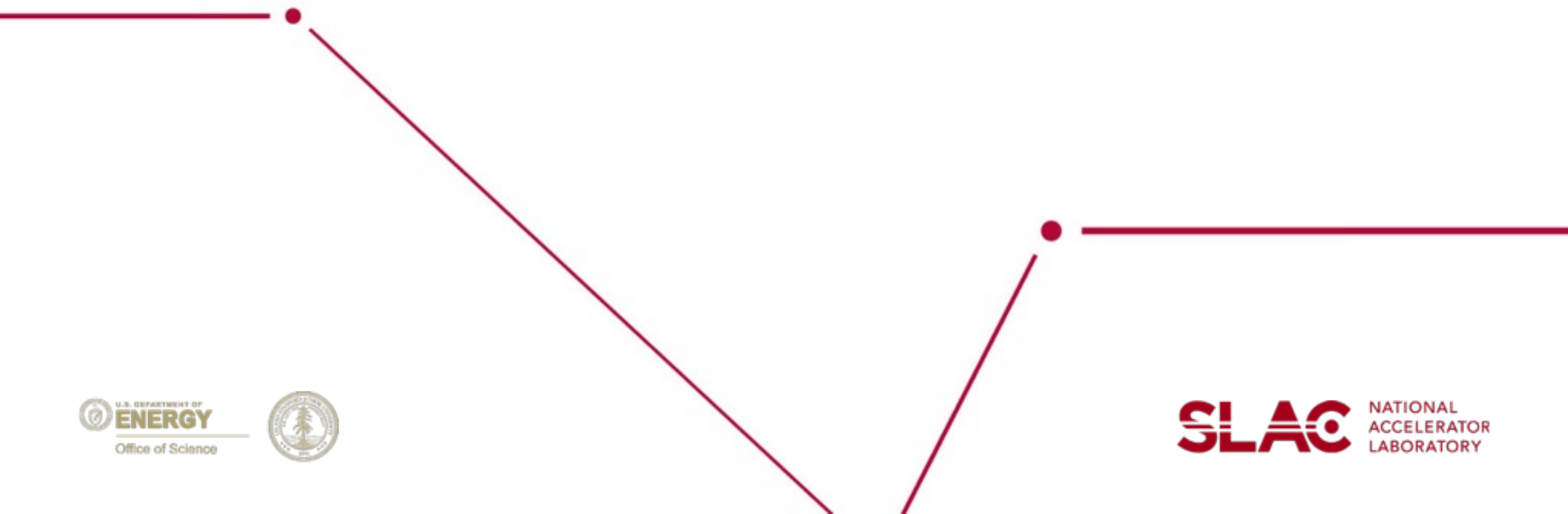
Data Quality Monitoring

Matt Graham, SLAC
HPS Collaboration Meeting
Monday, October 26, 2015



Online vs offline monitoring

- Different goals/needs between these two
  - Online: make sure the data we're writing to disk is usable; e.g, the detector is working correctly, good quality beam, etc. If something looks wrong, figure out what it is and fix it. Need rapid access to data & quick turnaround (doesn't mean we can't monitor high level things)
  - Offline: document the quality (or just the attributes) of the data; maybe, decide whether it is "physics quality"; facilitate observation of long term trends (or just run-to-run wonky stuff). Can be done anytime after a run, but quick turnaround is good; attributes (numbers & plots) should be saved and available for future reference…

Ingredients for quality data quality monitoring

- quick turnaround from the time the data is recorded to monitoring plots & numbers
- infrastructure and actual code that does the monitoring
- people willing & able to look at the output in a timely manner

Ingredients for quality data quality monitoring

- quick turnaround from the time the data is recorded to monitoring plots & numbers
  - data from control room to tape ASAP
    - can we read & write to the disk at same time?
  - once data on tape run recon & DQM right away
    - currently recon takes >12 hours/file (~250k events)
      - run on sub-set of events (first 50k/file?) to get plots out faster?
- infrastructure and actual code that does the monitoring
- people willing & able to look at the output in a timely manner

Ingredients for quality data quality monitoring

- quick turnaround from the time the data is recorded to monitoring plots & numbers
- infrastructure and actual code that does the monitoring
  - org.hps.analysis.dataquality has existed for ~2 years and has lots and lots (and lots) of plots
  - it is and has been run in the production data and MC recon scripts for every file
    - it runs on recon'ed files and is very fast
  - there are some hooks in there for calculating e.g. averages and putting them in a db (though we need to test it…it's been a while)
  - currently the output is a ROOT file with lots of histograms; can also spit out AIDA
    - In fact, Jeremy has found a really nice web-based browser that we should start using!
- people willing & able to look at the output in a timely manner

