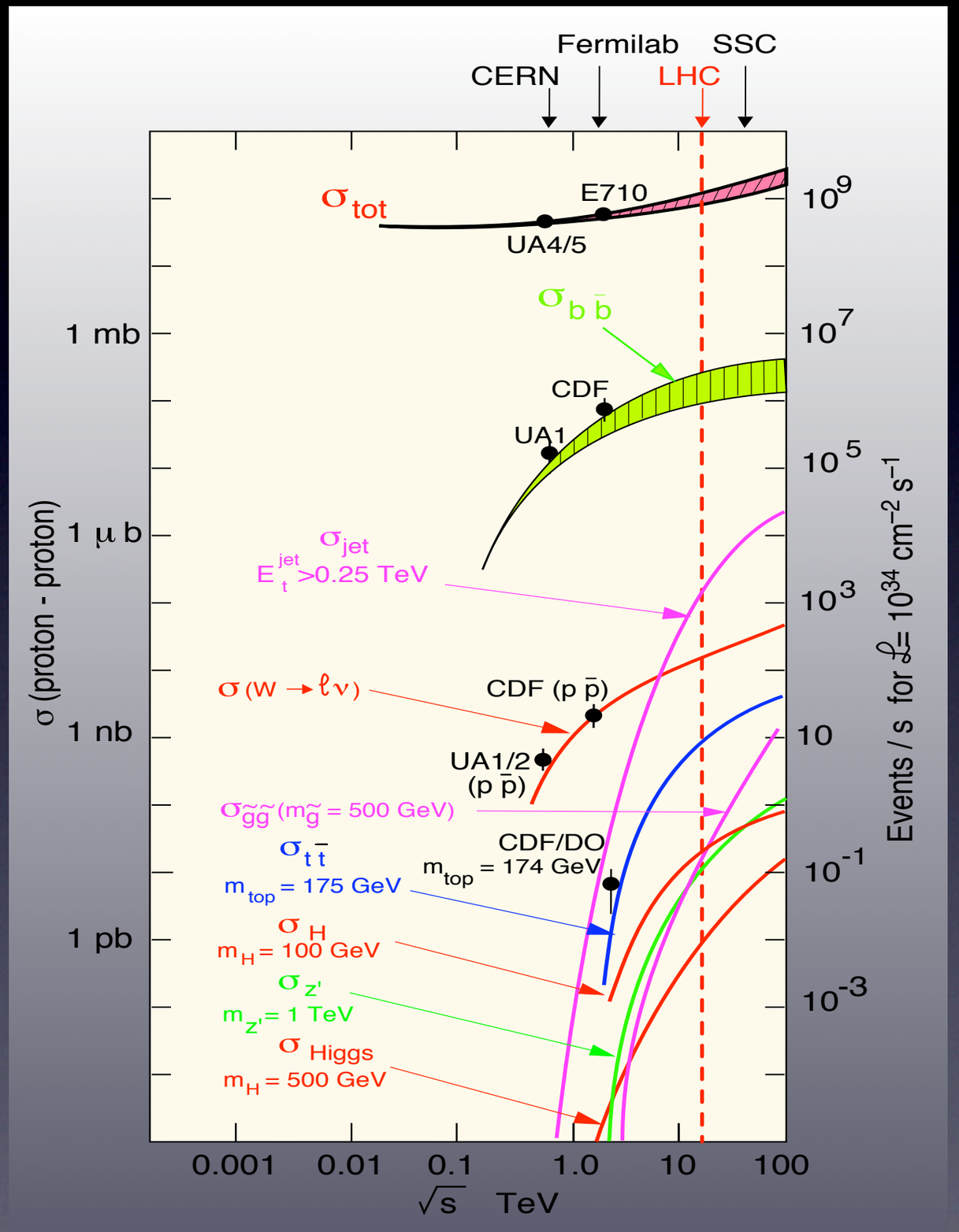
Analysis Computing

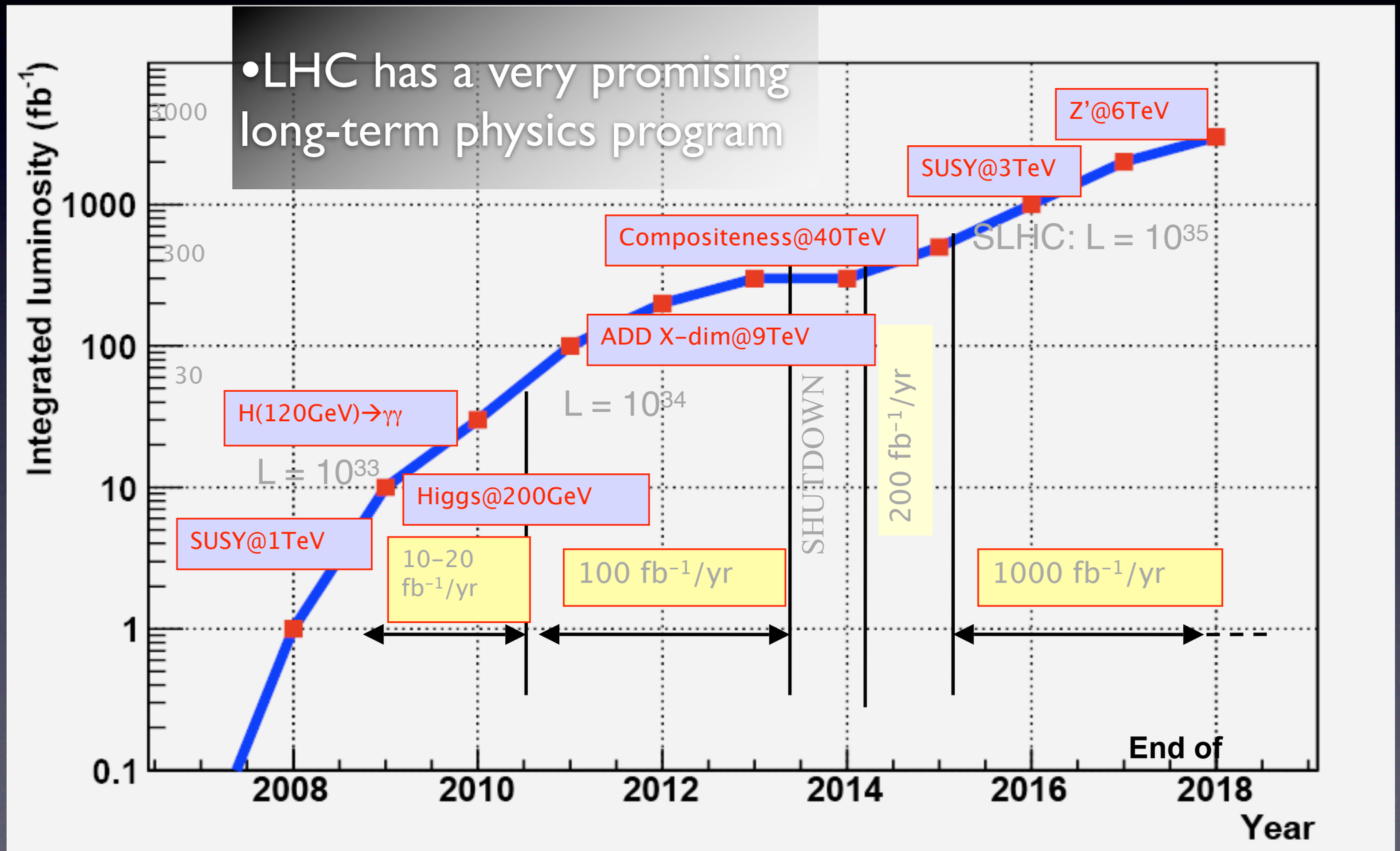
Amir Farbin
CERN

LHC Data

- At $L=10^{34}$ cm^{-2} s^{-1} (~100x less at startup):
  - $W \rightarrow l\nu$, $Z \rightarrow ll$ ~10^2 Hz
  - top at 10 Hz
  - Higgs at 1- 10^{-1} Hz (m_H=100 - 600 GeV)
  - SUSY up to 10 Hz (depending on scale)
- Significant increase in SM x-sections over Tevatron ⇒ Lots of control samples to quickly:
  - Understand detector
  - Tune MC to 14 TeV
- Great potential for early discovery.

