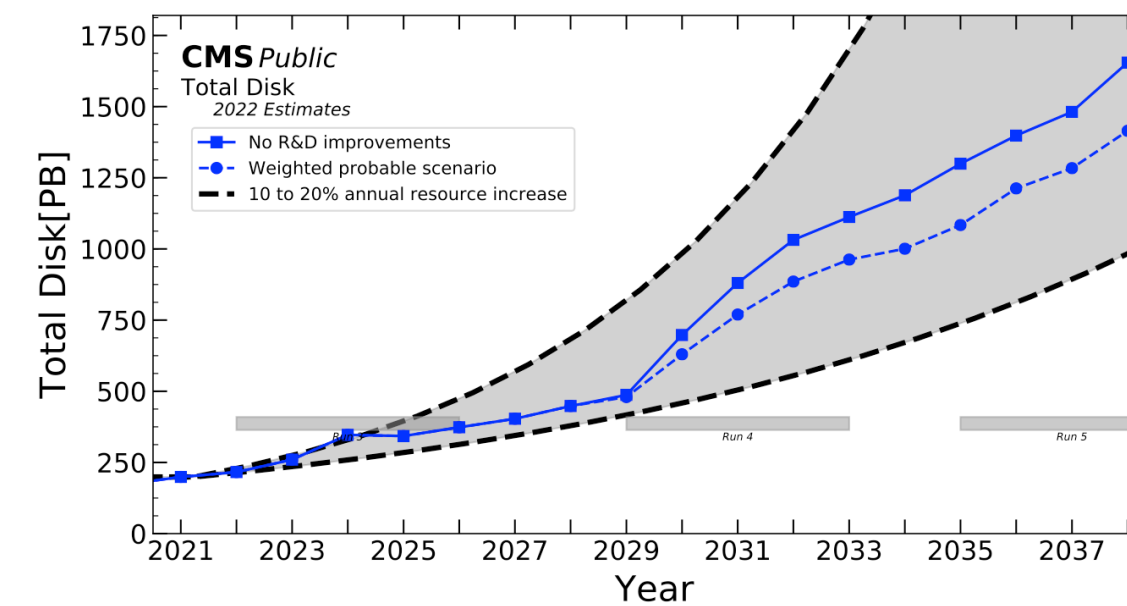
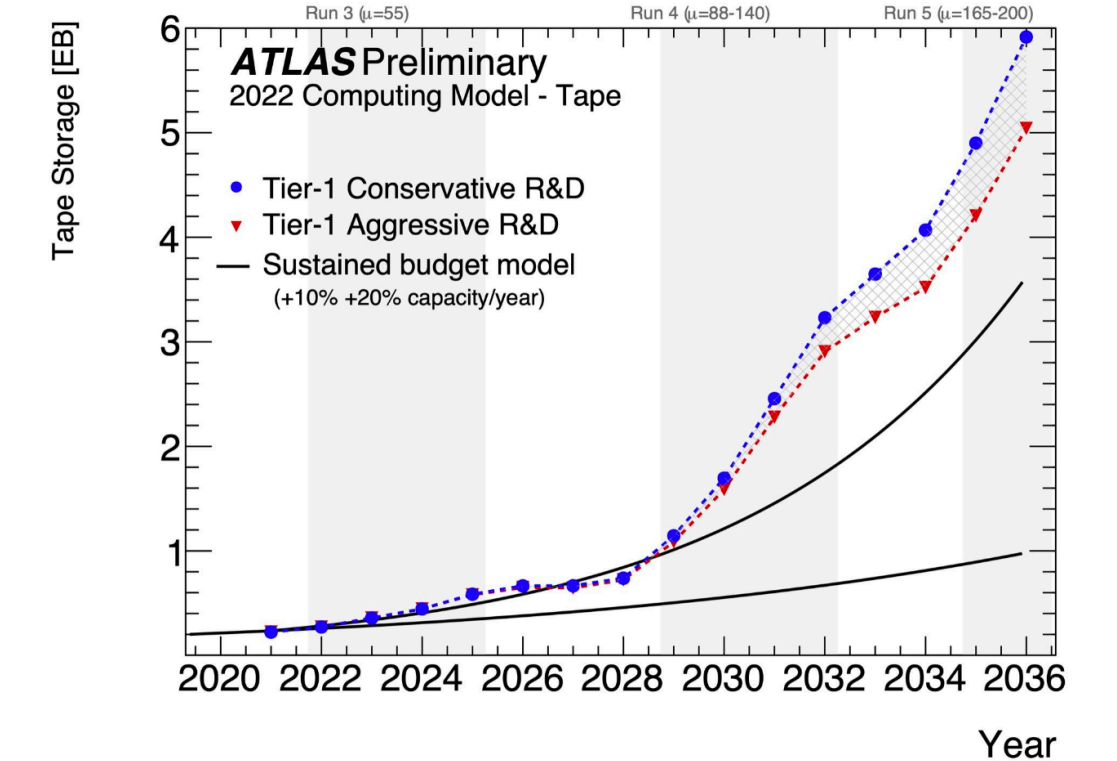


Figure 3: LHCb Tape Storage requirements. Blue lines show the FC-bands. Green shows actual pledges, red shows total projected requirements. Purple shows the requests made in the LHCb Upgrade Computing Model TDR.

HL-LHC Data Volume Projections

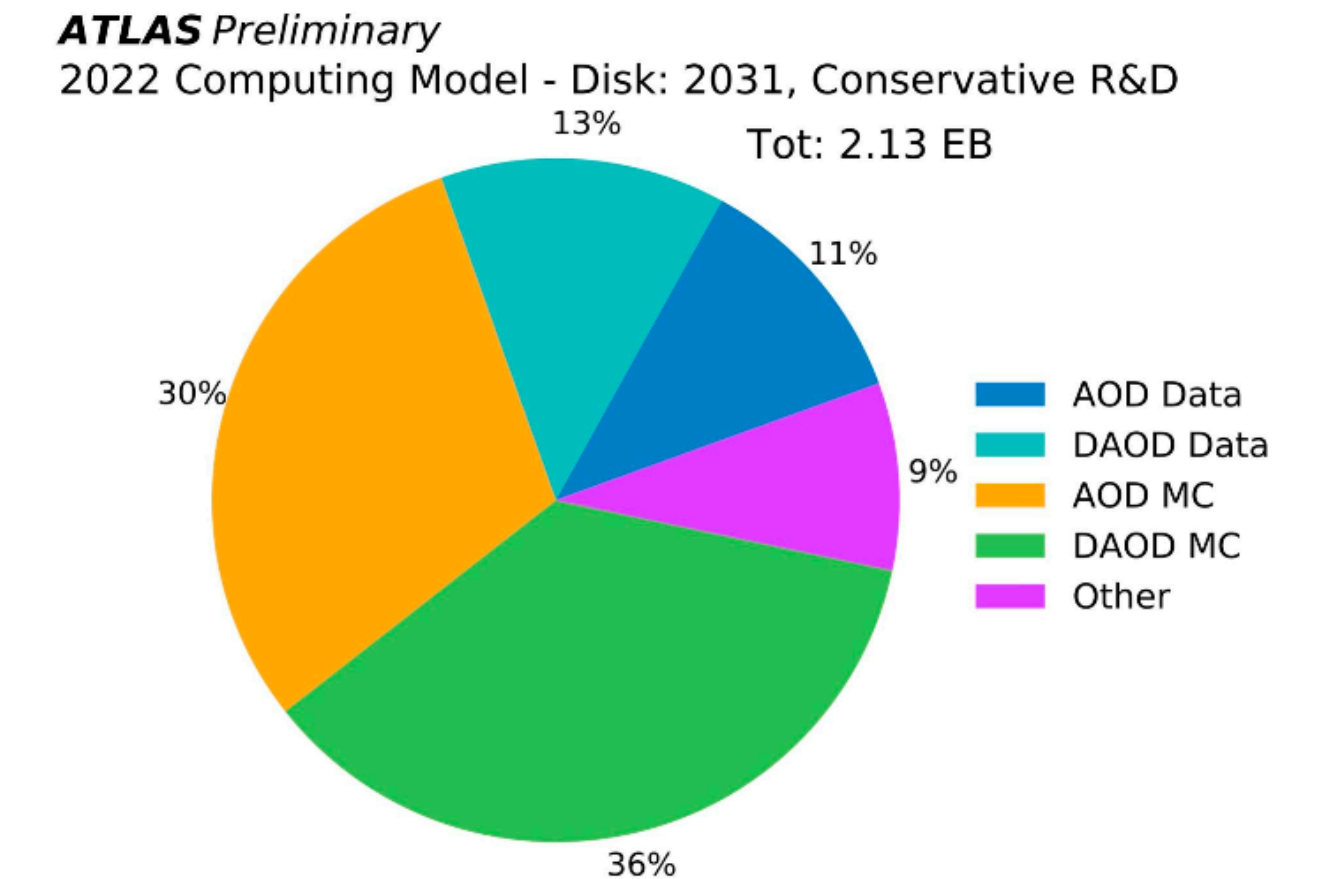
- Different data formats to be considered:

- Raw detector data
- **Simulated data (MC)**
- Reconstructed data (AOD)
- Derived analysis datasets

- **ATLAS, CMS:** ~7(~10) EB expected including disk and tape by the end of Run 5, depending on compression, trigger thresholds, and data formats.
- **LHCb:** O(1) EB by the end of Run 5. 10x pileup in Run 5 + high rate full software trigger
- **ALICE:** Continuous readout mode, 170 GB/s to the disk buffer. ~2 EB by the end of Run 5

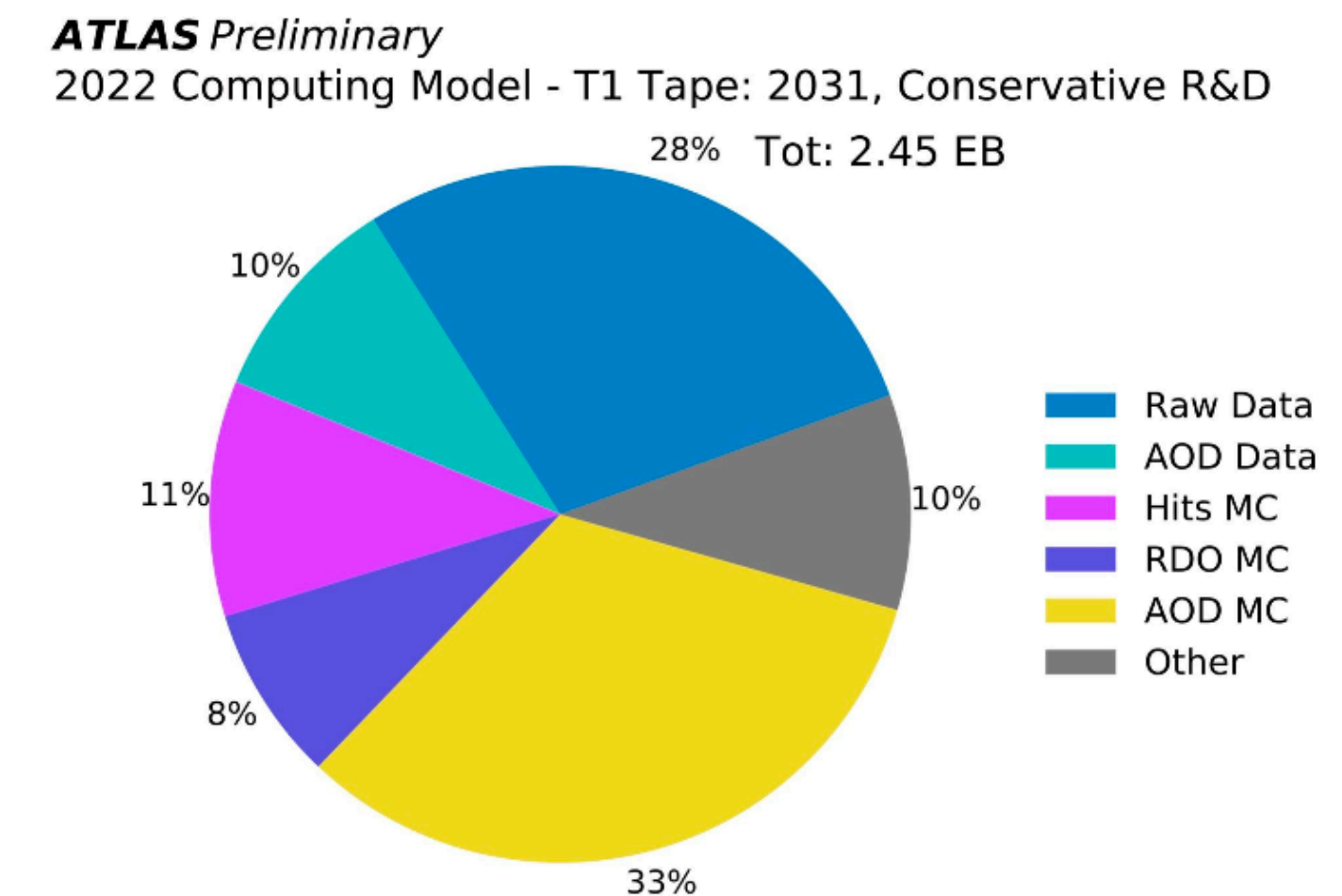
Disk

> 65% MC



Tape

> 50% MC



Architecture for the offline analysis

WORLDWIDE LHC COMPUTING GRID



A key tool for physics

The most sophisticated data-taking & analysis system ever built for science, providing near real-time access to LHC data.



