Argonne NATIONAL LABORATORY

… for a brighter future

U.S. Department of Energy

UChicago ▶ Argonne LLC

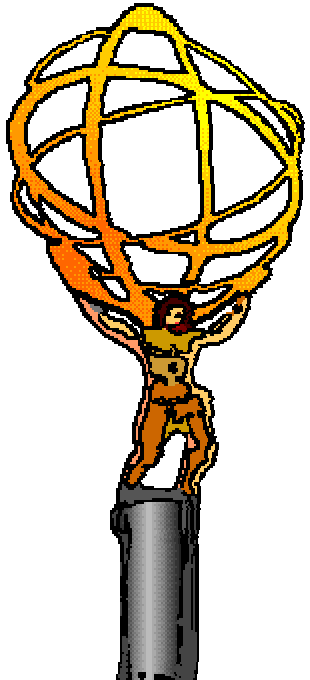
Office of Science U.S. DEPARTMENT OF ENERGY

A U.S. Department of Energy laboratory managed by UChicago Argonne, LLC

TAGS in the Analysis Model

Jack Cranshaw, Argonne National Lab

September 10, 2009



Objectives for this Talk

- This is not a tutorial, that would take several hours.
- What you should take from this presentation:
  - Know how to check the content of TAGS.
  - How to construct a query on the TAGS.
  - Understand the general changes needed to use TAGS with athena.
  - Understand plans for TAGS in real physics analyses.

Event Selection

- ***Observations***
  - ATLAS operates at a hadron collider where signal to noise is almost always low.
  - ATLAS will run for multiple years.
  - ATLAS event rates and data sizes impose limitations on the amount of data which can be hosted on any particular site.
  - Although running over a 1% subset of a large data sample will take more than 1% of time for the full data sample, it will take less than 100%.
  - Having multiple people run over the same data (RAW, ESD, AOD, …) to make similar selections or make similar rejections is a waste of resources.
- ***Consequences***
  - Event selection activities such as skimming, thinning, etc. are important and necessary activities.
  - Maximizing the information than can be used for these activities can have large leverage effects on computing and physicist resources.