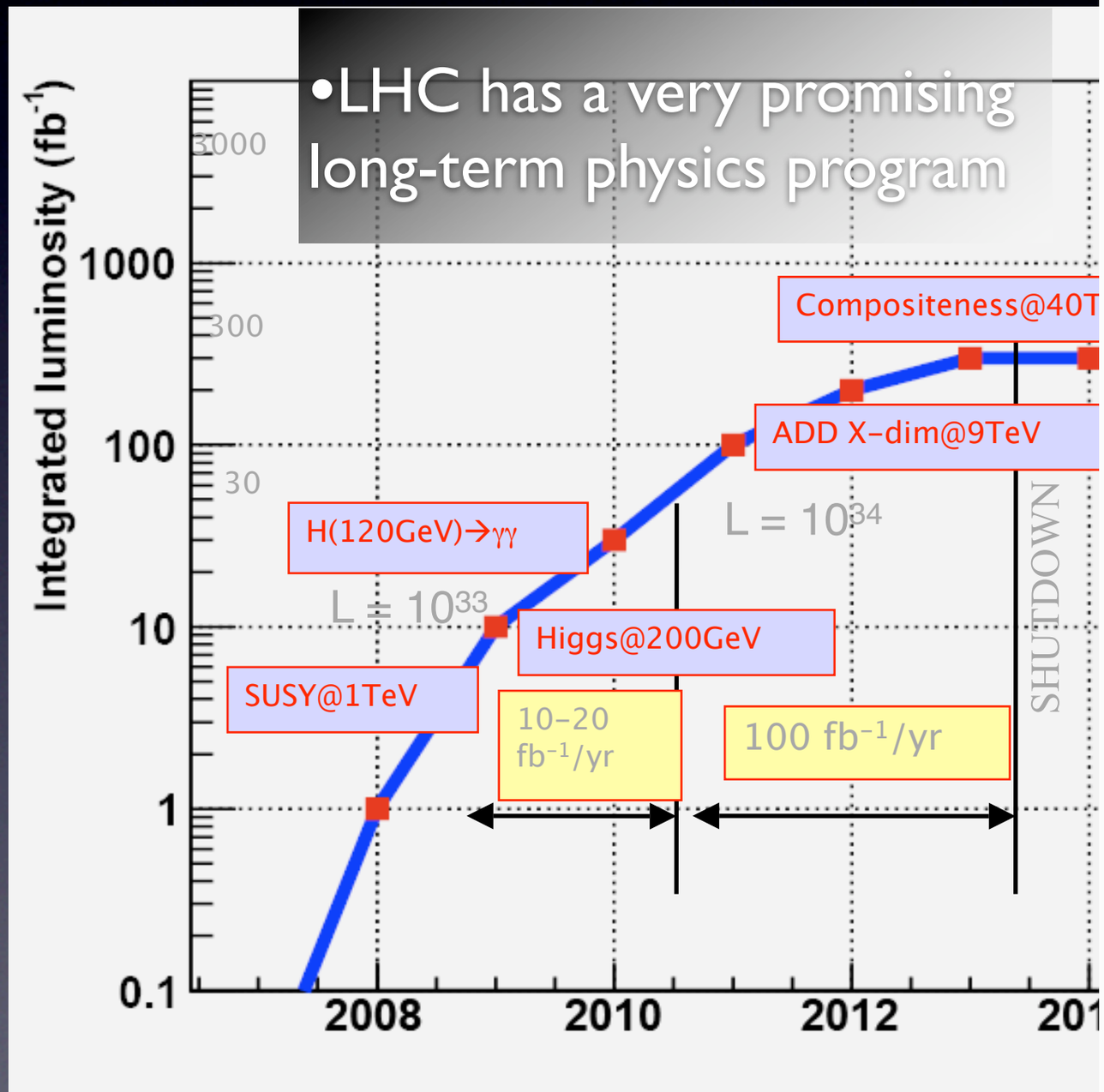


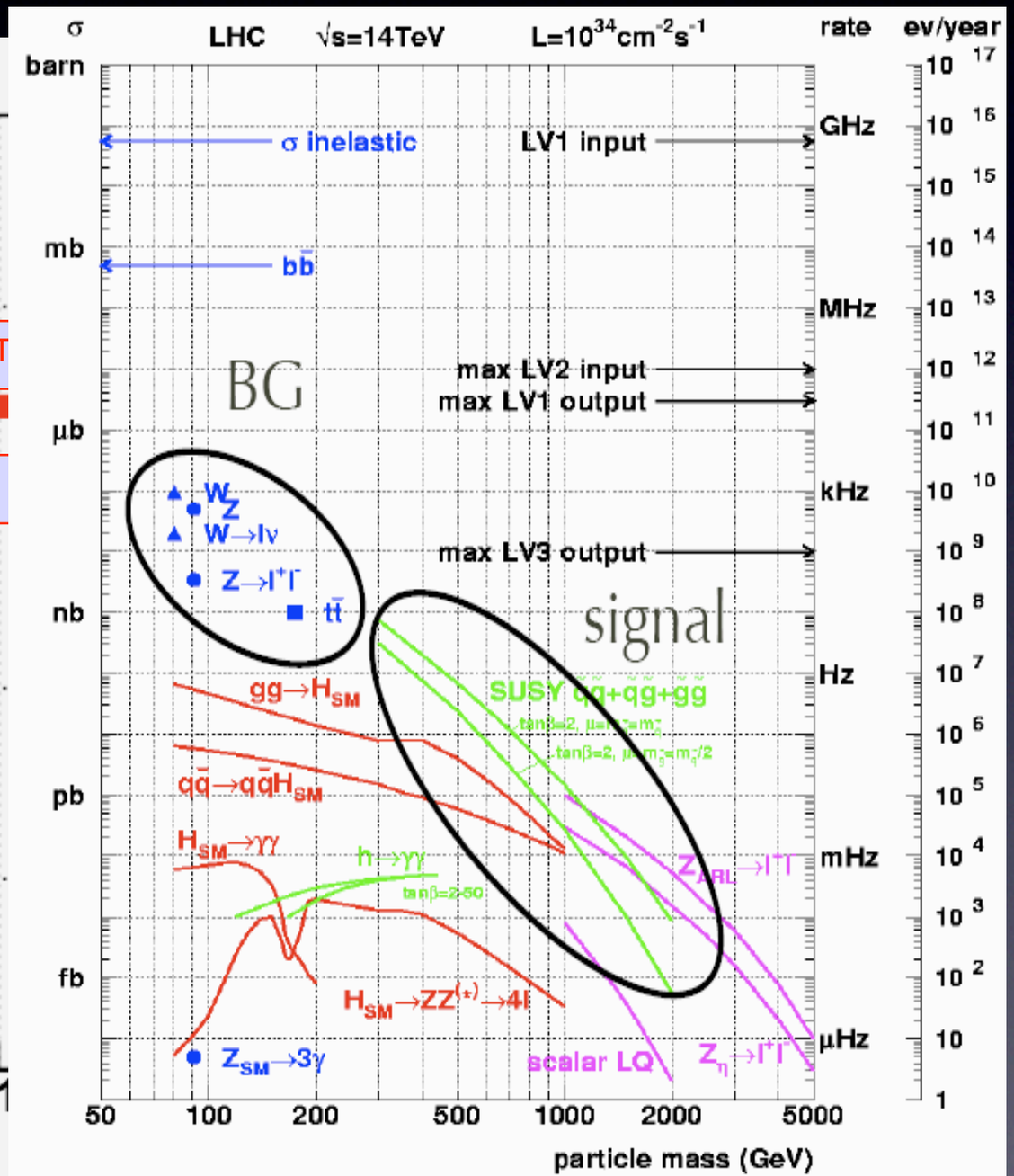
New Physics in 2008?



New Physics in 2008?

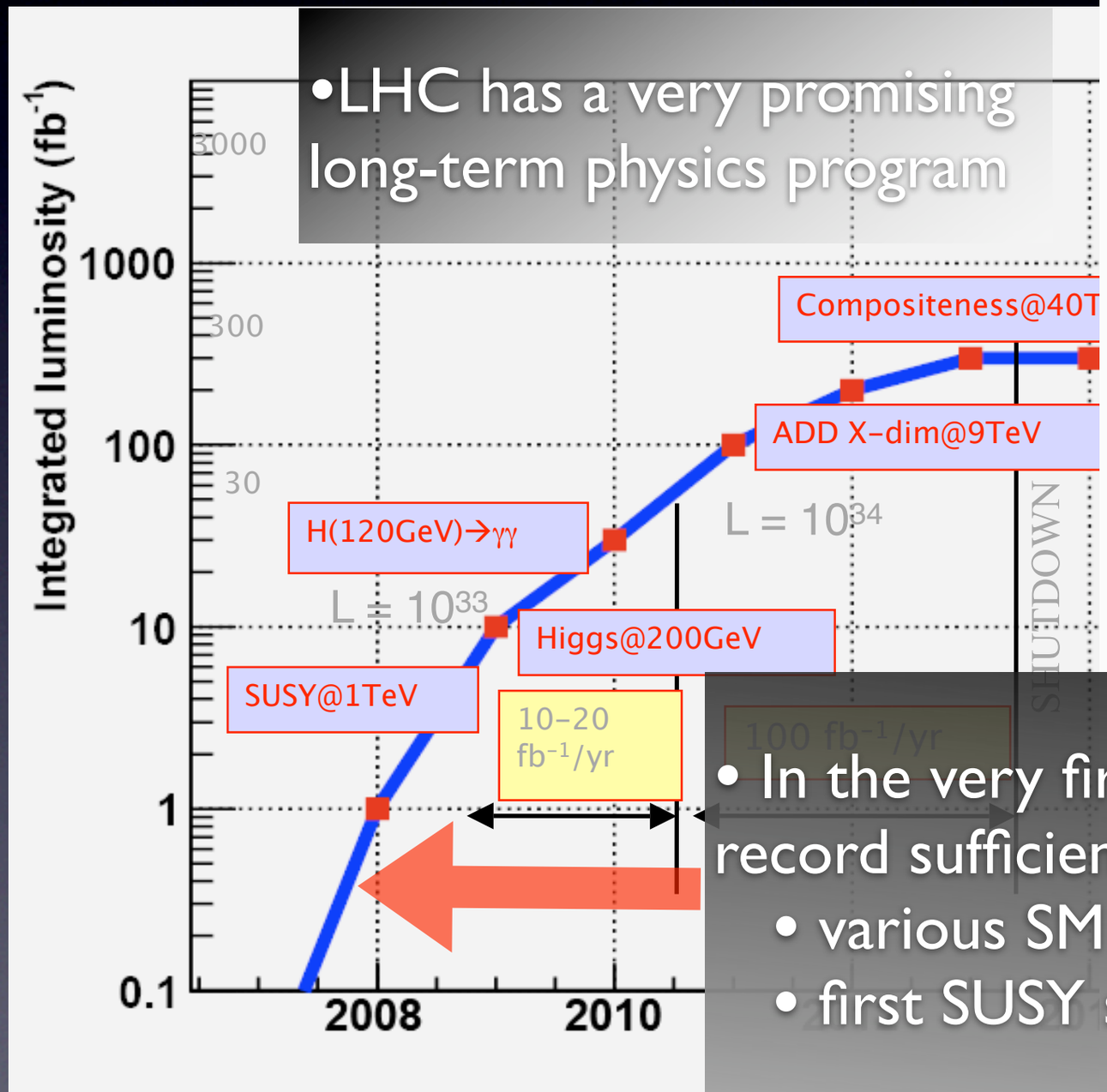
- LHC has a very promising long-term physics program