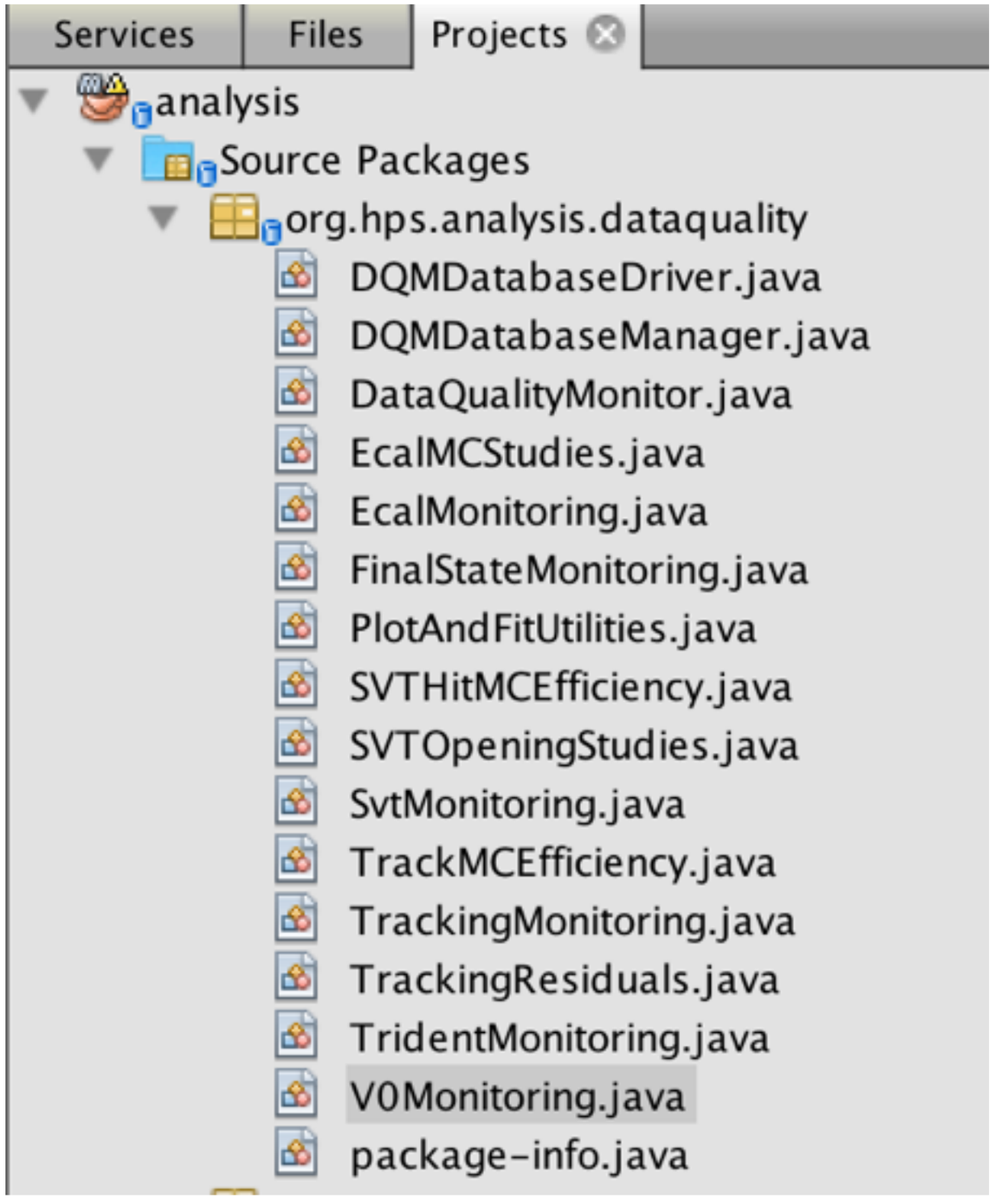
Ingredients for quality data quality monitoring

- quick turnaround from the time the data is recorded to monitoring plots & numbers
- infrastructure and actual code that does the monitoring
- people willing & able to look at the output in a timely manner
  - does the collaboration think this is an important thing?
  - this is where we really dropped the ball in the engineering run…we had discussed a plan prior to the run and never followed through

Here's a plan

- We should be looking at data (& recon) quality:
  - right after data is taken (pass-0)
  - after every recon pass
  - for all production MC samples (per pass as needed)
- We need to have some organization …. a proposal
  - DQM manager — oversees the organization, code infrastructure, common tools etc
  - sub-"system" experts: ECal, SVT, Trigger, Tracking, Analysis…a representative who's responsible for what gets monitored, the code for that, and making sure the data gets looked at (signs of on "quality").
- During data taking, there should be daily reports from each sub-systems DQM rep about the previous day's data quality ("all good"!)
  - this can be done remotely of course

org.hps.analysis.dataquality



As I mentioned, there is a lot of code that exists already; I've (and Sho, I think) used this package extensively

If anything, there are probably too many plots in here now. Sub-system reps should go through their part and organize as needed; separate vital DQ plots/ numbers vs supplementary ones