Seamless access

Computing resources which include data storage capacity, processing power, sensors, visualization tools and more.



Global collaboration

42 countries
170 computing centres
Over 1 million computer cores
2 exabytes of storage



Enabling discovery

WLCG computing enabled physicists to announce the discovery of the Higgs Boson on 4 July 2012.

<https://wlcg-public.web.cern.ch/>

Architecture for the offline analysis

Tier-0

CERN Data Center, located in Geneva

All data from the LHC passes through this central hub

20% of the total compute capacity

Raw data (first copy), Reconstruction and distribution to Tier-1s

Tier-1

14 large computing centers

Safe-keeping part of the raw data and simulation

Reconstruction and distribution to Tier-2s

Tier-2

140 Universities or other institutes

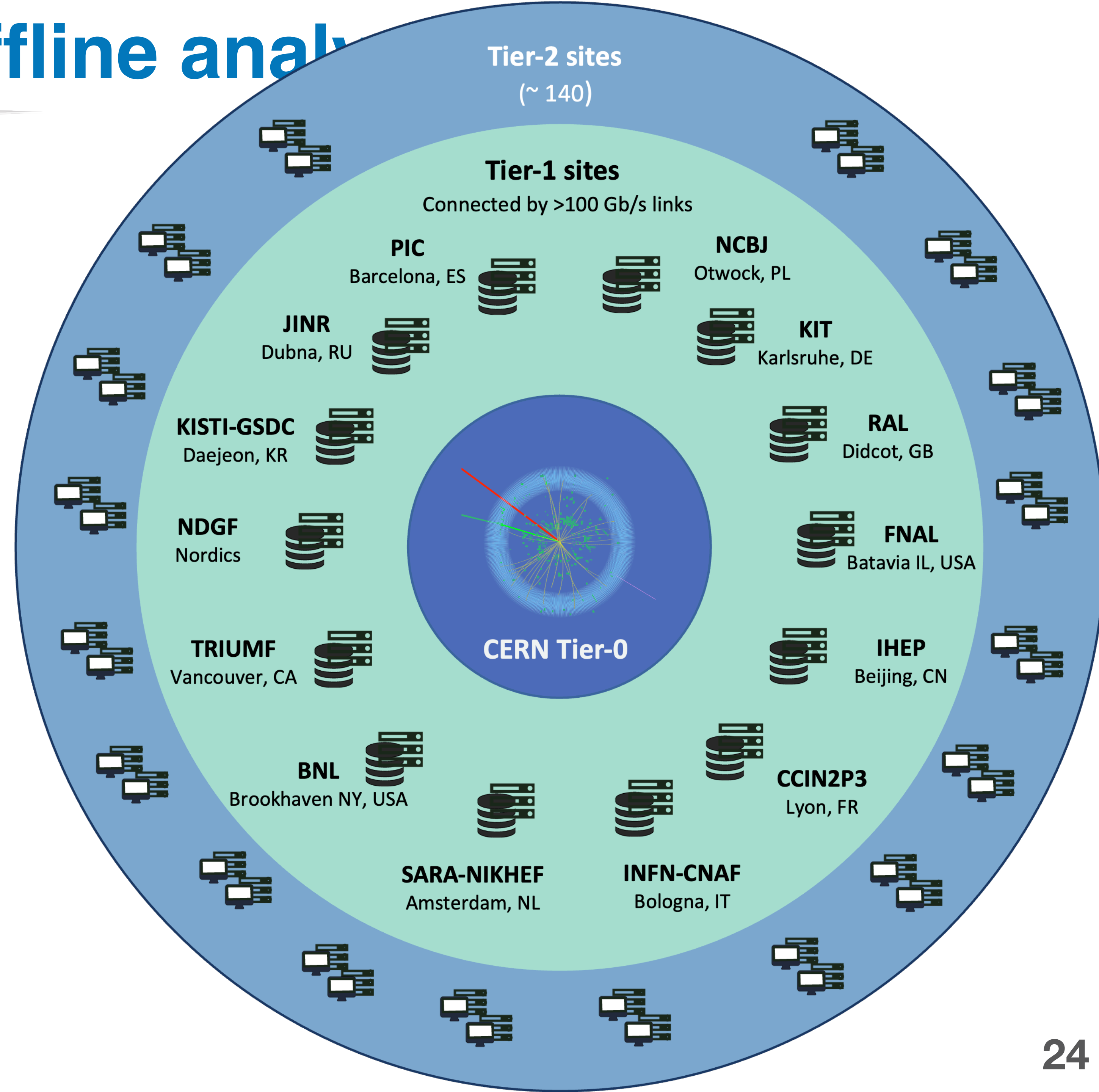
Store data and provide computer power to analysers

Simulation, reconstruction

Tier-3

Individual centers and local facilities

No formal engagement



Architecture for the offline analysis

Tier-0

CERN Data Center, located in Geneva

All data from the LHC passes through this center

20% of the total compute capacity

Raw data (first copy), Reconstruction and distribution to Tier-1s

Tier-1

14 large computing centers

Safe-keeping part of the raw data and simulation

Reconstruction and distribution to Tier-2s

Tier-2

160 Universities or other institutes

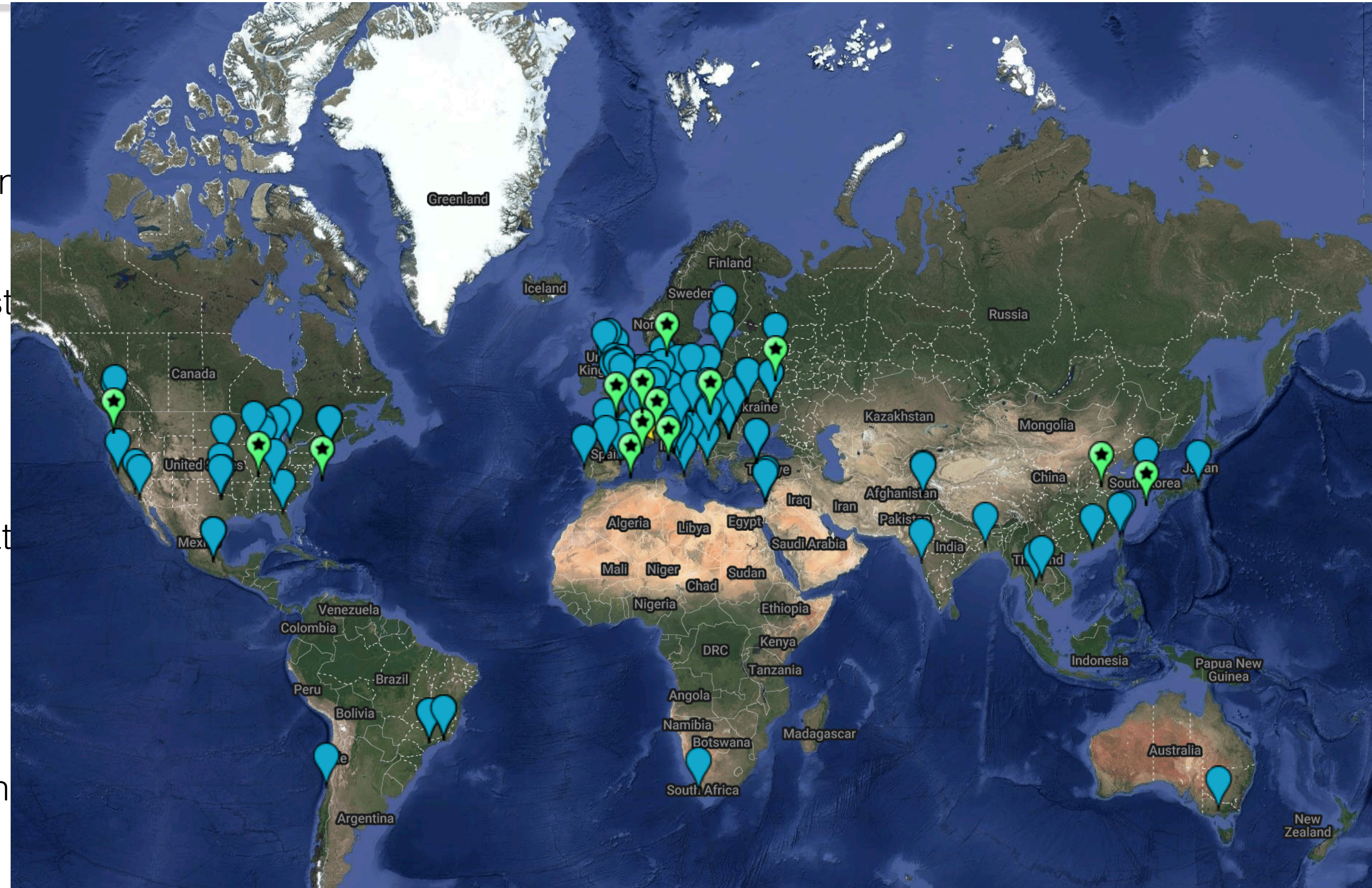
Store data and provide computer power to analysts

Simulation, reconstruction

Tier-3

Individual centers and local facilities

No formal engagement



[https://www.google.com/maps/d/u/0/viewer?](https://www.google.com/maps/d/u/0/viewer?mid=1uqabSfnzpBZ_fS_RfBFhRXx9nws&femb=1&ll=39.10707364257659%2C-27.421875&z=2)

[mid=1uqabSfnzpBZ_fS_RfBFhRXx9nws&femb=1&ll=39.10707364257659%2C-27.421875&z=2](https://www.google.com/maps/d/u/0/viewer?mid=1uqabSfnzpBZ_fS_RfBFhRXx9nws&femb=1&ll=39.10707364257659%2C-27.421875&z=2)