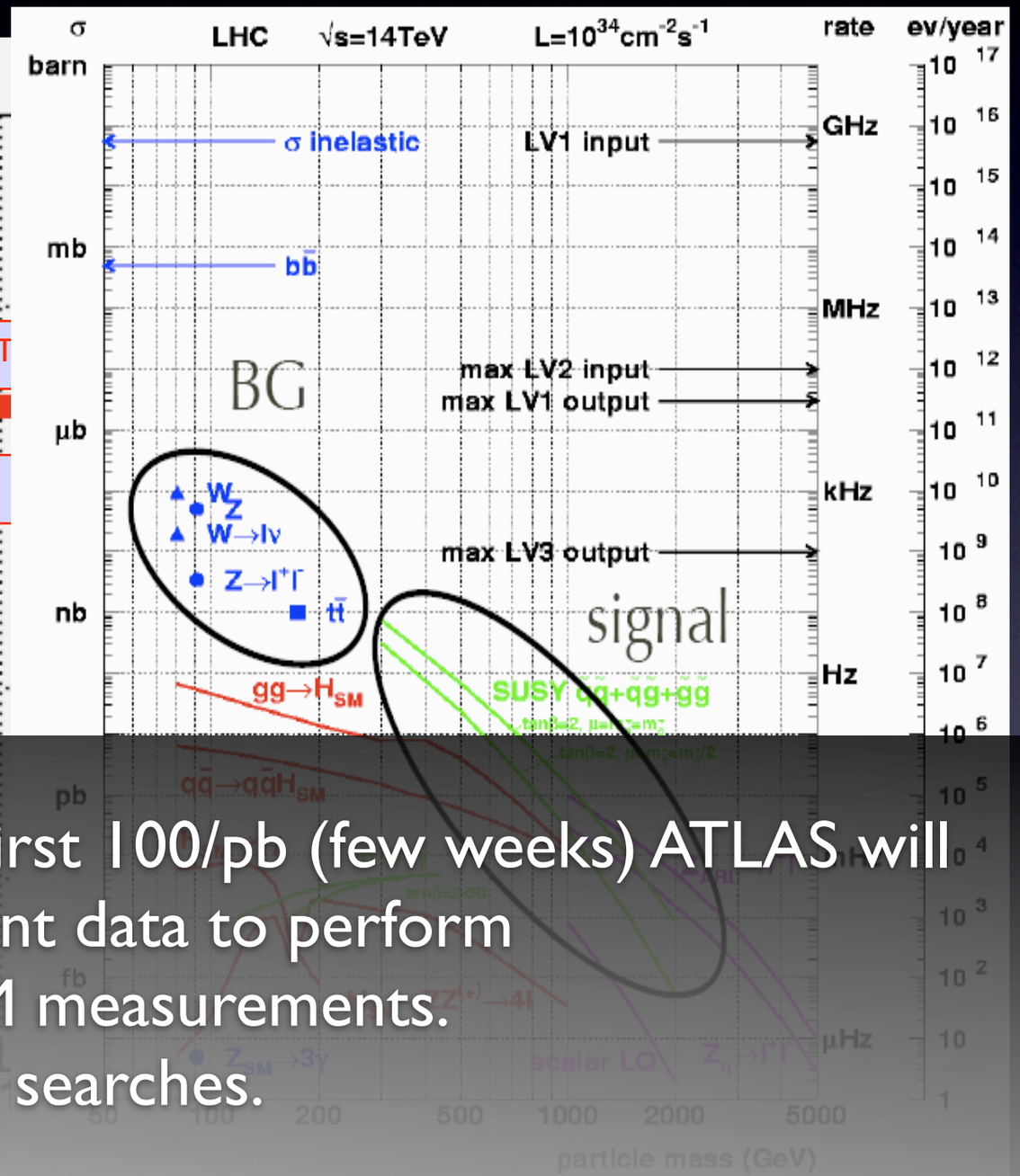




New Physics in 2008?

- LHC has a very promising long-term physics program





- In the very first 100/pb (few weeks) ATLAS will record sufficient data to perform
  - various SM measurements.
  - first SUSY searches.

From Data to Measurement

- Immediately after first data
  - Must understand/optimize efficiencies/acceptance, fake rates, scale, resolution, tails.... trigger.
  - Understand Standard Model "backgrounds": reconstruct W, Z, top events.
  - Ultimately analyze much of the collected data samples: W (+Jets), Z (+Jets), top, Jet+Jet, γ+Jet, ...

- Once the data is recorded extracting results is a matter of
  - man-power + organization
  - software + computing infrastructure
- Tevatron took years to publish Run II results...
  - many D0/CDF colleagues have commented on insufficient software preparation
- Can we do better?