Event-level Metadata

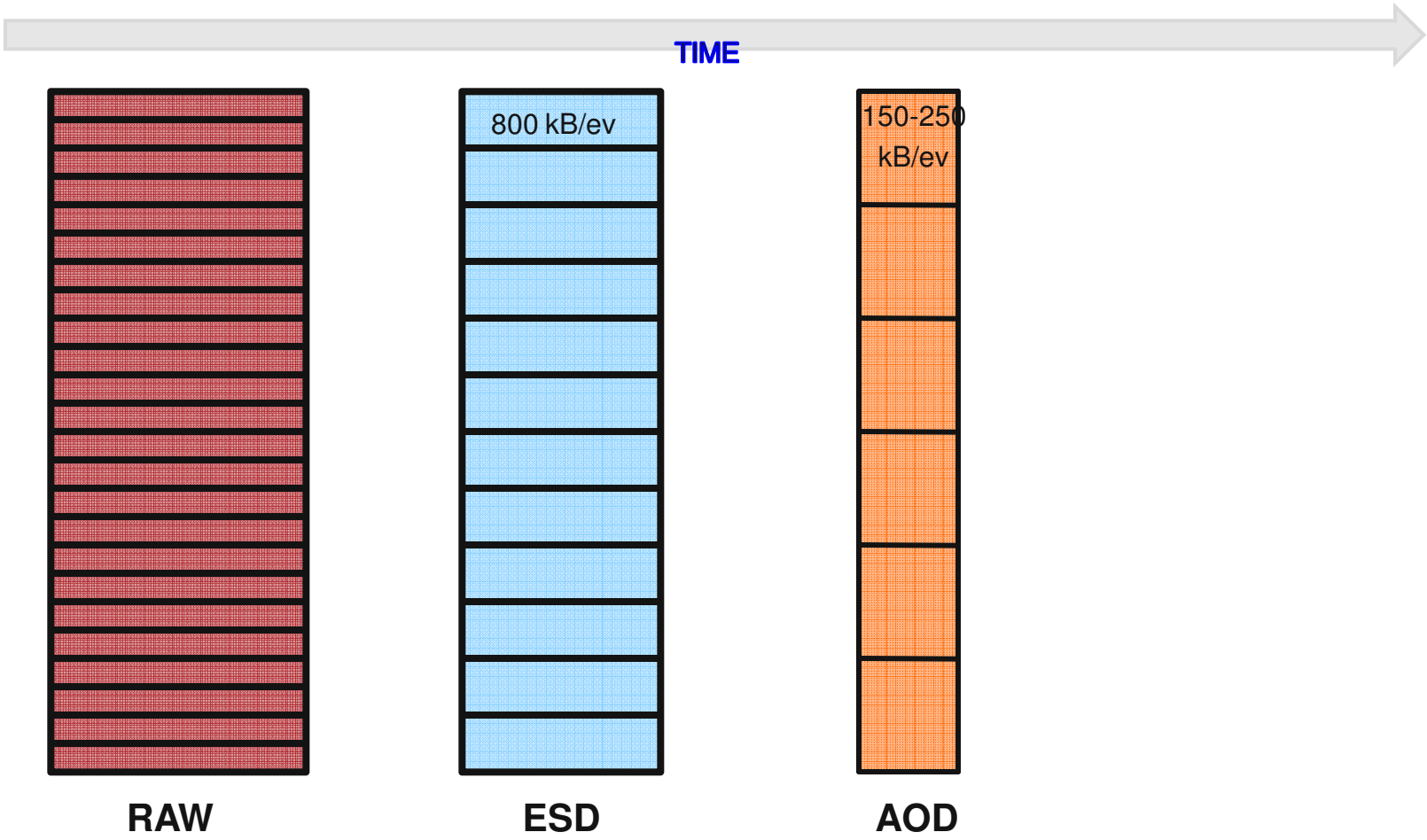
- Information about data is called metadata.
- In ATLAS, metadata exists at many levels (dataset, run, lumiblock, event, …)
- The TAGS are event-level metadata.
  - TAGS must work with metadata at the other levels as well. You will learn more about this in the following talk.
- TAGS contain
  - Event identification
  - Trigger information
  - Stream information
  - Detector status
  - Physics quantities
    - *Photon, Electron, Jet, Muon, Tau Jet*
  - Physics decisions
    - *BPhys, Exotic, Jet Tagging, …*
  - https://twiki.cern.ch/twiki/bin/view/AtlasProtected/TagForEventSelection

Are TAGS an ntuple?

- TAGS look like an ntuple: a set of related quantities in rows by event.
  - TAGS are stored in multiple formats (relational database and ROOT)
  - The ROOT storage is via a TTree, and the contents can be viewed using standard ROOT tools.
- Data in TAGS, though are designed for *event selection*, not data monitoring, and not physics analysis.
  - Nevertheless, the storage in a TTree aids in doing simple validations, and there have been cases where errors in data processing have been found first by simple analyses of TAG content.
- TAGS are *more* than an ntuple.
  - TAGS contain navigational information which allow users
    - *To identify datasets and files*
    - *To use them directly as input to athena jobs*
- Many time ***TAG queries*** can look exactly like ntuple selections
  - "RunNumber<430000 && NLooseElectron>4 && triggers(EFmu_20)=1"
  - *Caveat*: triggers and other bitmasks stored in the TAGS require decoding (sometimes time dependent).