Architecture for the offline analysis

Tier-0



CERN Data Center, located in Geneva

All data from the LHC passes through this center

20% of the total compute capacity

Raw data (first copy), Reconstruction and distribution to Tier-1s

Tier-1



14 large computing centers

Safe-keeping part of the raw data and simulation

Reconstruction and distribution to Tier-2s

Tier-2



160 Universities or other institutes

Store data and provide computer power to analysts

Simulation, reconstruction

Tier-3

Individual centers and local facilities

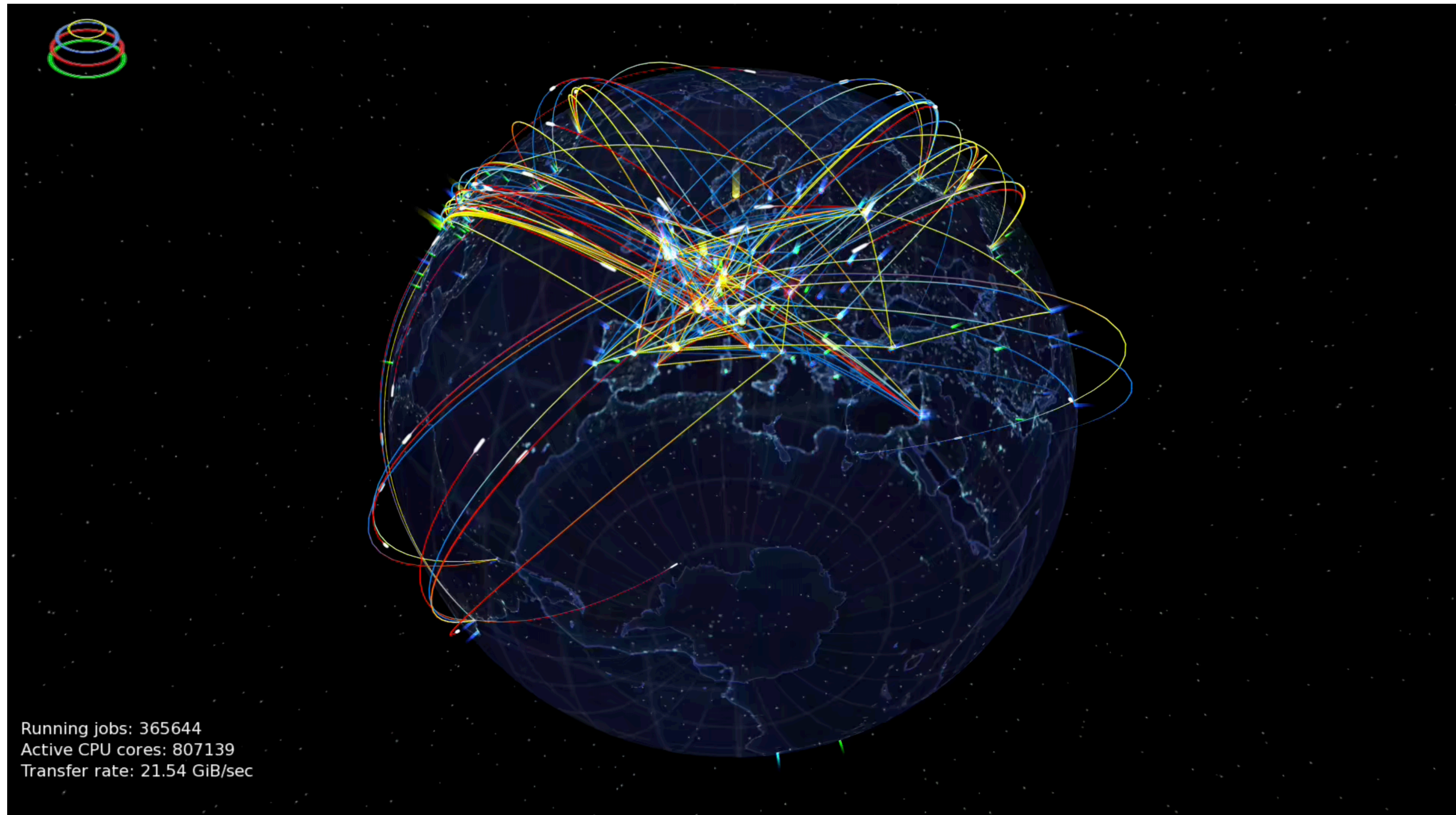
No formal engagement



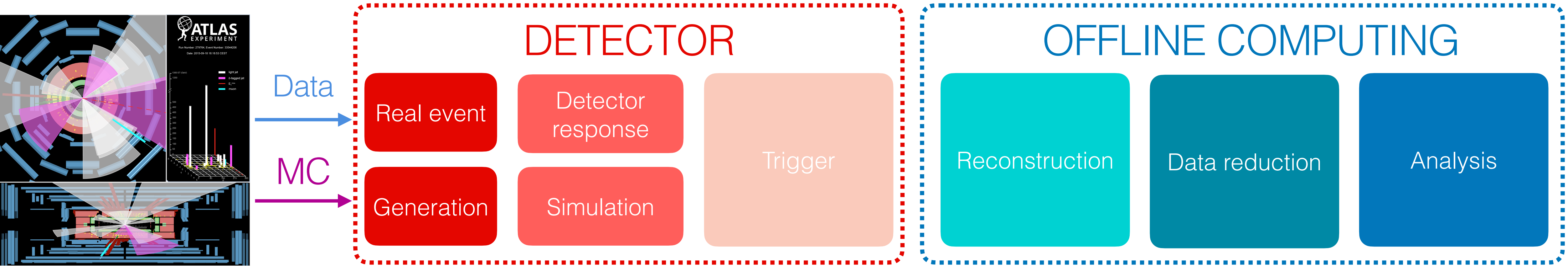
[https://www.google.com/maps/d/u/0/viewer?](https://www.google.com/maps/d/u/0/viewer?mid=1uqabSfnzpBZ_fS_RfBFhRXx9nws&femb=1&ll=39.10707364257659%2C-27.421875&z=2)

[mid=1uqabSfnzpBZ_fS_RfBFhRXx9nws&femb=1&ll=39.10707364257659%2C-27.421875&z=2](https://www.google.com/maps/d/u/0/viewer?mid=1uqabSfnzpBZ_fS_RfBFhRXx9nws&femb=1&ll=39.10707364257659%2C-27.421875&z=2)

Architecture for the offline analysis



Analysis



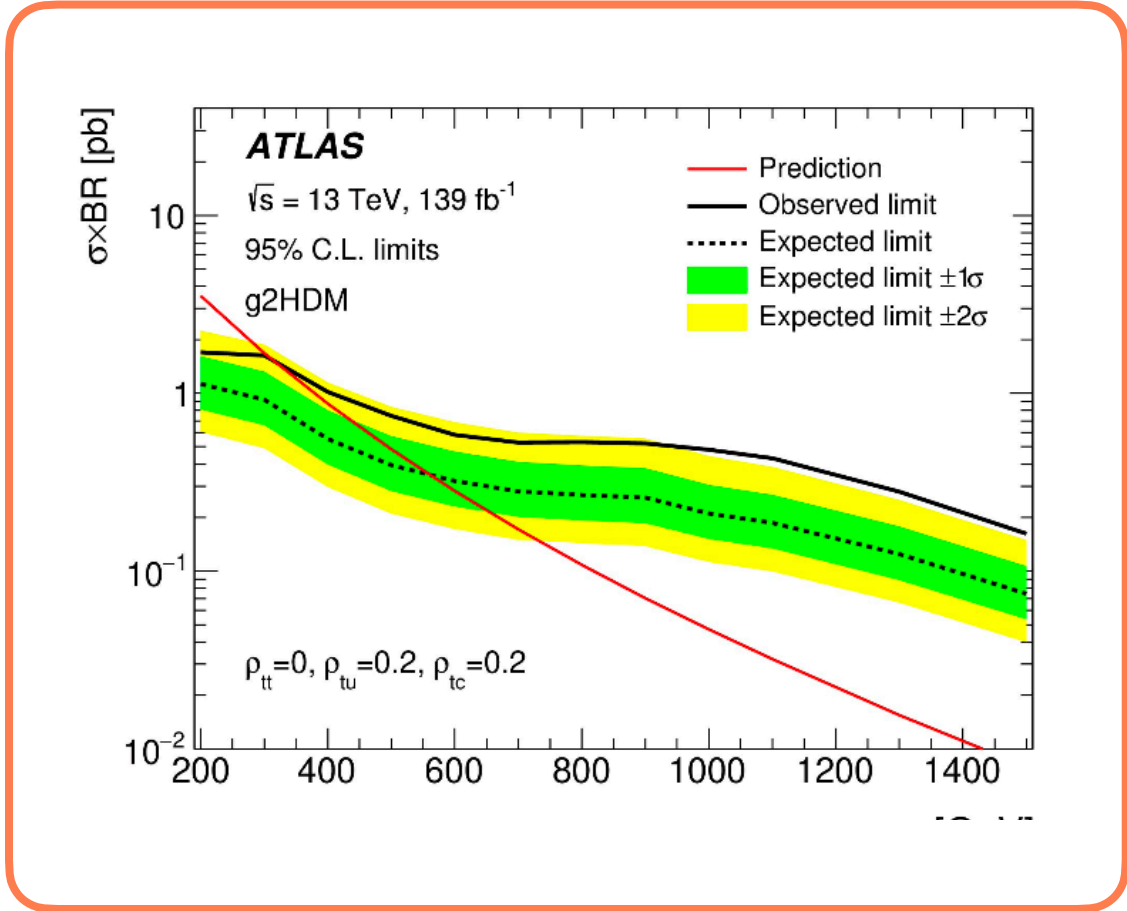
Data analysis

Project with the goal of making a scientific statement based on experimentation (ATLAS data)

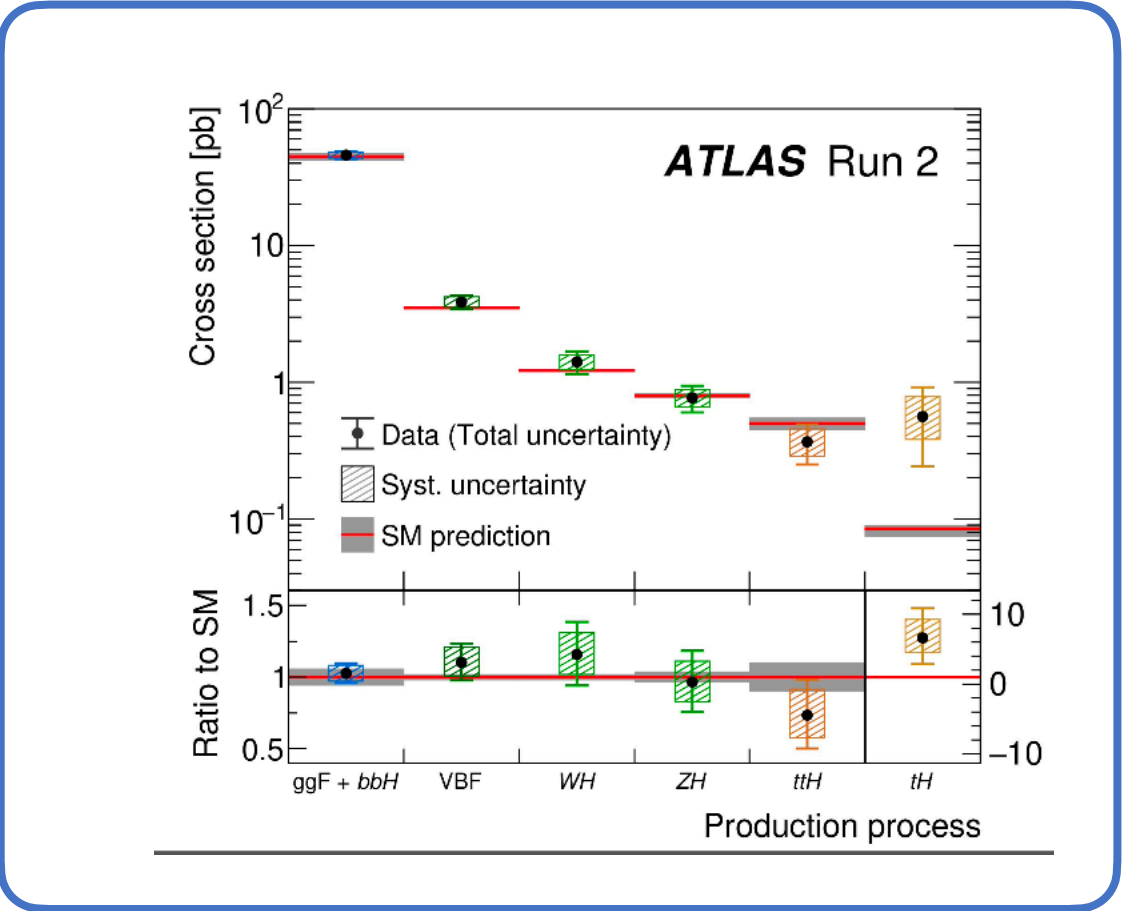
Output: Results (numbers) with associated uncertainties

Most common analyses in ATLAS:

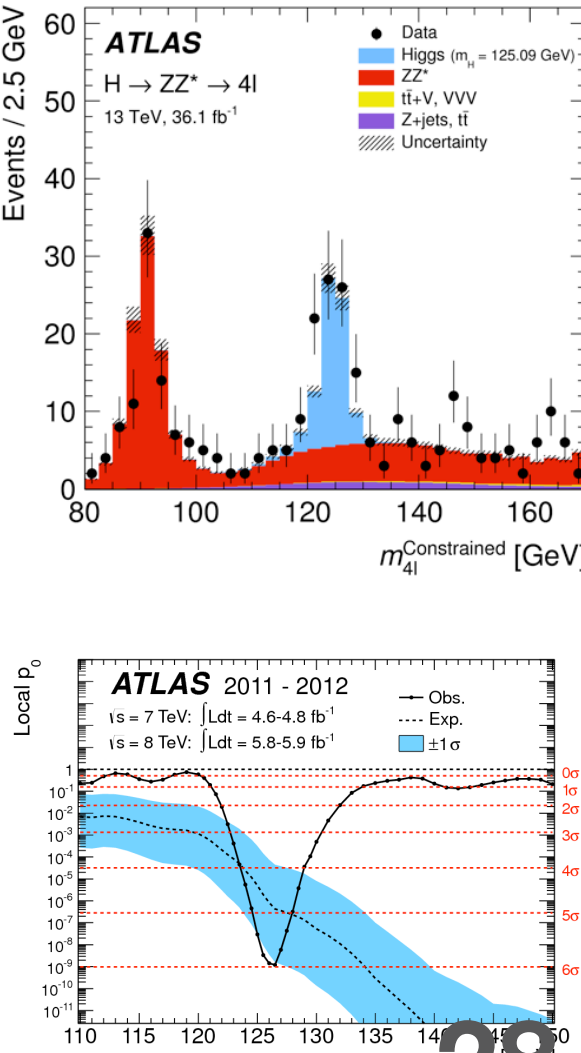
searches

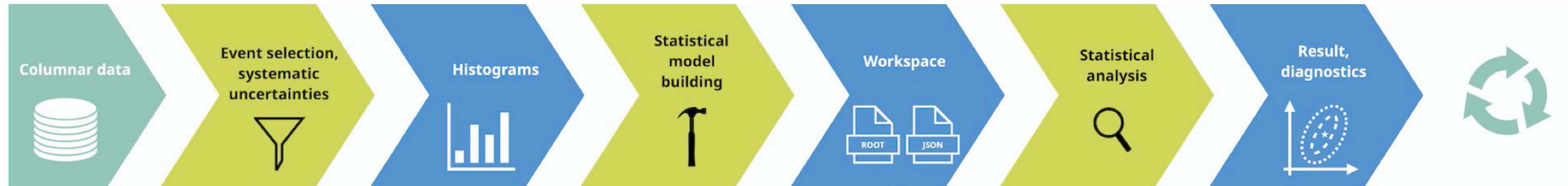


measurements



Local production

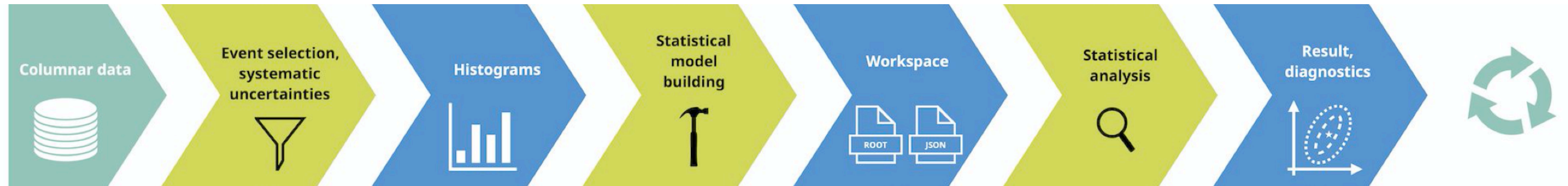




data query

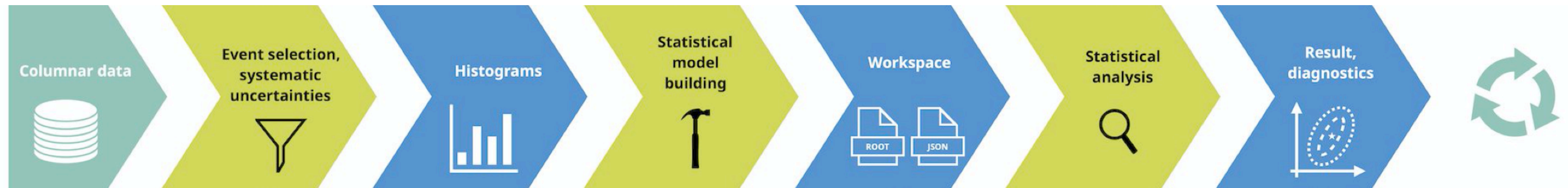
<https://iris-hep.org/projects/servicex.html>

<https://iris-hep.org/projects/func-adl.html>

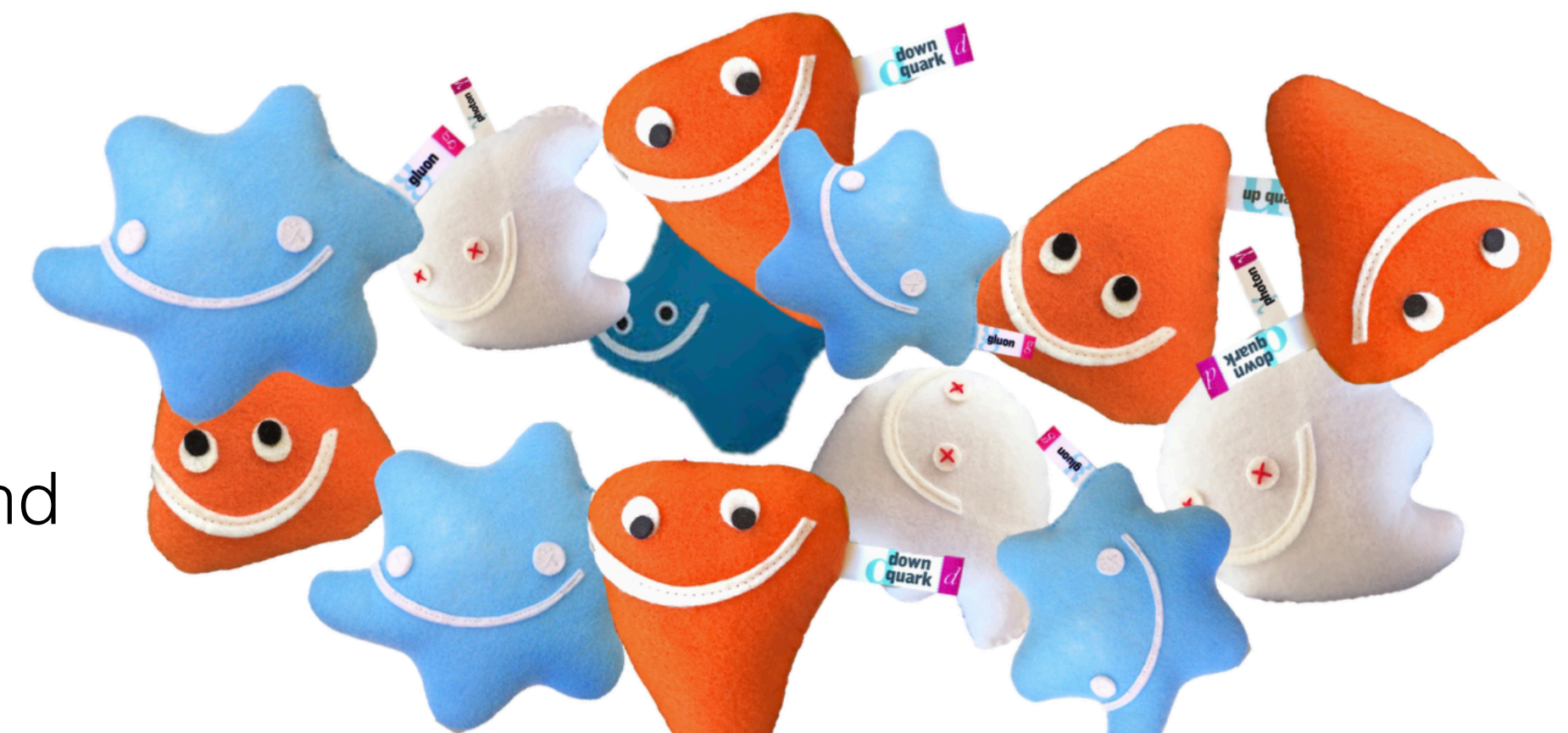


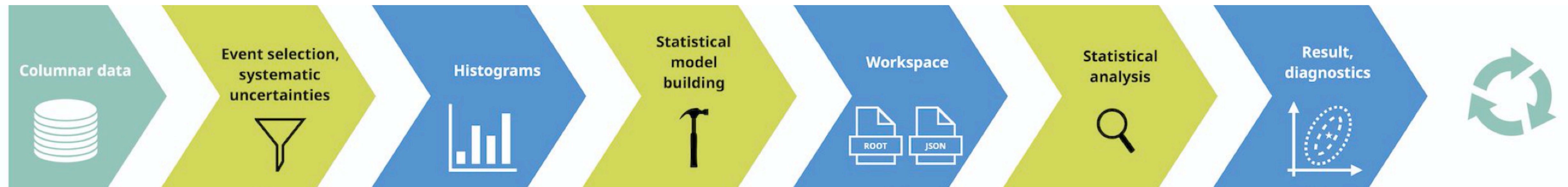
- Event selection
 - LHC collision data contain a lot of physics processes!
 - Depending on the analysis, we only care about a specific process we call “our signal” — e.g. the Higgs boson





- Event selection
 - LHC collision data contain a lot of physics processes!
 - Depending on the analysis, we only care about a specific process we call “our signal” — e.g. the Higgs boson
 - We don’t care about other processes we call “our background”
- Our sensitivity improves if we isolate our signal:
 - Apply a selection (phase space) to keep more signal than background





- Event selection

<https://iris-hep.org/projects/uproot.html>

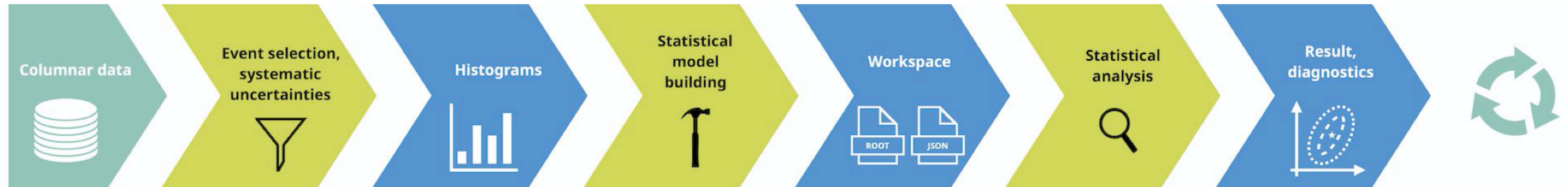
<https://iris-hep.org/projects/awkward.html>