An example of Jeremy's web viewer

AIDA Plots

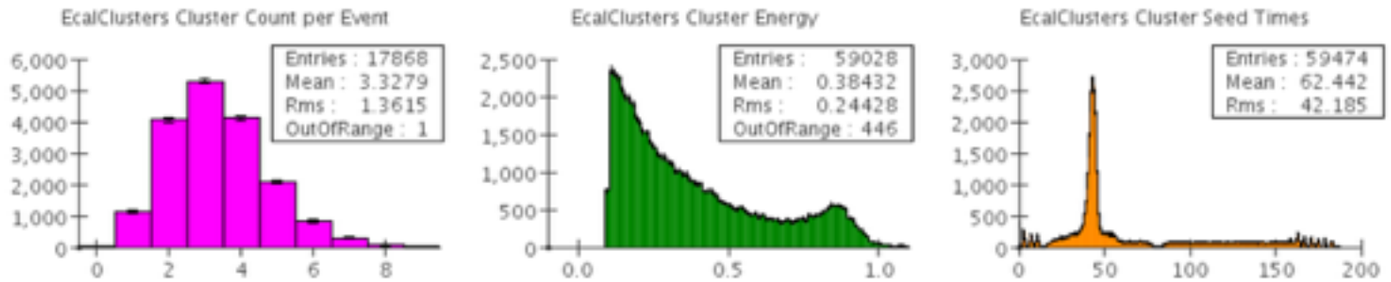
File: /work/hps/recon/dqm_test.aida

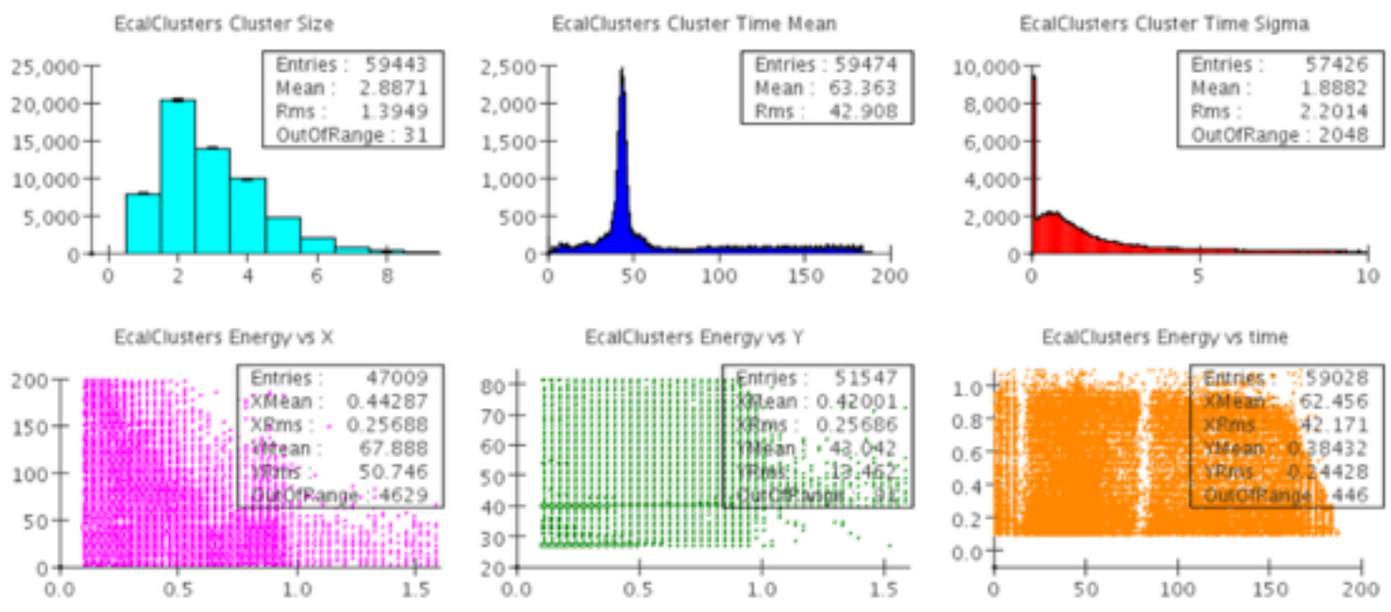
/
- EcalClusters (0 / 100)
  - all (15 / 20)
  - pairs0 (15 / 20)
  - pairs1 (15 / 20)
  - singles0 (15 / 20)
  - singles1 (15 / 20)
- EcalHits (0 / 25)
- FinalStateParticles (0 / 120)
- SvtMonitoring (0 / 504)
- TrackResiduals (0 / 360)
- Tracks (0 / 850)
- TridentMonitoring (0 / 105)
- V0Monitoring (0 / 260)

Path: /EcalClusters/all

Style Editor

15 plots selected. Showing 1 to 9. [First/ Previous] 1 , 2 [Next / Last]





Download: (*vector*) eps svg pdf swf emf ps , (*bitmap*) jpg png ppm gif

Work to do

- If we are serious about this, get organization together soon
- additions/subtractions/re-org of plots to look at
- set up DQM database for mean/sigma/counts-per-file tracking (if needed)
- set up online (as in, the internet) DQM plot browser
- make sure we follow through with the plan