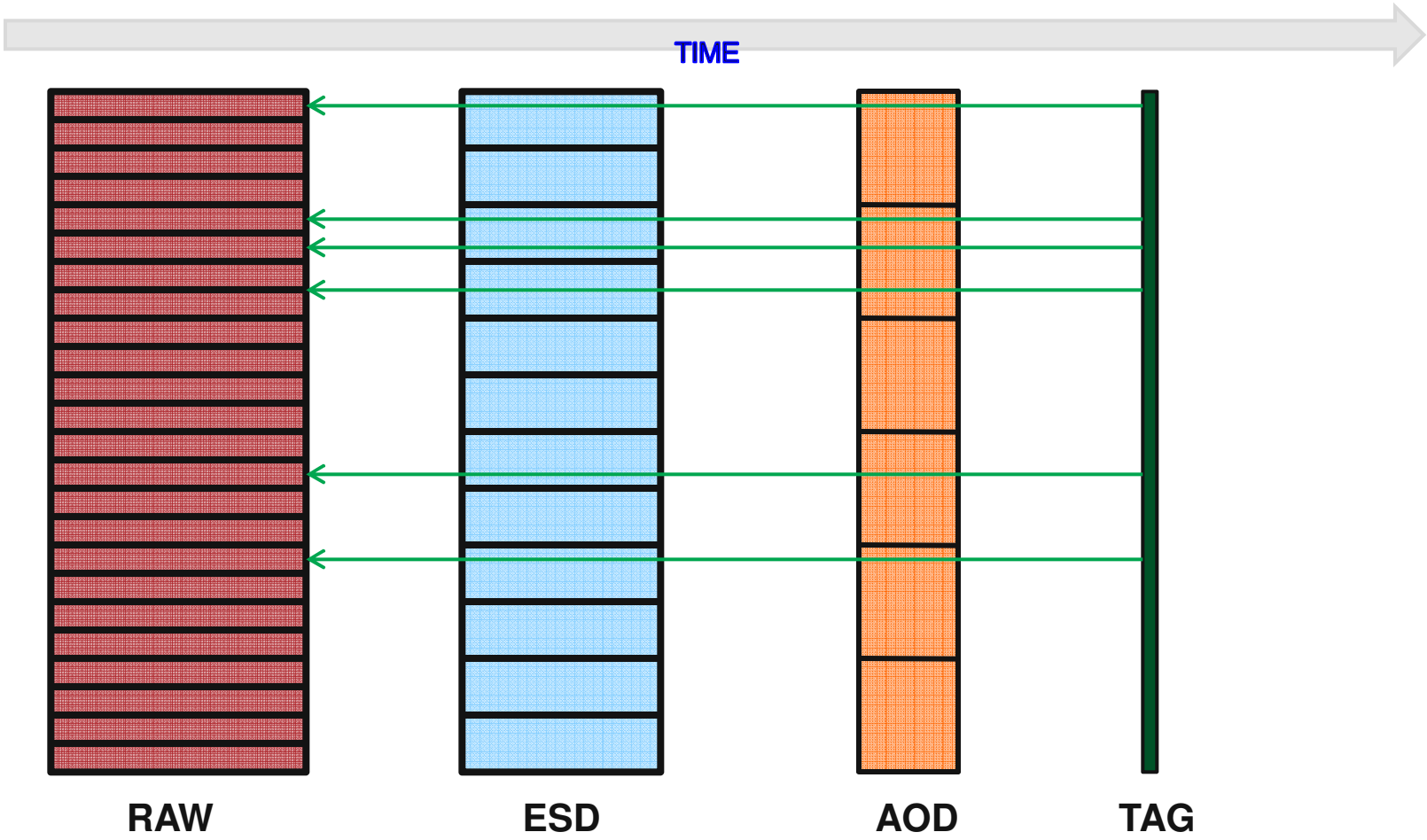
Data processing and TAG building

- ATLAS data goes through several stages of processing.