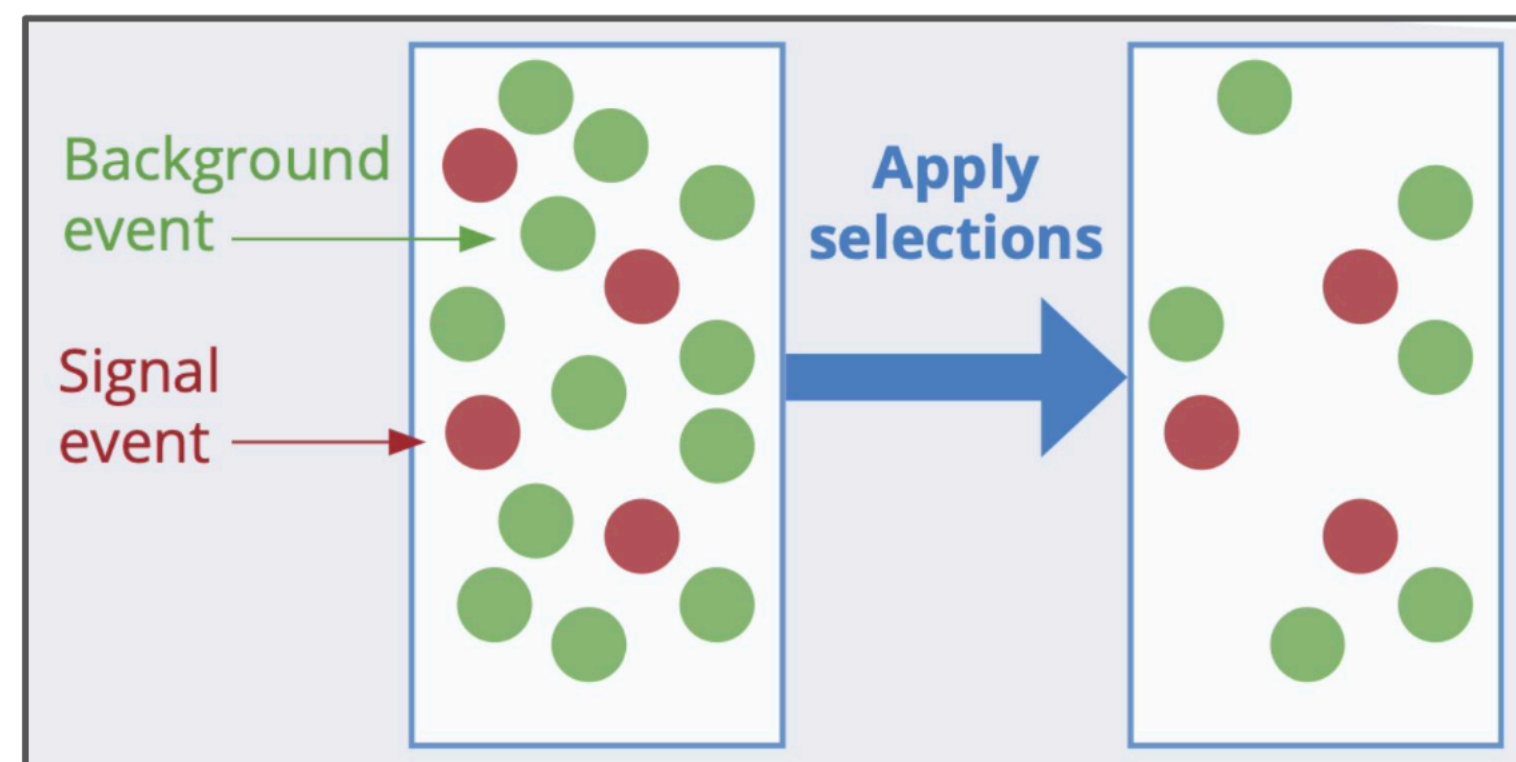
data access and selection

<https://github.com/scikit-hep/coffea>

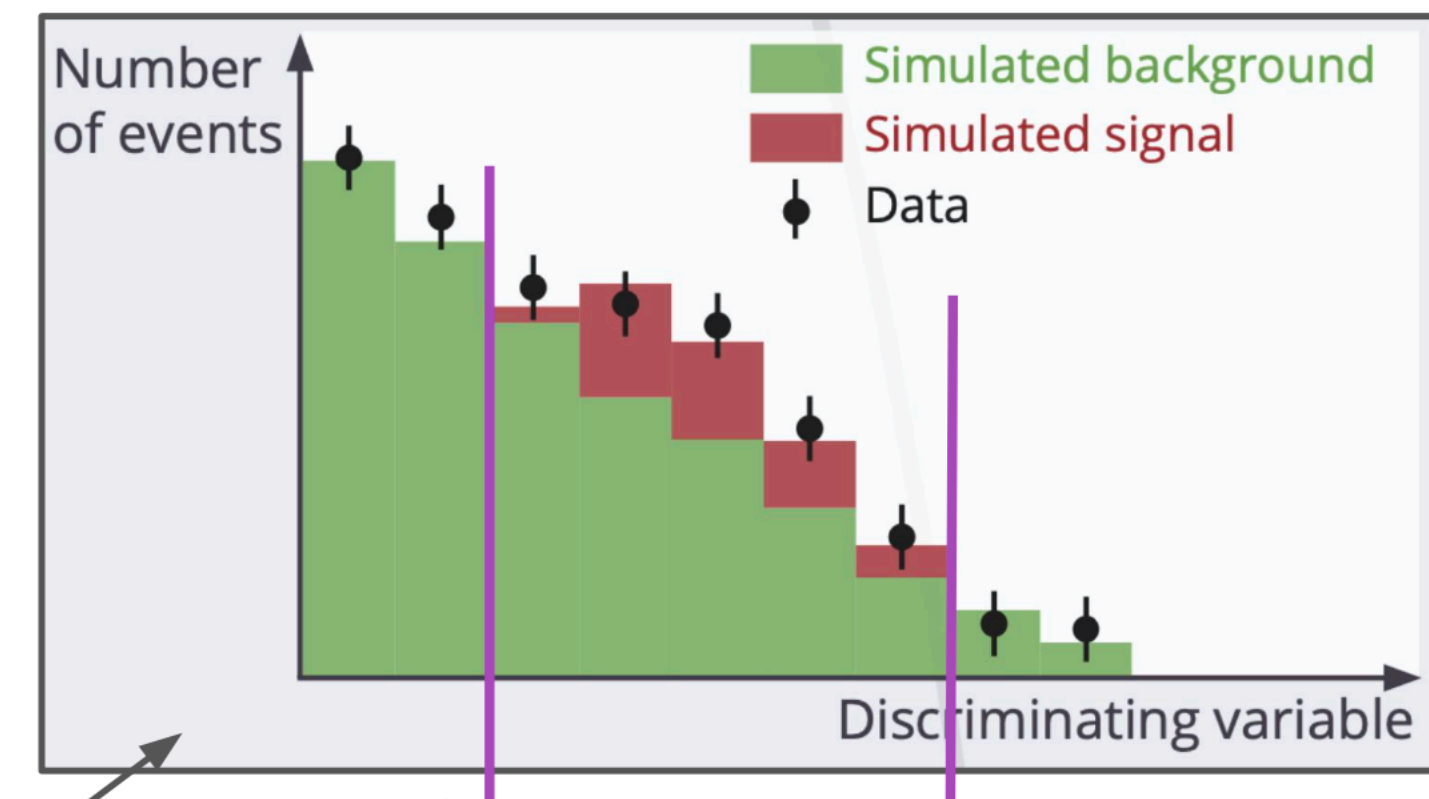
https://root.cern/manual/data_frame/

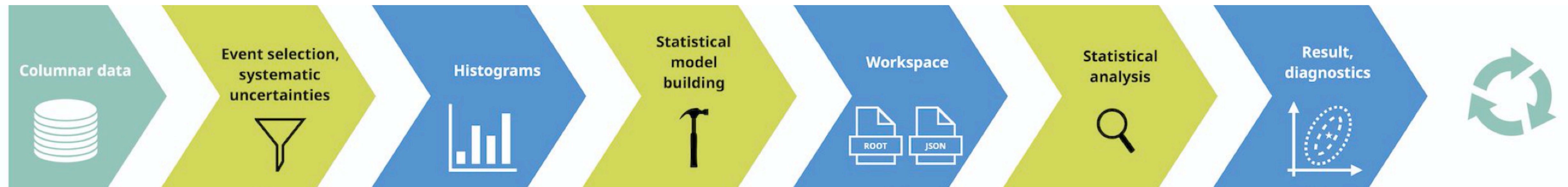


- Histograms
 - Refining event selection to maximise efficiency



- Keeping information on filled histograms
- Since this is a multidimensional problem, machine learning can provide large improvements





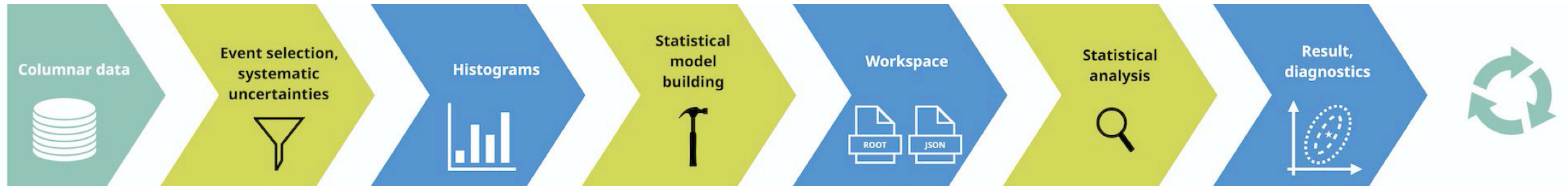
- Histograms



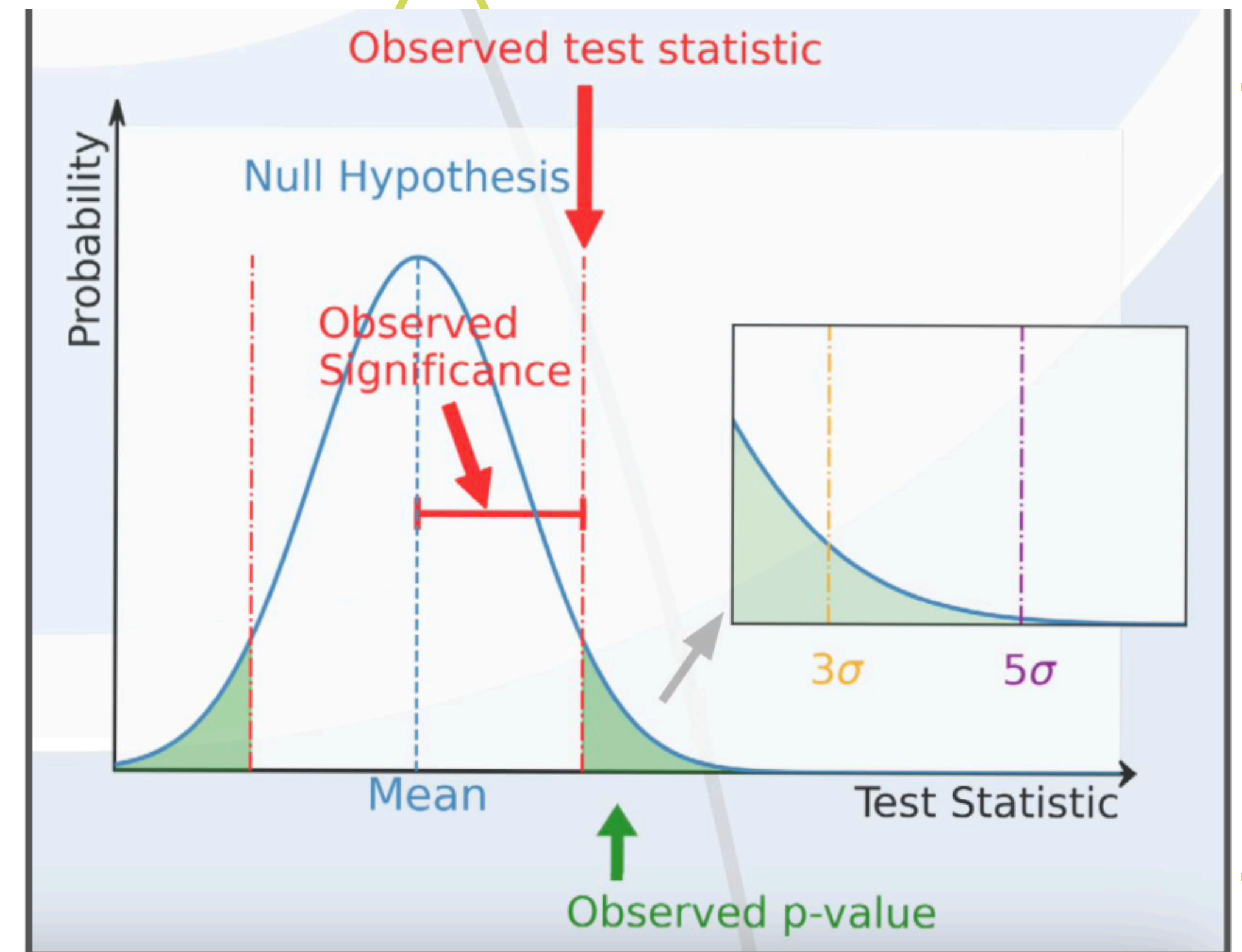
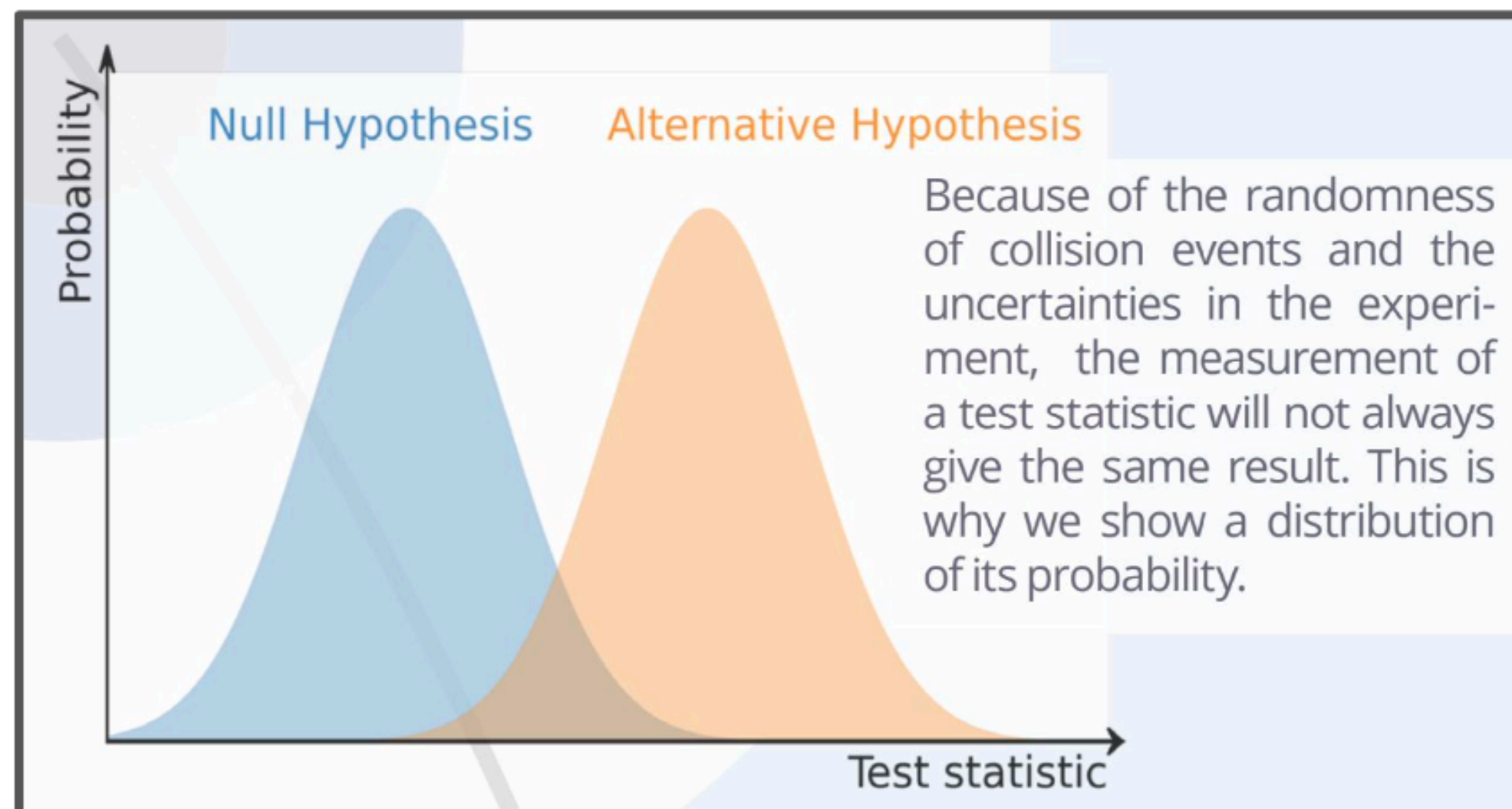
histogramming