Computing in HEP

Full Simulation

KHz

Generation

mHz

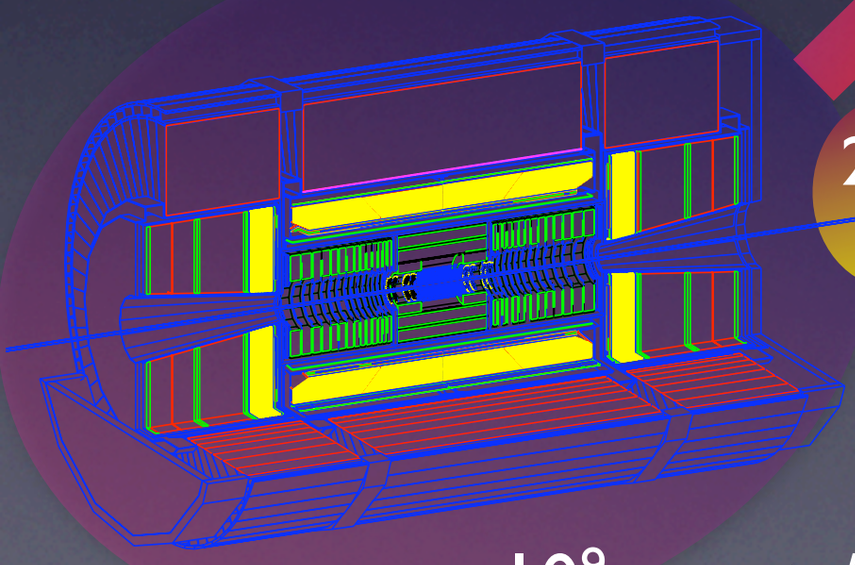
Simulation

Hz

Digitization

Will only simulate 20% of data

High-level Trigger



200 Hz

10^9 events/year

Data Store

cHz

Reconstruction

Data Base

Fast Simulation

KHz

Generation

Hz

Fast Simulation

KHz

Hz

Algorithmic Analysis

MHz

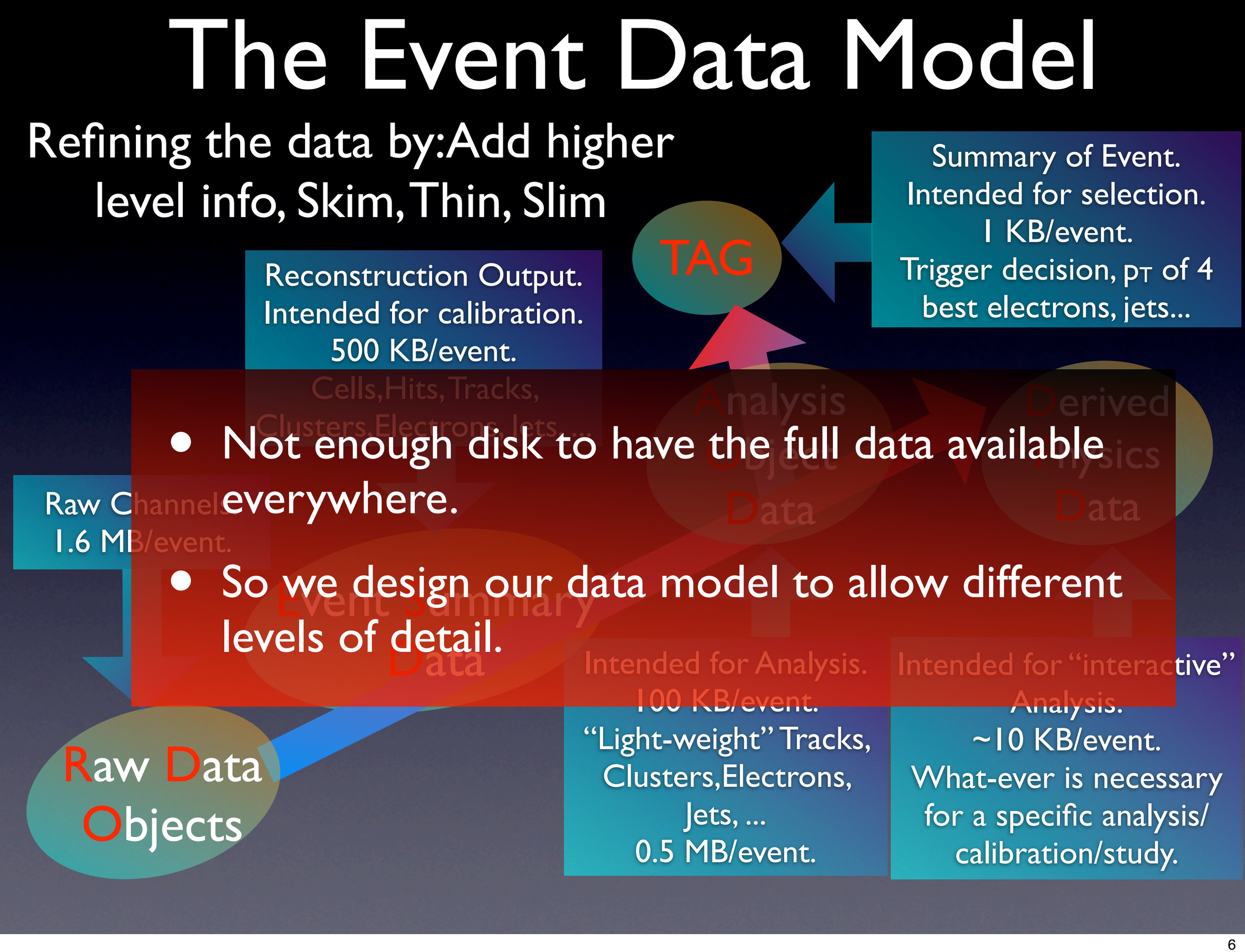
Interactive Analysis

Statistical Analysis

Data Analysis & Calibration

The Event Data Model

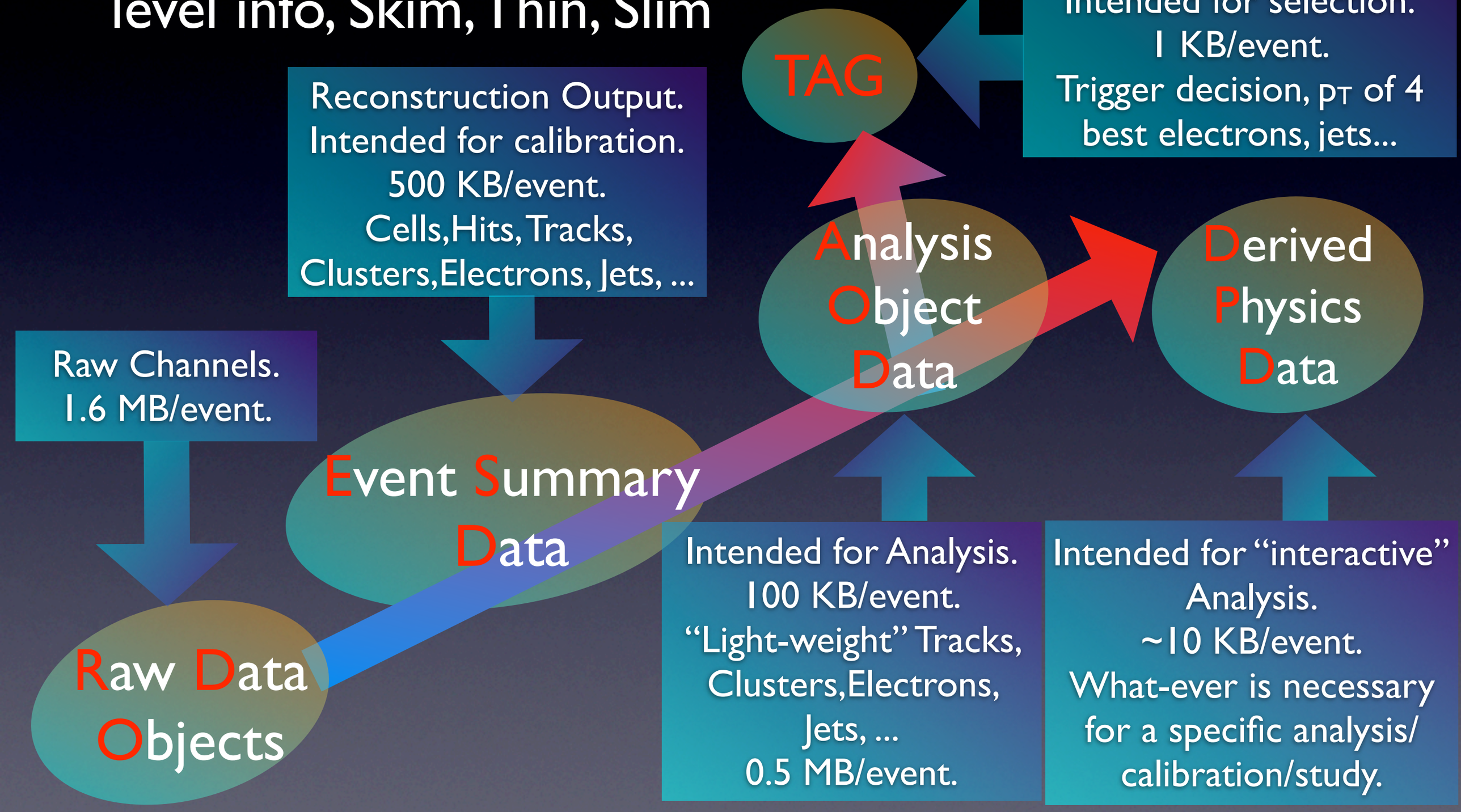
Refining the data by:Add higher level info, Skim, Thin, Slim



- Not enough disk to have the full data available everywhere.
- So we design our data model to allow different levels of detail.

The Event Data Model

Refining the data by:Add higher level info, Skim, Thin, Slim



Event Data Model (EDM)

EDM Level	Contents	Primary Intent	Size/ Event (KB)	Max Ideal Input rate (Hz)	Access-ibility
Raw Data Objects	Raw Channels	Reconstruction (calibration)	1600	N/A	Central Reco/ Reprocessing: Tier 0/1
Event Summary Data	Cells, Hits, Clusters, Tracks, MET, Electron, Jet, Muon, Tau, Truth	Re-reconstruction, Re-calibration	500		CERN CAF (access limited), Tier 1 (on tape)
Analysis Object Data	Clusters, Tracks, MET, Electron, Jet, Muon, Tau, Slimmed Truth	Limited Re-reconstruction (eg Jets, b-tag), limited re-calibration, Analysis	100	1000	Full: Tier 1,2 (disk) Subset: Tier 3
Derived Physics Data	Any of the above + composites (eg top) + derived quantities (sphericity)	Interactive Analysis: Making plots, performing studies	Typically ~10	10^6	Tier 3: eg your laptop
TAG	Summary. Ex: p_T, η of 4 best e,γ,μ,τ,jet	Selection Events for analysis	1	10^8	Everywhere

The GRID

- Resources Spread Around GRID

•Derive 1st pass calibrations within 24 hours.
•Reconstruct rest of the data keeping up with data taking.

•Reprocessing of full data with improved calibrations 2 months after data taking.
•Managed Tape Access: RAW, ESD
•Disk Access: AOD, fraction of ESD

30 Sites Worldwide
Tier 2

AOD

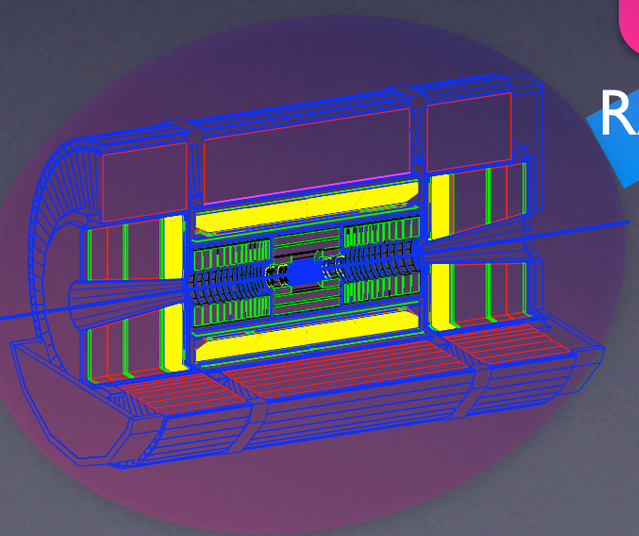
Tier 1
10 Sites Worldwide

•Production of simulated events.
•User Analysis: 12 CPU/ Analyzer
•Disk Store: AOD

RAW/ AOD/ ESD

Tier 0

RAW



CERN Analysis Facility

•Primary purpose: calibrations
•Small subset of collaboration will have access to full ESD.
•Limited Access to RAW Data.

The GRID

- Resources Spread Around GRID

•Reprocessing of full data with improved calibrations 2 months after data taking.
•Managed Tape Access: RAW, ESD
•Disk Access: AOD, fraction of ESD

30 Sites Worldwide

Tier 2

•Derive 1st pass calibrations within 24 hours.
•Reconstruct rest of the data keeping up with data taking.

• The ATLAS Computing Model cannot handle analysis activity on the ESD.
• Analysis must be performed on the AOD.

➡ Important to make sure that the AOD meets analysis requirements.

➡ Important to provide sufficient redundancy and flexibility in AOD to recover from unexpected problems.

AOD

RAW/ ESD/ AOD

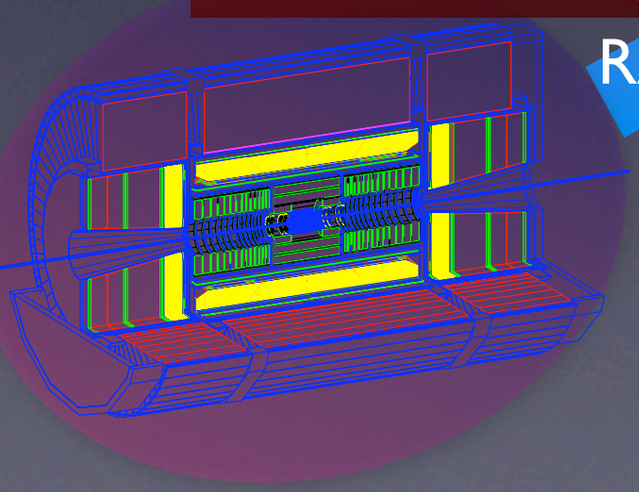
Tier 1

10 Sites Worldwide

•Production of simulated events.
•User Analysis: 12 CPU/ Analyzer
•Disk Store: AOD

Tier 0

RAW



CERN Analysis Facility

•Primary purpose: calibrations
•Small subset of collaboration will have access to full ESD.
•Limited Access to RAW Data.

Delayed Response

- ATLAS will collect data at constant rate (200Hz) regardless of luminosity. ⇒ 10^9 events/year.
- Optimistic estimate of % of 10^9 events (1year) analyzed in realistic analysis (~100x slower today):

Analysis job run time	1 CPU (1 MB/s) Laptop	25 CPUs (25 MB/s) 2 people	100 CPUs (100 MB/s) 4 people	1000 CPUs (1 GB/s) WG	
1 hr	.016%	.41%	1.7%	16%	Interactive Analysis (final step)
1 day (12 hrs)	.2%	5%	20%	All	Batch Interactive Analysis (final step)
1 wk. (150 hrs)	2.7%	70%	All	All	Batch Analysis (intermediate step)
1 mo. (700 hrs)	12%	All	All	All	Centralize Skim (intermediate step)

- Scaling issues at computing sights will also be an important factor.
  - ➡ Running analysis on every available CPU will break the system.