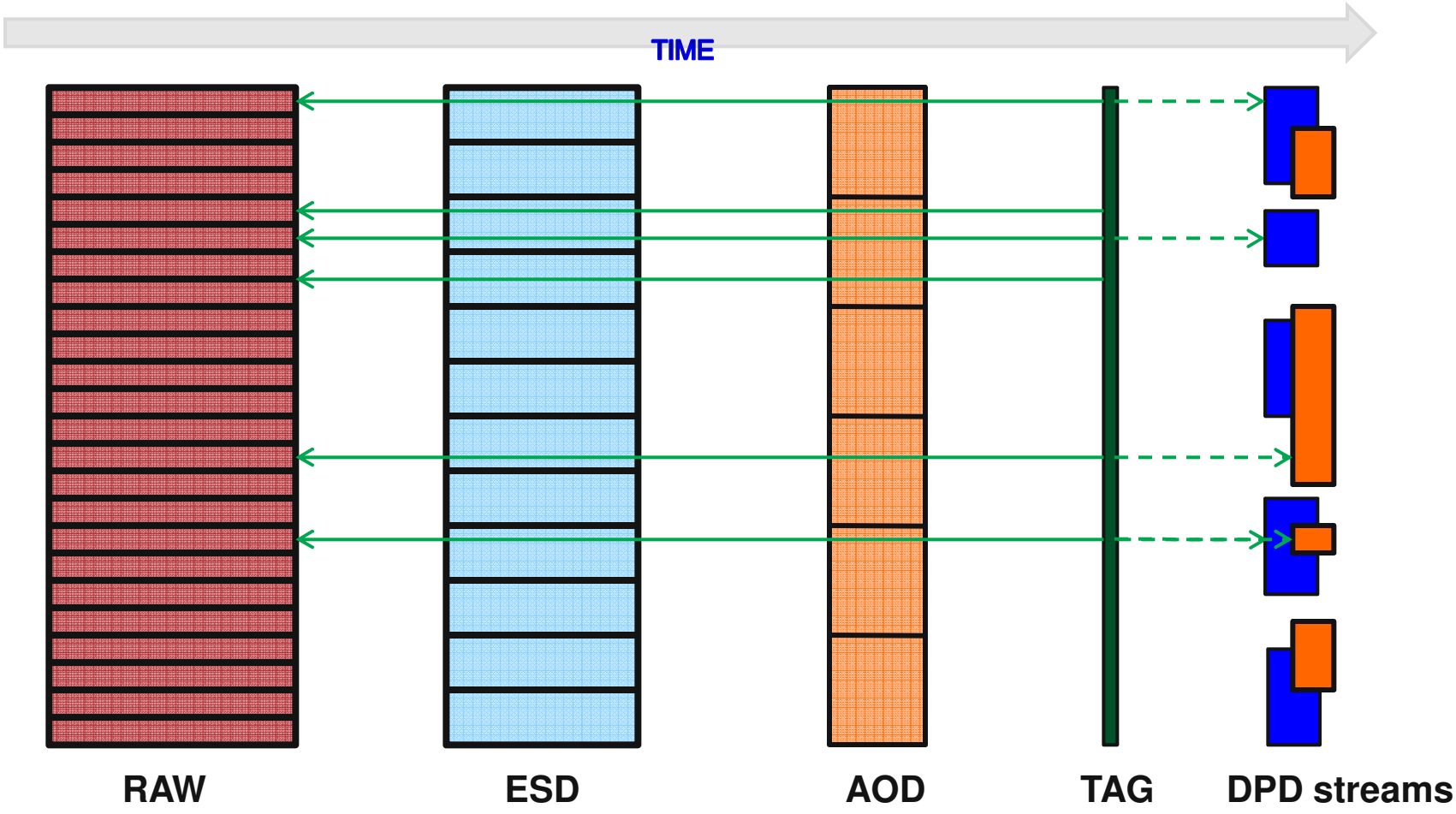
Data processing and TAG building

- TAGS are built at the same time as AOD and contain navigation information for all previous stages.
- http://indico.cern.ch/materialDisplay.py?contribId=38&sessionId=3&materialId=slides&confId=50976