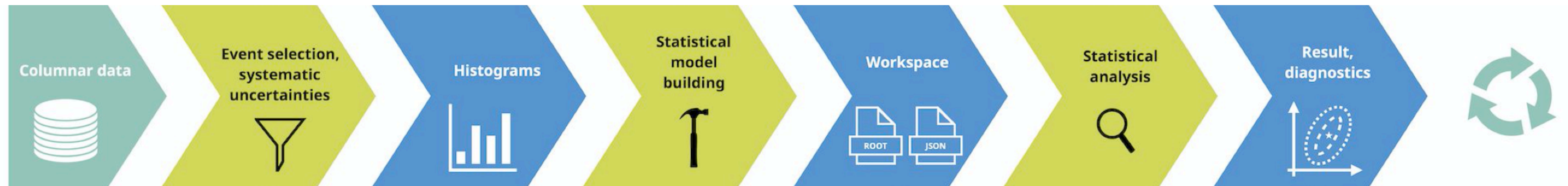
<https://boost-histogram.readthedocs.io/en/latest/>

<https://hist.readthedocs.io/en/latest/>



- Statistical analysis
 - Likelihood fits in signal and control regions to extract parameter of interest

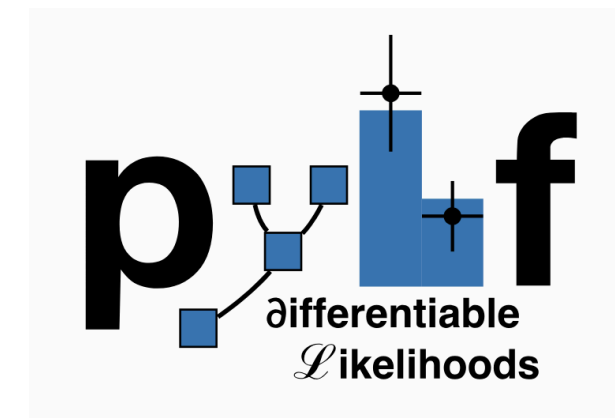




- Statistical analysis

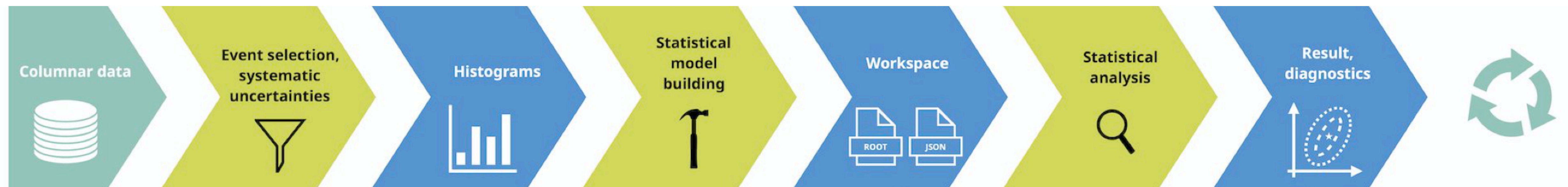
<https://pyhf.readthedocs.io/en/v0.7.6/>