➡ Users need to be smart about their analysis strategy:

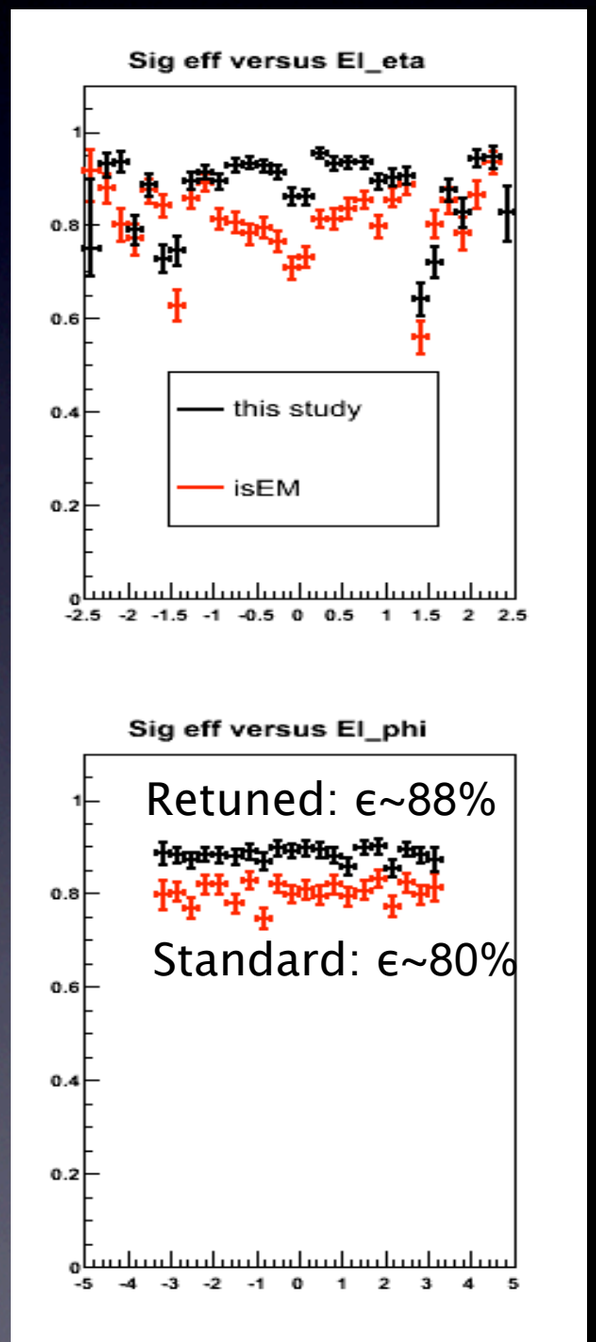
- Perform analyses collectively
- Analyze in multiple steps: slow steps a few times ⇒ DPD ⇒ fast steps many times

Analysis Model

- Given these constraints and the complexity of the tasks ahead... it is important to have a plan of how to analyze LHC data.
- Analysis Model: An attempt to ensure physics needs are met by the ATLAS software.
- Lots of recent developments based on experience from other experiments:
  - Fundamental software framework features
  - Organization of our data
  - Tools to collaboratively tackle complex tasks

Analysis Model Basics

The Right Data in the Right Place



- A Simple Example:
  - The standard ATLAS Electron identification selection is coded into the Electron reconstruction and stored with the Electron

➡ Difficult to retune.

➡ Remained the same for 2 years while software/understanding improved.

- So we made the necessary electron variables available at analysis time (AOD).

➡ Electron Selection tuned in context of analysis.

➡ 8% better selection efficiency for same jet rejection on SUSY events.

➡ Improvement can be distributed to others w/o reprocessing the data.

Analysis Model Basics

Redundant Solutions

	Jets	Electrons	Missing Et
ESD All Calo Cells (not	Default Calibrate clusters to	Default Calibrate cells to EM	Default Build Missing Et from calibrated clusters + rgy in
Al (C +u sam cells	Build jets from uncalibrated clusters, calibrate based on energy samplings	Choose electron cluster, recalibrate cells, re-calc shower shapes, re-calibrate electron	from re-calibrated hard objects (eg jet, electron) + remaining contributions.
clusters (available for analysis)			

- Hypothetical Scenario:
  - 2 months from target conference, ATLAS discovers low level calorimeter calibration problem which hinders various measurements.
  - Not enough time to correct, reprocess, and redistribute data.

Analysis Model Basics

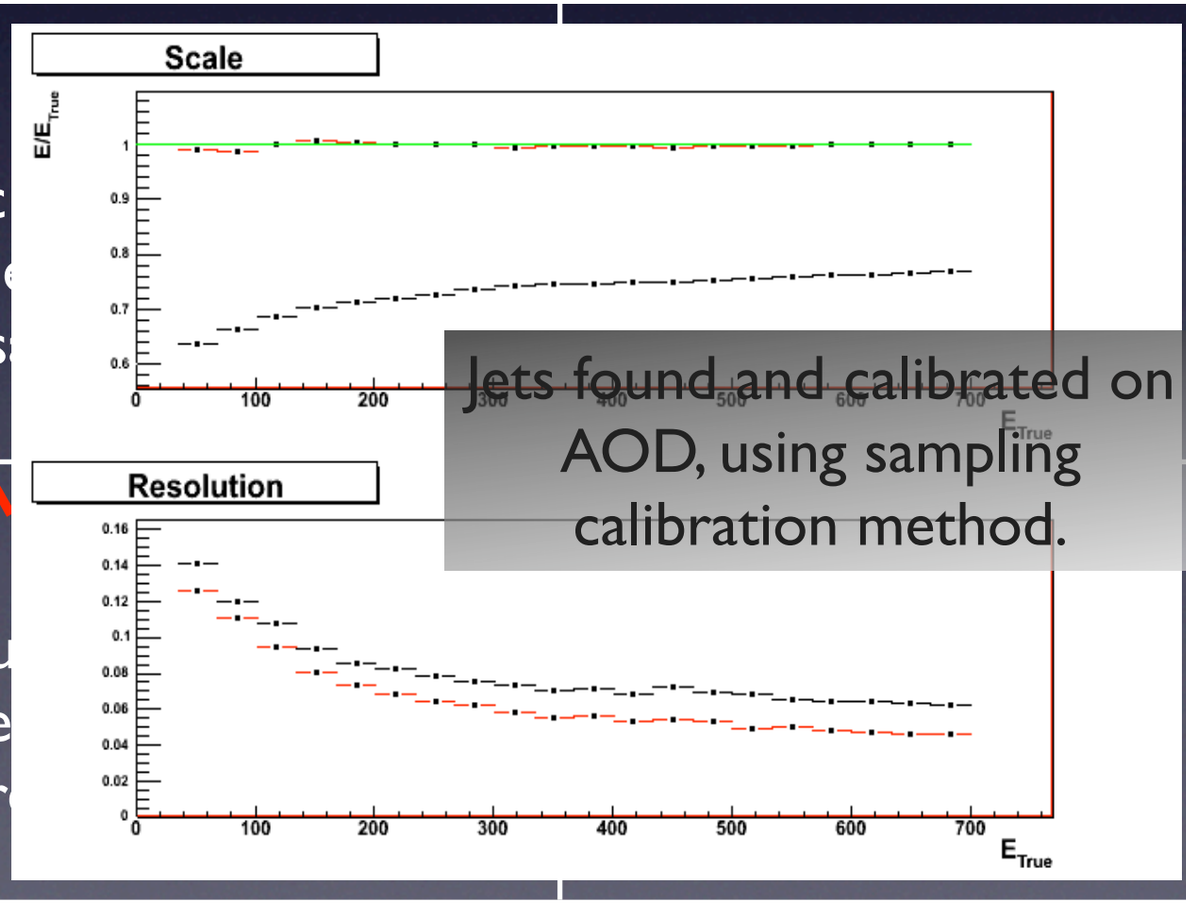
Redundant Solutions

	Jets	Electrons	Missing Et
ESD All Calo Cells (not available for analysis)	Default Calibrate clusters to hadronic scale based on cells	Default Calibrate cells to EM scale	Default Build Missing Et from calibrated clusters + out of cluster energy in cells. Save in components.
AOD All Clusters (Calibrated +uncalibrated samplings), All cells in electron clusters (available for analysis)	Build jets from calibrated clusters, apply "out-of-cone"/Jet Alg Corrections	Choose electron cluster size, calibrate electrons based on samplings in clusters	Build Missing Et from individual contributions.
	Plan B Build Jets From uncalibrated clusters, calibrate based on energy samplings	Plan B Choose electron cluster, recalibrate cells, re-calc shower shapes, re-calibrate electron	Plan B Build Missing Et from re-calibrated hard objects (eg jet, electron) + remaining contributions.