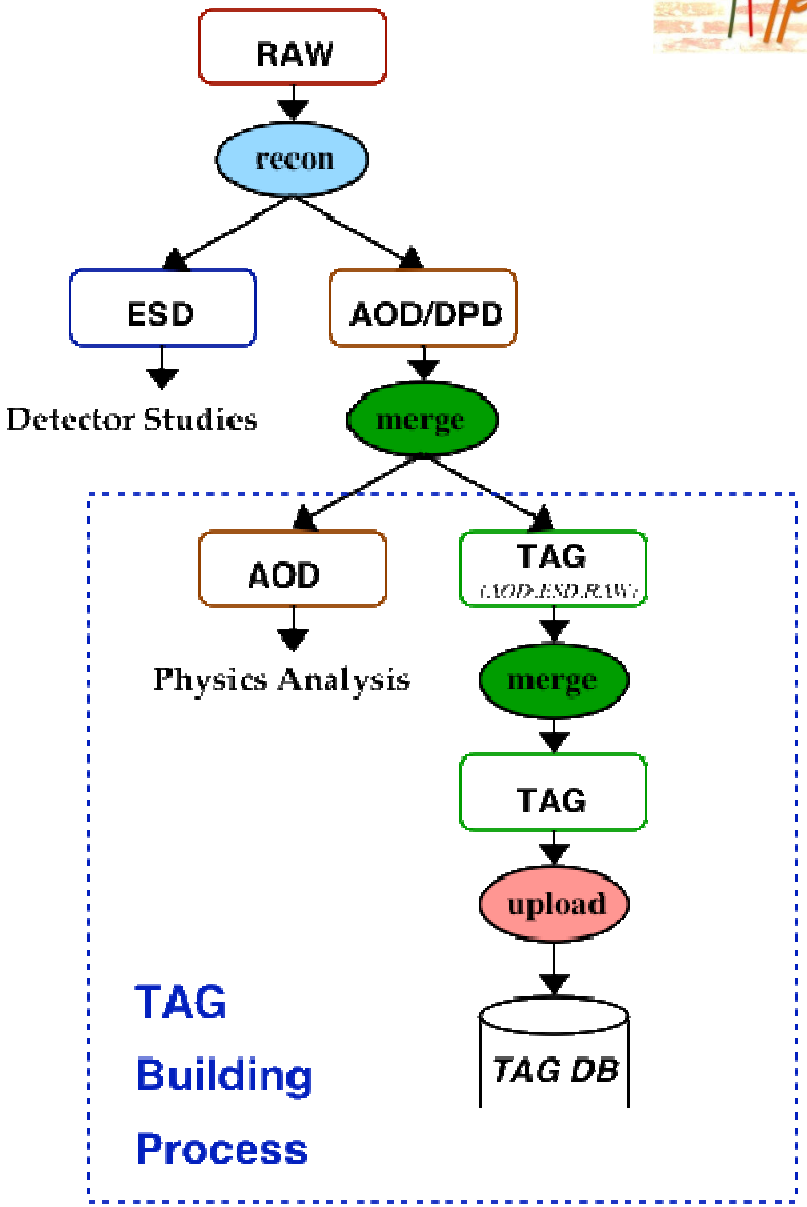
Data processing and TAG building

- TAGS are built at the same time as AOD and contain navigation information for all previous stages.
- TAG capabilities have been and will continue to evolve rapidly.



More on TAG Processing

- TAG files are produced in ROOT files.
- These ROOT files are grouped into datasets and made available through DQ2.
- These files are also uploaded into relational database.
- TAG data in the relational database is made available through various services which aid in event selection and metadata browsing.