<https://iris-hep.org/projects/cabinetry.html>

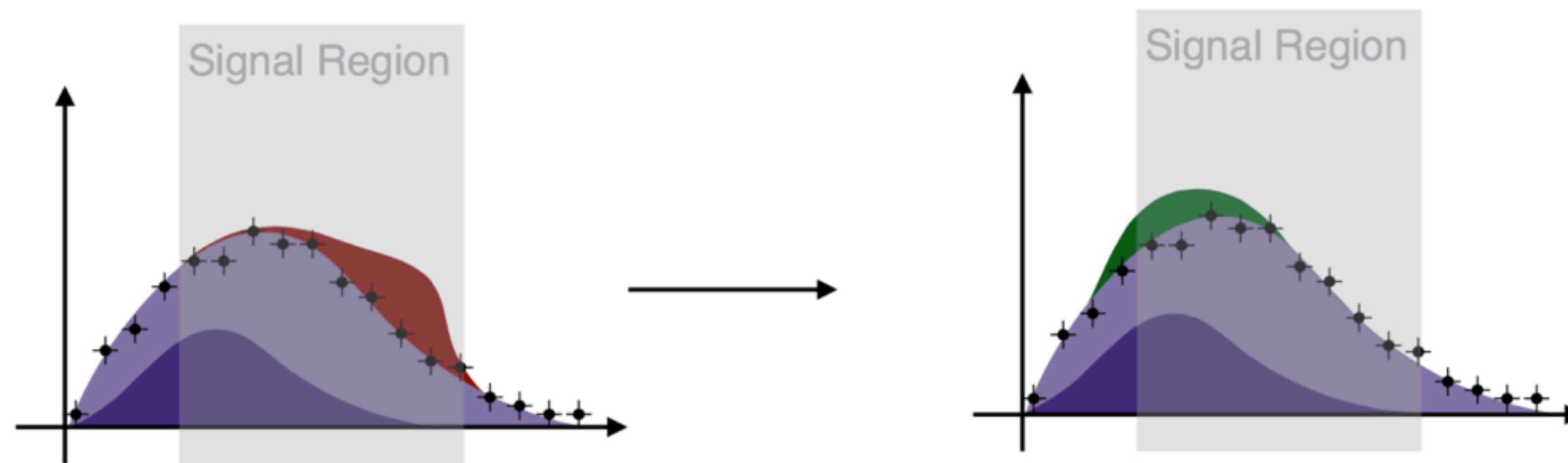


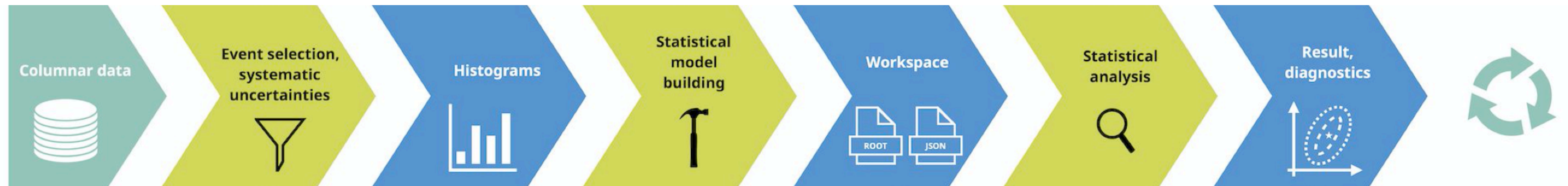
 **cabinetry**

statistics and modelling



- Preservation and reinterpretation



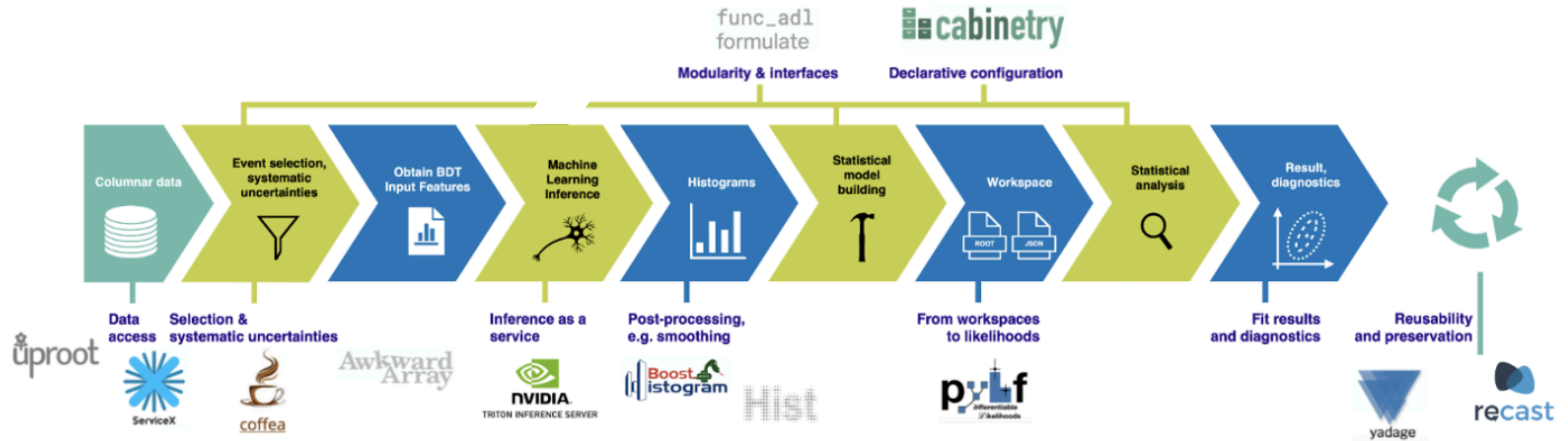


<https://github.com/recast-hep>



preservation

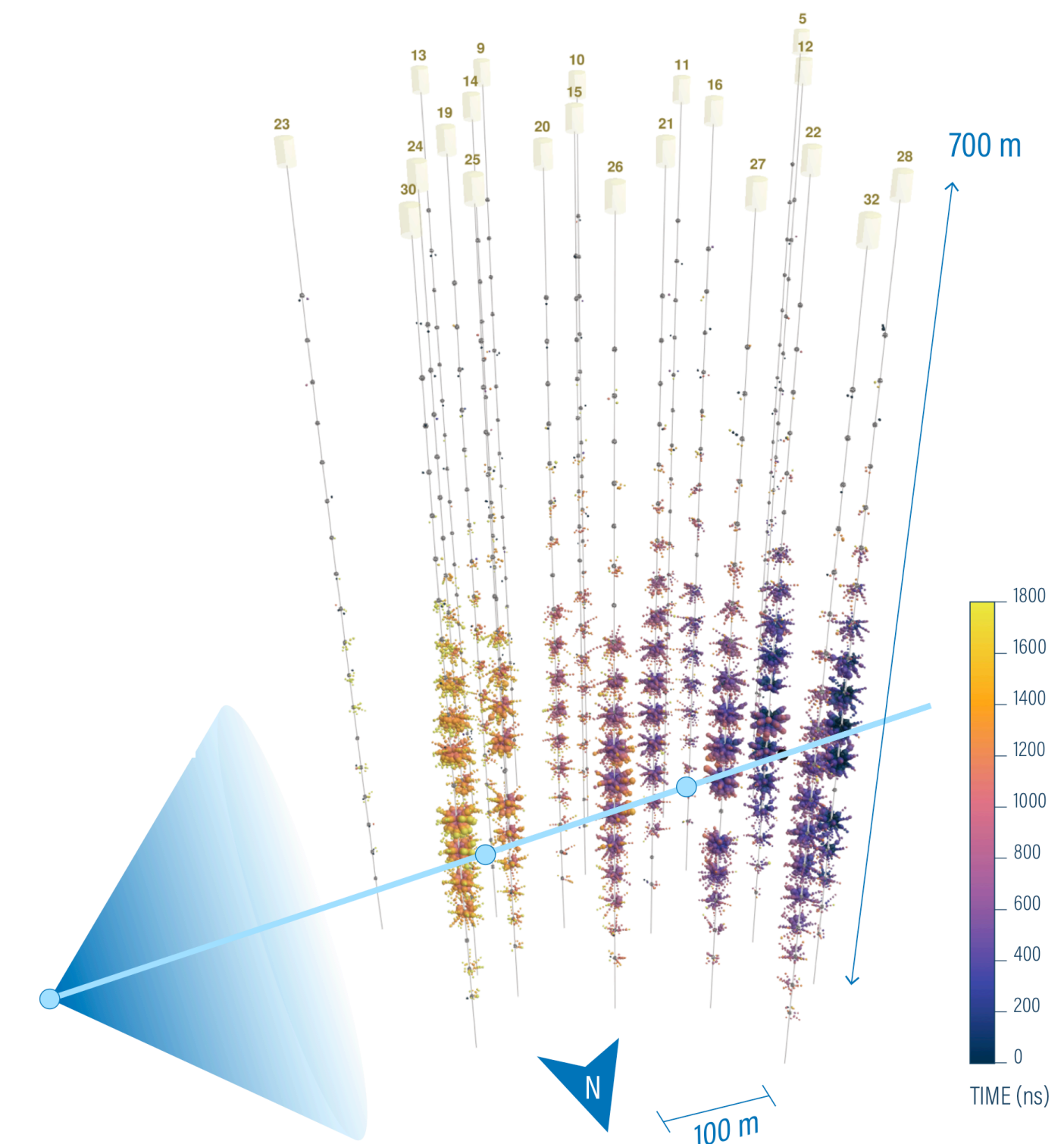
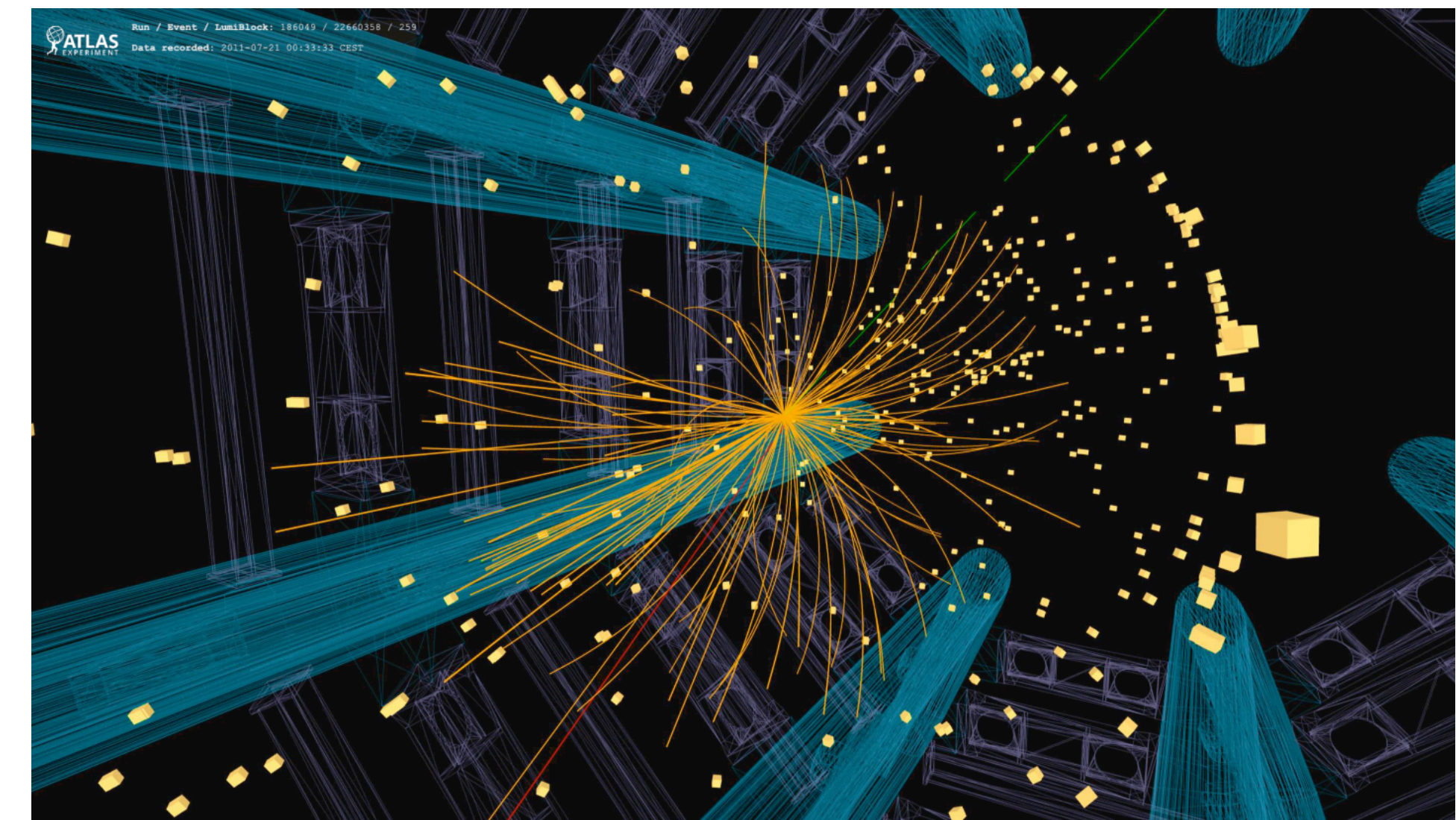
Analysis tools



Summary

- At the LHC, lots of data; in astrophysics, very little
- Same challenge: we don't observe physics directly, we observe signals in a detector.
 - everything we call 'physics' comes from reconstruction and interpretation of signals
 - that transformation happens offline!!
- **Offline computing is the bridge between detector and discovery**
- Complex pipeline with different steps
 - Parallel path in HEP and astrophysics
- Many computing challenges in the upcoming years
 - Finding new solutions in hardware and software
- Plethora of modern tools, try them out for an efficient analysis!

**“In modern physics,
discoveries are not only made in detectors,
they are made in algorithms.”**

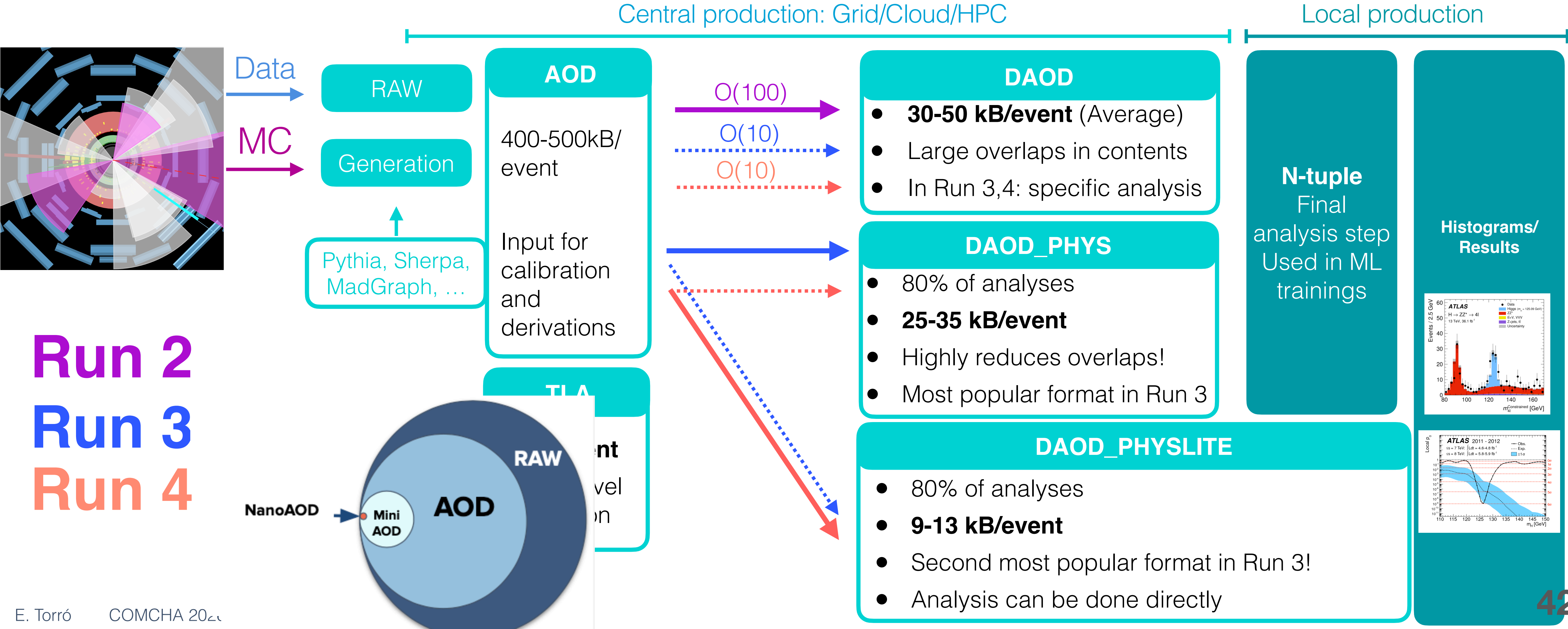
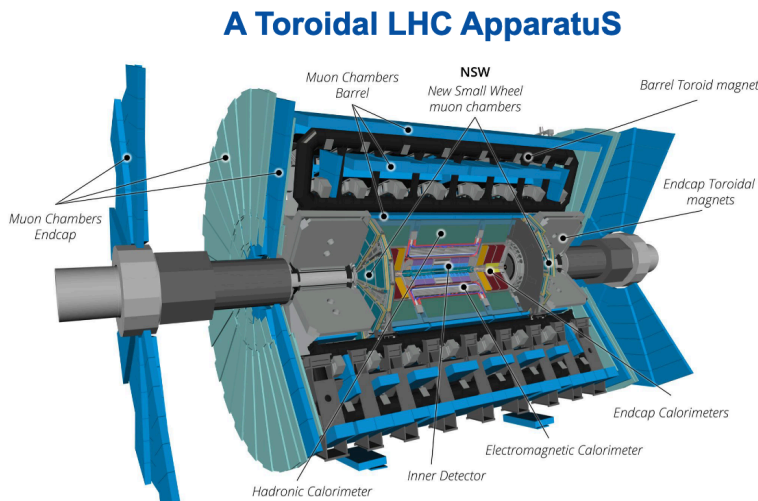




Backup

Data reduction steps

- Reducing datasets sizes to facilitate data analysis
- Selecting subsets of events/tracks/objects relevant for analysis.
- **Challenge:** keeping all interesting physics, scalable, possibly in real-time or quasi-real-time.

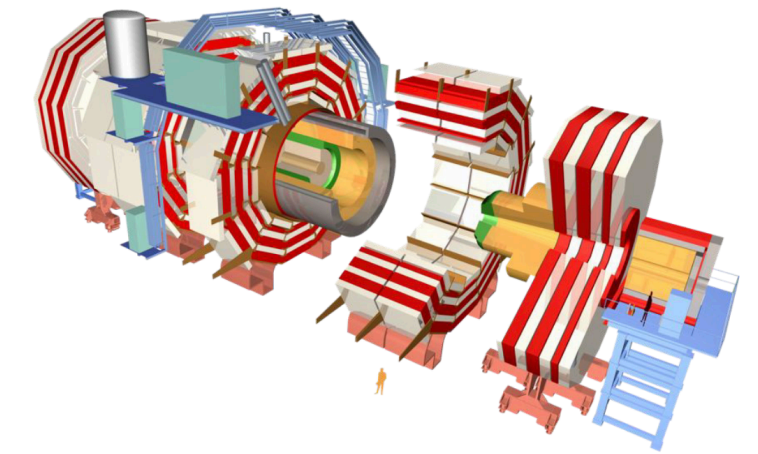


Data reduction steps

- Reducing datasets sizes to facilitate data analysis
- Selecting subsets of events/tracks/objects relevant for analysis.
- Challenge:** keeping all interesting physics, scalable, possibly in real-time or quasi-real-time.