Analysis Model Basics

Redundant Solutions

	Jets	Electrons	Missing Et
ESD All Calo Cells (not available for analysis)	Default Calibrate clusters to hadronic scale based on cells	Default Calibrate cells to EM scale	Default Build Missing Et from calibrated clusters + out of cluster energy in cells. Save in components.
AOD All Clusters (Calibrated +uncalibrated samplings), All cells in electron clusters (available for analysis)	Build jets from calibrated clusters, apply "out-of-cone"/Jet Alg Corrections Plan B Build Jets From uncalibrated clusters, calibrate based on energy samplings	c s Plan clu re r	

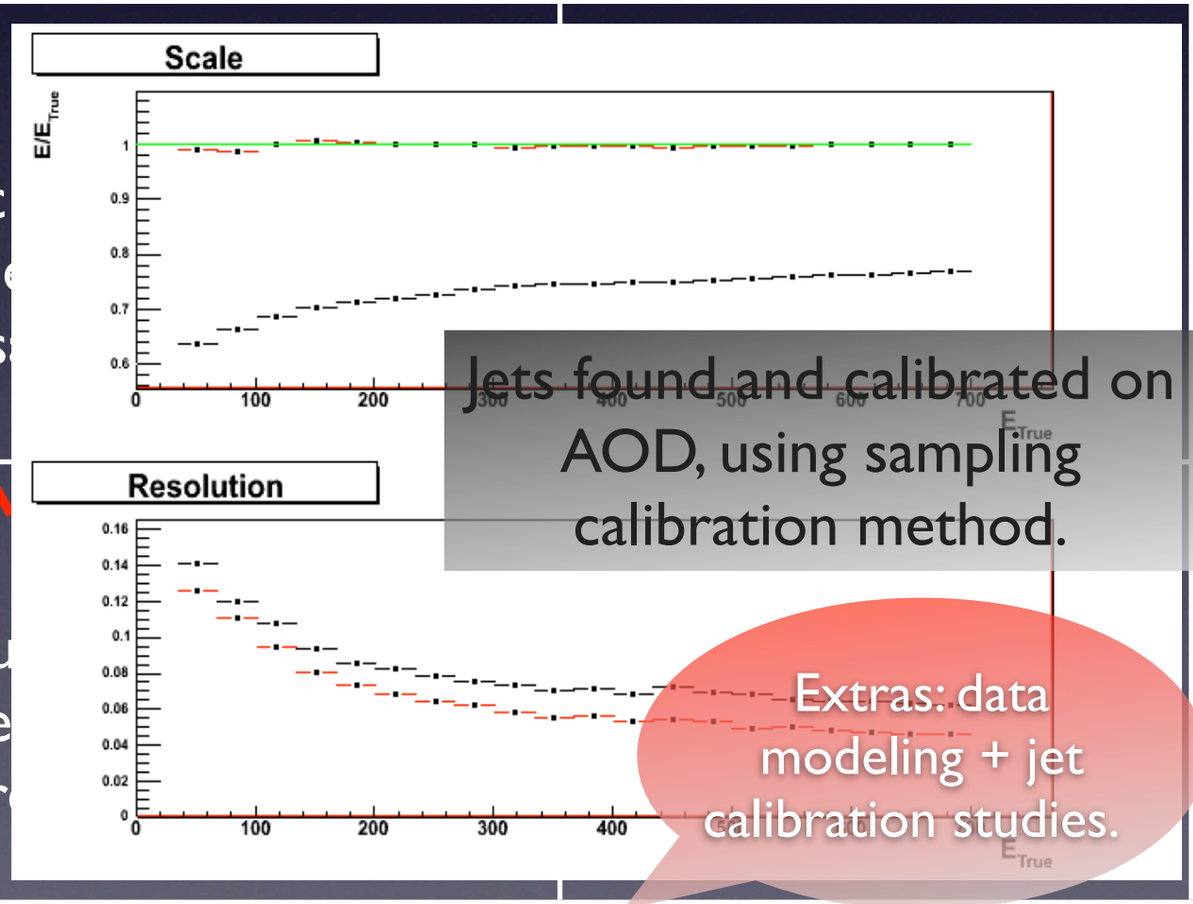


Jets found and calibrated on AOD, using sampling calibration method.

Analysis Model Basics

Redundant Solutions

	Jets	Electrons	Missing Et
ESD All Calo Cells (not available for analysis)	Default Calibrate clusters to hadronic scale based on cells	Default Calibrate cells to EM scale	Default Build Missing Et from calibrated clusters + out of cluster energy in cells. Save in components.
AOD All Clusters (Calibrated +uncalibrated samplings), All cells in electron clusters (available for analysis)	Build jets from calibrated clusters, apply "out-of-cone"/Jet Alg Corrections	c s	
	Plan B Build Jets From uncalibrated clusters, calibrate based on energy samplings	Plan clu re r	



Jets found and calibrated on AOD, using sampling calibration method.

Extras: data modeling + jet calibration studies.

Analysis Model Basics

Redundant Solutions

	Jets	Electrons	Missing Et
ESD All Calo Cells (not available for analysis)	Default Calibrate clusters to hadronic scale based on cells	Default Calibrate cells to EM scale	Default Build Missing Et from calibrated clusters + out of cluster energy in cells. Save in components.
AOD All Clusters (Calibrated +uncalibrated samplings), All cells in electron clusters (available for analysis)	Build jets from calibrated clusters, apply "out-of-cone"/Jet Alg Corrections	Choose electron cluster size, calibrate electrons based on samplings in clusters	Build Missing Et from individual contributions.
	Plan B Build Jets From uncalibrated clusters, calibrate based on energy samplings	Plan B Choose electron cluster, recalibrate cells, re-calc shower shapes, re-calibrate electron	Plan B Build Missing Et from re-calibrated hard objects (eg jet, electron) + remaining contributions.

EDM Lessons from Other Experiments I

Observations from:
BaBar, CDF, D0, H1
ATLAS Analysis Model Workshop (Oct 2006)

- *Observation:* Speed is the most important factor in the Analysis Model adopted by users... no matter what the management says or sw-developers provide.
- AOD access speed (few Hz) has been a concern for a long time.
- It has been impractical to repeatedly iterate analyses on AOD, so users often dump large ntuples which mostly copy AOD contents... and perform analysis outside athena.
- Solution: Transient/Persistent split
  - Transient version of data: the format in memory... optimized for manipulation... stays constant so client code doesn't change when data changes.
  - Persistent version of data: the format stored on disk... optimized for size and speed... can change as deemed necessary.
- It now appears that AOD speed can close to the ROOT limit (10MB/s).
- In release 13: AOD and ntuple speed should be comparable.

EDM Lessons from Other Experiments II

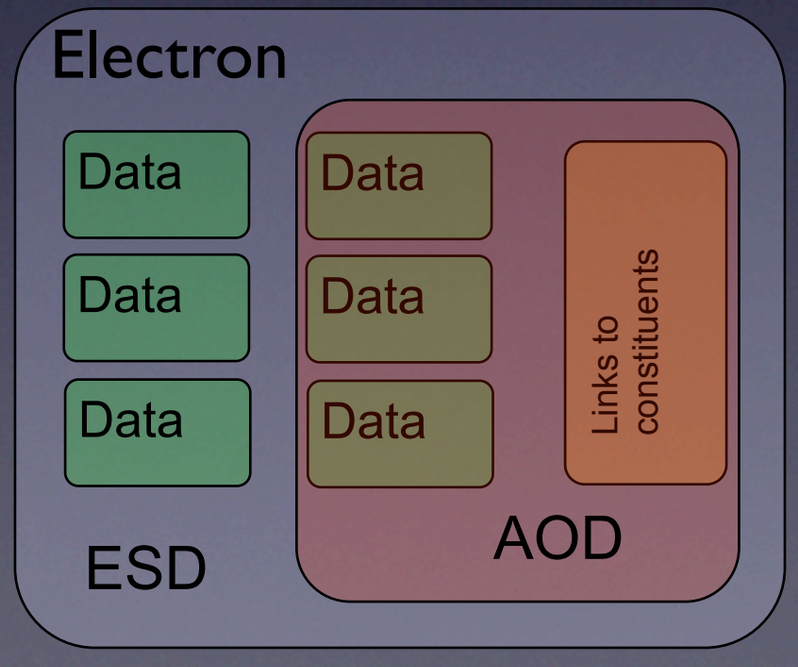
Observations from:
BaBar, CDF, D0, H1
ATLAS Analysis Model Workshop (Oct 2006)

- *Observation:* Tasks naively thought to be addressed by "ESD"-based analysis or reprocessing (eg: calibration, alignment, track-fit, re-clustering) are routinely performed in the highest level of analysis.

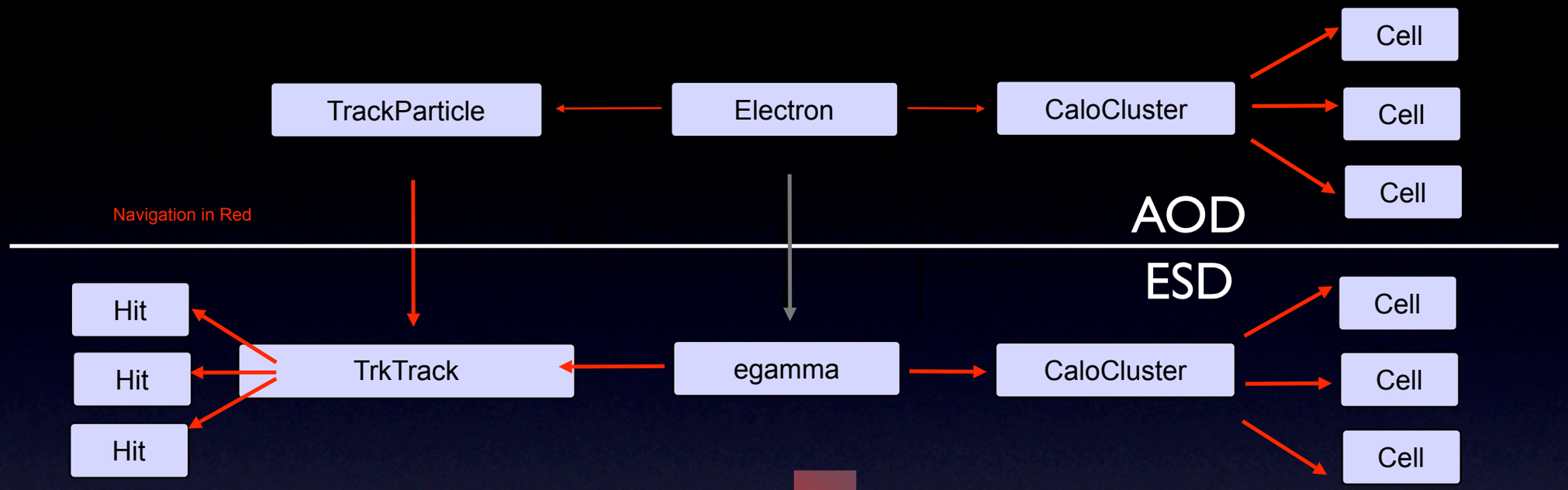
➡ As experiments evolve:

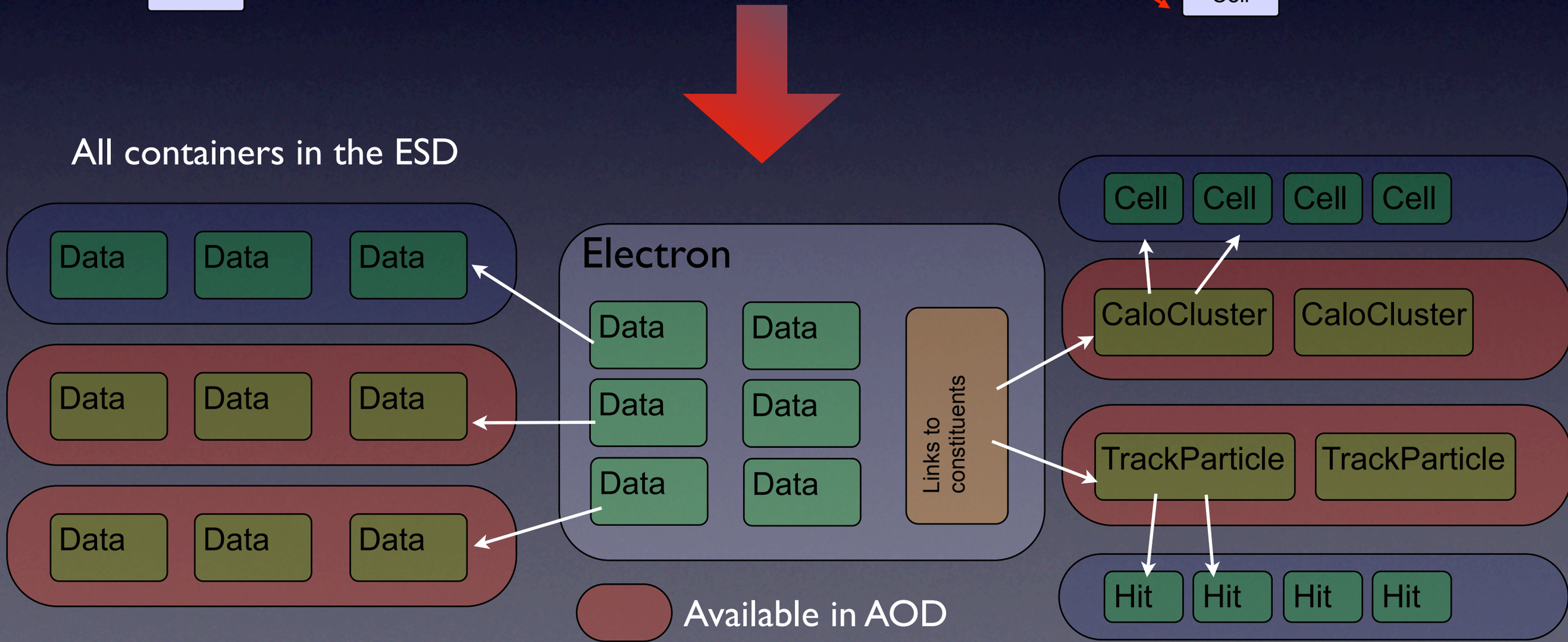
- "ESD" bloated and too difficult to access ⇒ dropped
- "AOD" is gradually augmented with some "ESD" quantities (eg: hits in roads/cells) to provide greater functionality at analysis time.

- Build a flexible data model by merging ESD/AOD format... but keeping separate levels of detail:
  - Analyzers can seamlessly switch between ESD/AOD.
  - Jobs read on demand... speed
  - Anyone can reconfigure data model w/o schema change
    - Move data between levels by changing configuration.
    - No compiled code involved!
    - Seamlessly read data before/after the change.

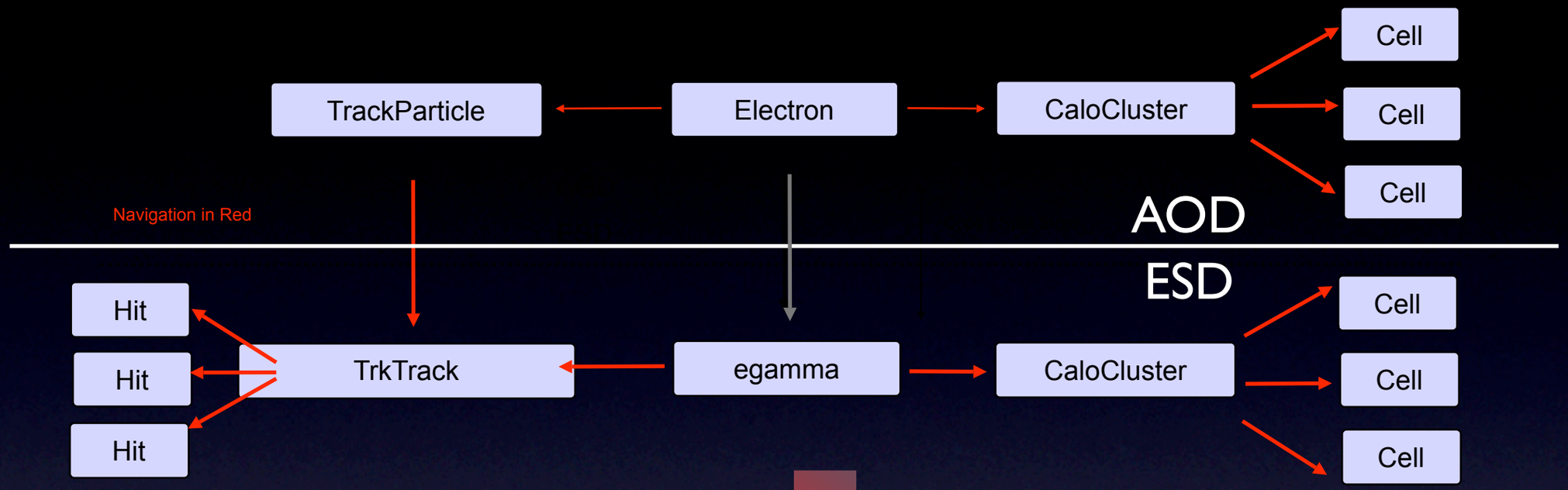


AOD/ESD Merger



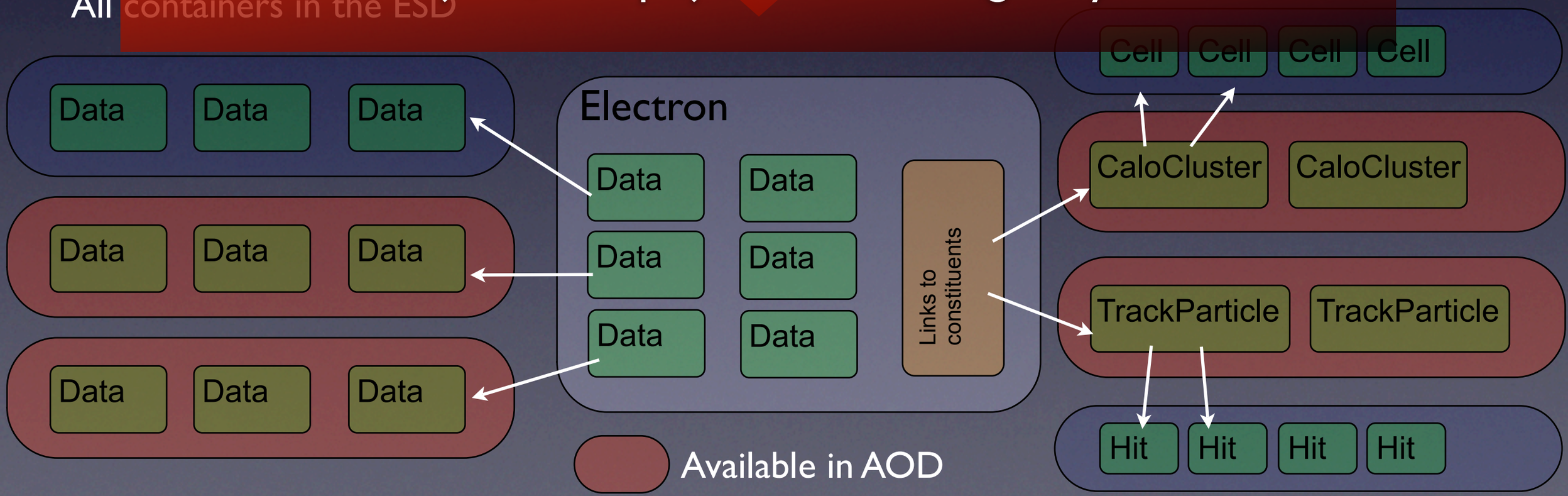


AOD/ESD Merger



- All Particle Objects, except jets will be merged by release 13.

All containers in the ESD



EDM Lessons from Other Experiments III

Observations from:
BaBar, CDF, D0, H1
ATLAS Analysis Model Workshop (Oct 2006)

- Derived Physics Data (DPD) is Traditionally an "ntuple" which can analyzed standalone (eg in ROOT) without the experiment's software framework.
- *Observation:* Any hick-up the experiment software or computing, and physicists by-pass the framework & copy all of the data into DPD format:
  - BaBar: more data in proprietary DPD than AOD. A primary contributor to a complete redesign of computing model.
  - Tevatron: DPD became the AOD. Proprietary frameworks developed by users.
- BaBar (CM2), CMS, and *now* ATLAS solutions (AOD/DPD merger):
  - allow the EDM to be easily extendible with UserData
  - allow the EDM to be read in both framework and ROOT.