Use Case I : Simple use by athena

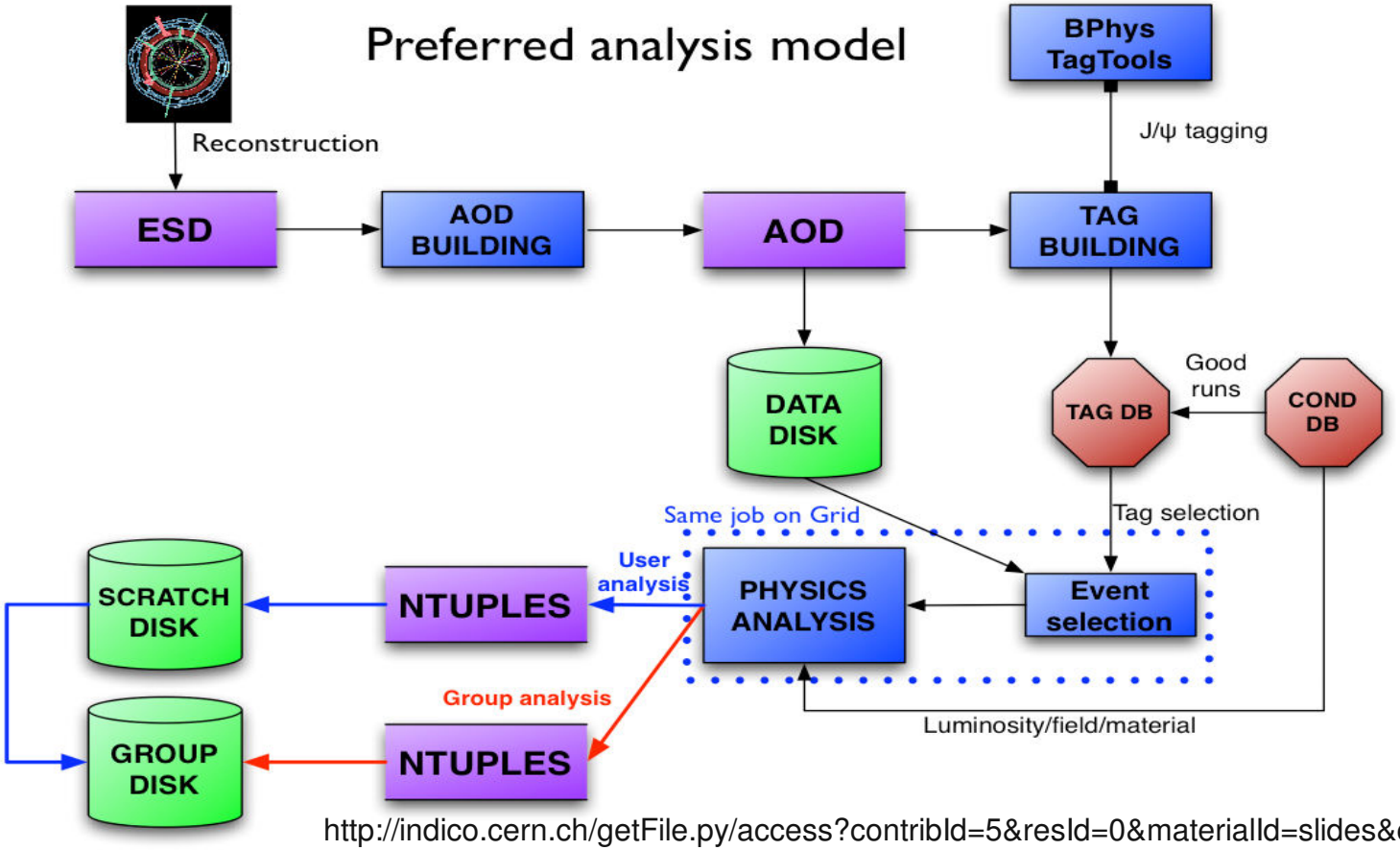
- As you've already learned, athena input is set using the EventSelector. This includes using TAGS as input.
- Simple, circular example
  - Take a POOL file and build a TAG file from it using
    - *athena.py –c "In=['yourPOOLFile.root']; Out=myTAG.root" TagCollectionTest/MakeRunEventCollection.py*
  - This will create a TAG file myTAG.root with a TAG with only the run and event numbers. Go ahead and open it in root and take a look. Pick an event number.
  - TAGS only contain file identifiers, so they need a catalog to find the files. So now do
    - *pool_insertFileToCatalog yourPOOLFile.root*
  - Take the job options you were using to analyze yourPOOLFile.root and change the EventSelector to have
    - *EventSelector.InputCollections = ["PFN:myTAG.root"]*
    - *EventSelector.CollectionType = "ExplicitROOT"*
    - *EventSelector.Query = "EventNumber==100"*

Use Case II: Analyzing commission data

- Put down your tricycle, it's time to drive the big rig.
- TAGS are meant as a tool for examining the cumulative ATLAS event store, so you will develop tools using the first use case, but the real work will occur on the grid.
- *Specific Example*: The TAGS described previously have been developed for ATLAS physics running, but a specialized TAGCOMM has been developed and is being used for looking at commissioning data.
- This is being used by users at CERN to pick out interesting and anomalous events that have shown up during cosmic data taking.
- The US uses a grid job submission system called Panda, and a tool exists for running athena using Panda called pathena. The transition from using athena locally to using pathena has been made as simple as possible. Most cases involve the switch
  - athena myJobOptions.py
  - pathena –inDS=allTheATLASData myJobOptions.py
- Now that you're sold, here are the disclaimers. Many things are not automatic, and jobs *will* fail. There is a learning curve.
- An example, https://twiki.cern.ch/twiki/bin/view/Atlas/PathenaTagComm09

Use Case III: Plans for first data

- Plans for usage of TAGS with first data are being developed by the physics and performance groups. Here is an example from J/Psi, Upsilon walkthrough last week.



http://indico.cern.ch/getFile.py/access?contribId=5&resId=0&materialId=slides&confId=67059

What isn't in TAGS and Why

- TAGS don't contain data quality information.
  - Data quality is assigned after TAGS are made and may be analysis specific.
  - Data quality is assigned at the run or lumiblock level.
  - This information *is* entered into COOL and can be made available in an integrated metadata interface (see next talk).
- TAGS don't contain all physics objects in AOD.
  - Space and other technical limitations.

Other Resources

- ATLAS software tutorials
  - http://indico.cern.ch/conferenceDisplay.py?confId=65526
- Metadata mailing list
  - atlas-event-metadata@cern.ch