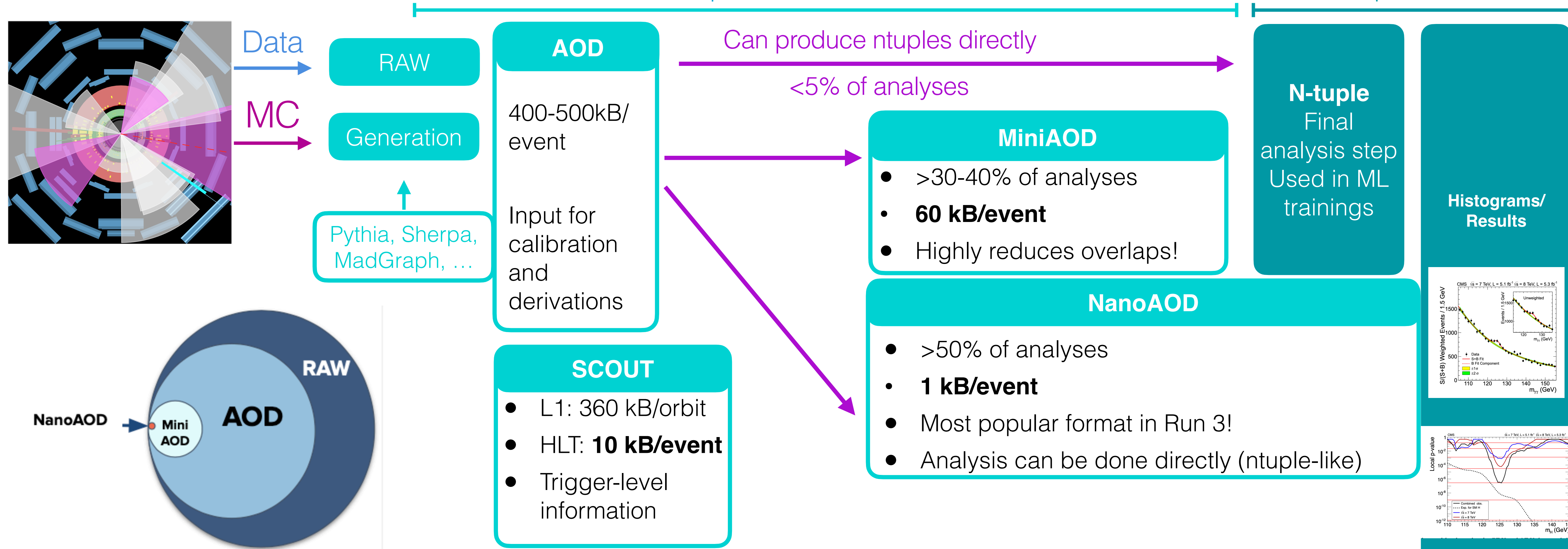
CMS

Compact Muon Solenoid



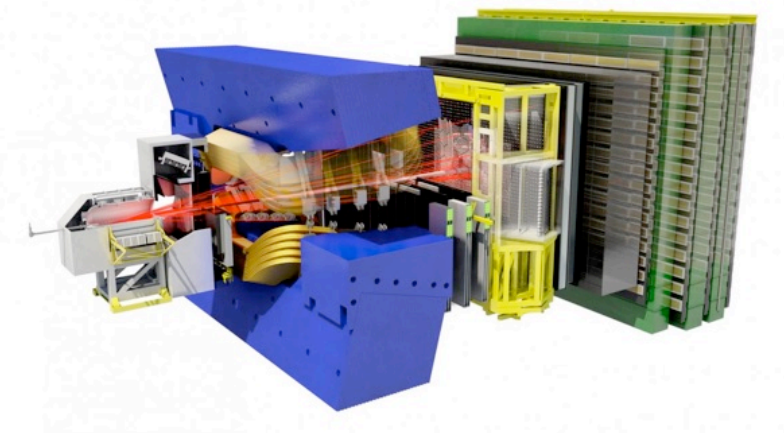
Central production: Grid/Cloud/HPC

Local production

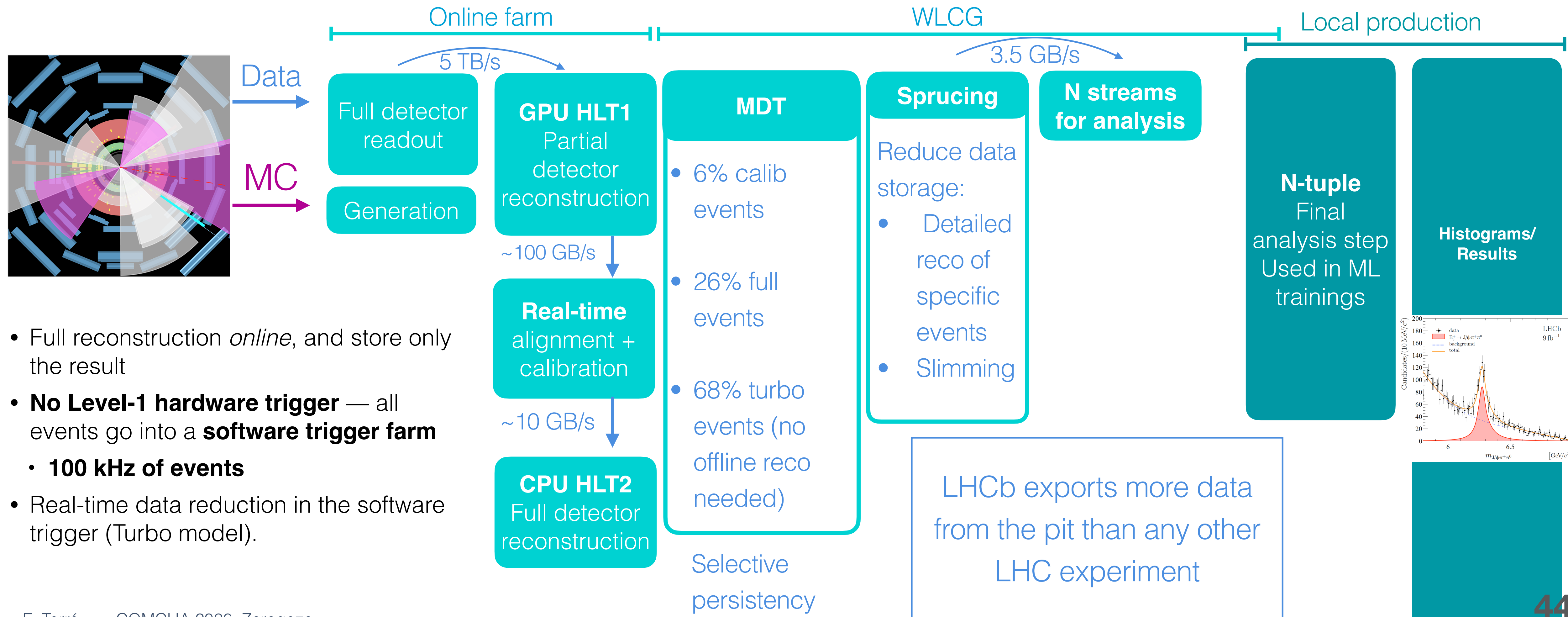


Data reduction steps

LHCb



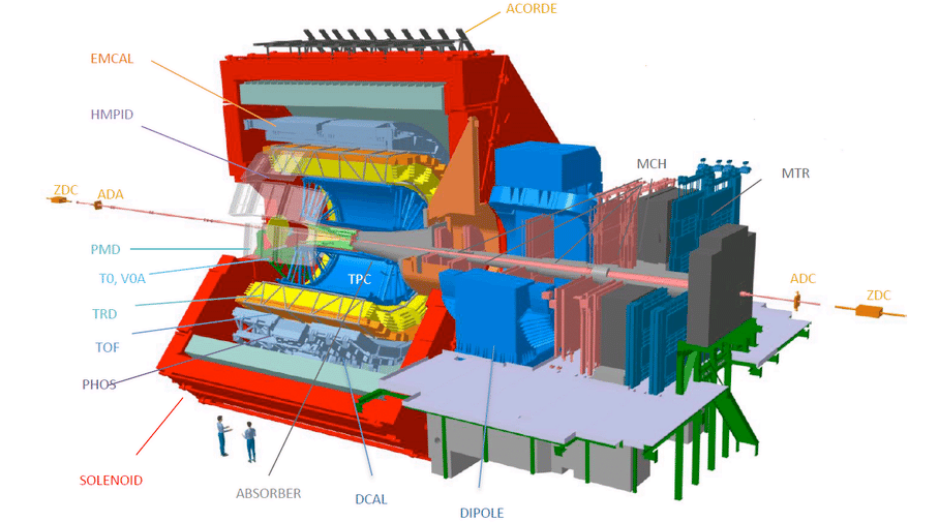
- Reducing datasets sizes to facilitate data analysis
- Selecting subsets of events/tracks/objects relevant for analysis.
- **Challenge:** keeping all interesting physics, scalable, possibly in real-time or quasi-real-time.



Data reduction steps

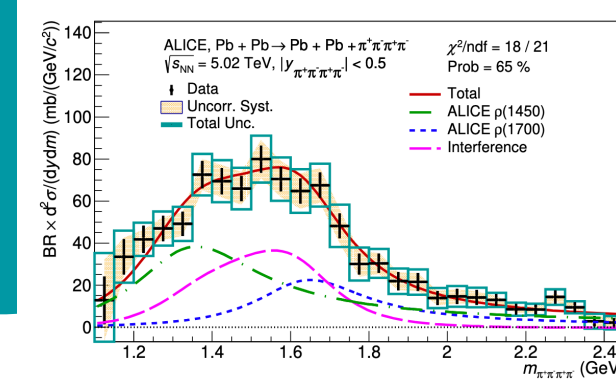
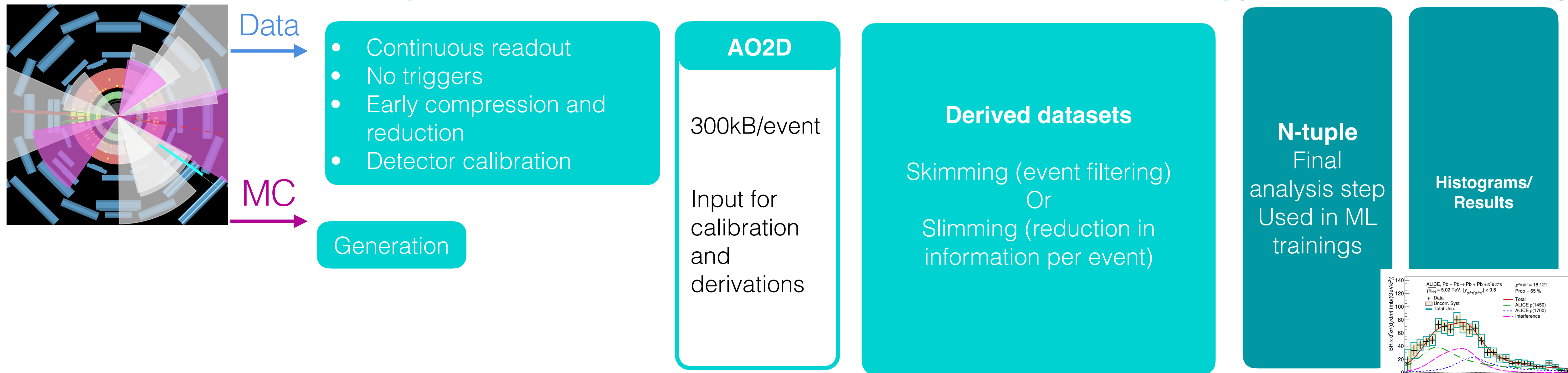
- Reducing datasets sizes to facilitate data analysis
- Selecting subsets of events/tracks/objects relevant for analysis.
- **Challenge:** keeping all interesting physics, scalable, possibly in real-time or quasi-real-time.

ALICE



Central production: Grid/Cloud/HPC

Local production



- Need to keep all rare heavy-ion events.
- Instead of triggering, early **compression**
- O² computing framework: combining online + offline in one.
- Data skimming in the Quality Control layer.