EDM Lessons from Other Experiments III

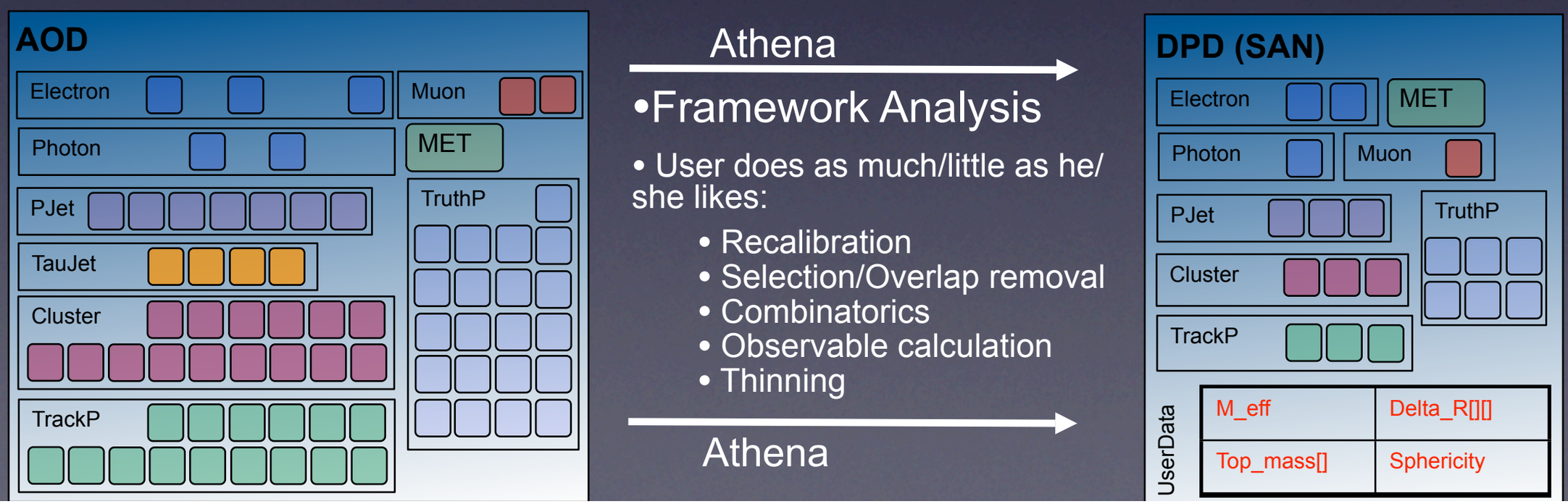
Observations from:
BaBar, CDF, D0, H1
ATLAS Analysis Model Workshop (Oct 2006)

- Derived Physics Data (DPD) is Traditionally an "ntuple" which can analyzed standalone (eg in ROOT) without the experiment's software framework.
- Observation: Any back up for experiment software or computing and physicists by-pass the framework & copy all of the data into DPD format:
  - BaBar: more data in proprietary DPD than AOD. A primary contributor to a complete redesign of computing model.
  - Tevatron: DPD became the AOD. Proprietary frameworks developed by users.
- BaBar (CM2), CMS, and *now* ATLAS solutions (AOD/DPD merger):
  - allow the EDM to be easily extendible with UserData
  - allow the EDM to be read in both framework and ROOT.

- Quick Comment (for "experts"):
  - ATLAS's original plan for interactive analysis was Interactive Athena.
  - Interactive Athena: python prompt + Gaudi + PyRoot + AOD convertors provide a ROOT like analysis environment.
  - Effort stalled because of AOD access speed made Interactive Athena unattractive to users... SAN is a consequence.
  - With faster AOD, we should revisit Interactive Athena... SPyRoot is prototype of how this can look.

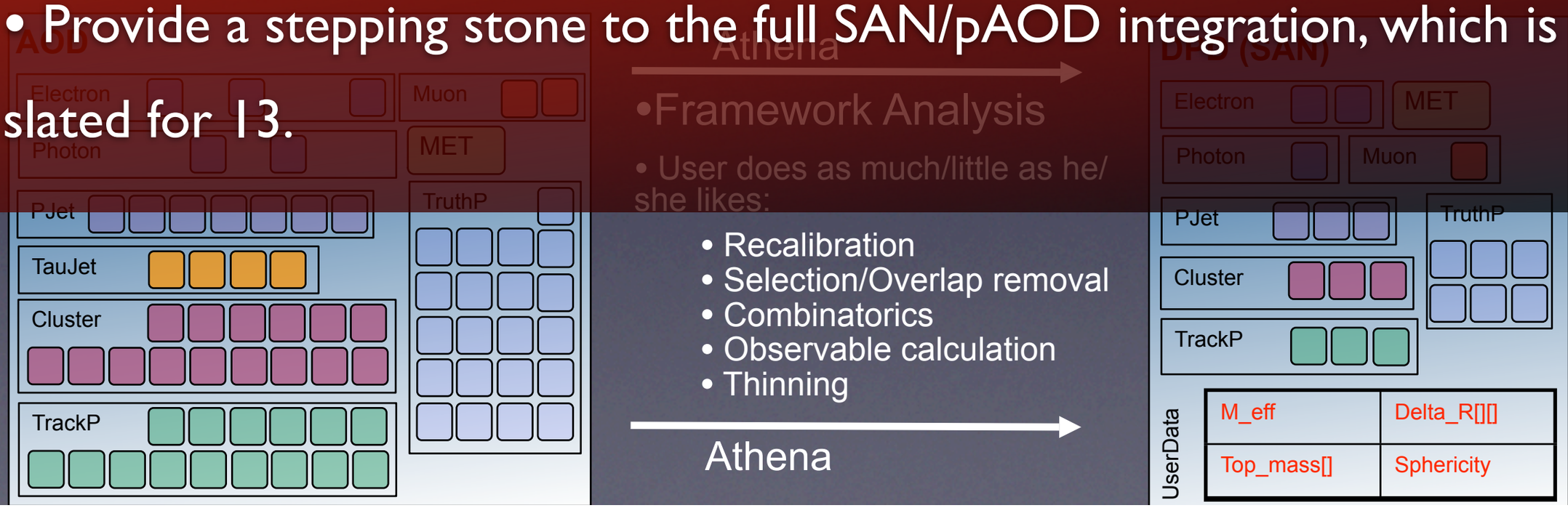
SAN & pAOD

- Until now, ntuples in ATLAS have had no structure (aka *flat*).
- SAN (Structured Athena Ntuple) prototype is a complete copy of AOD classes into format which is directly readable in ROOT (on any platform).
  - Cannot be read back into Athena...
  - Must make a SAN copy of AOD (we cannot keep AOD and SAN copies of analysis data).
- AOD Format Taskforce: make SAN the persistent version of the AOD.
  - AOD looks like SAN when opened directly in ROOT.
- Full AOD functionality preserved.
- No more "ntuples"!

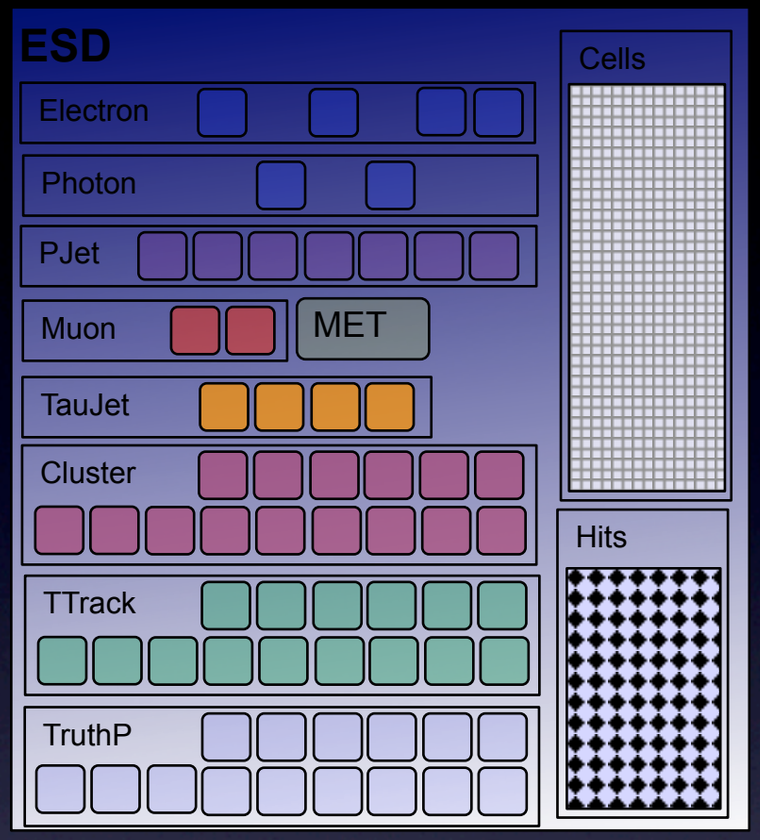


SAN & pAOD

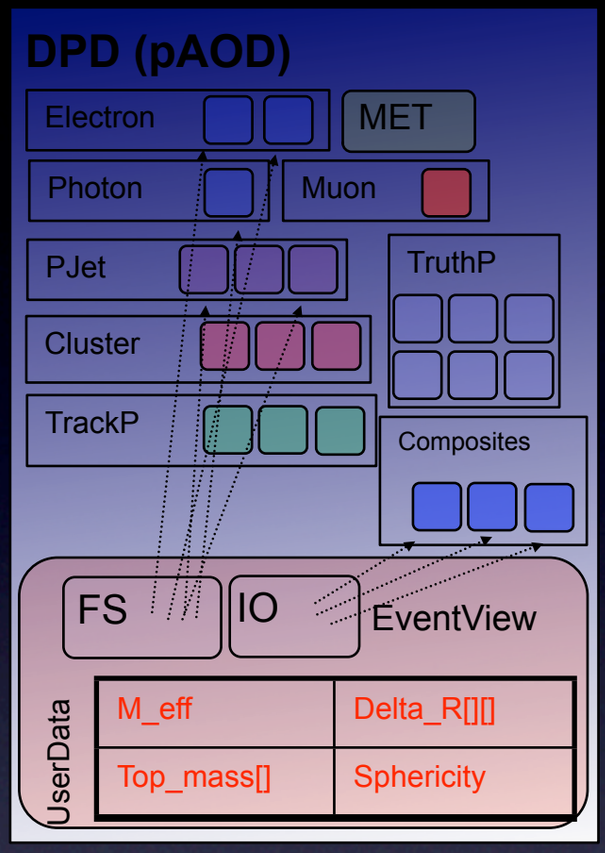
- Until now, ntuples in ATLAS have had no structure (aka *flat*).
- SAN (Structured Athena Ntuple) prototype is a complete copy of AOD classes into format which is directly readable in ROOT (on any platform).
  - Cannot be read back into Athena...
  - Must make a SAN copy of AOD (we cannot keep AOD and SAN copies of analysis data).
  - AOD Format Taskforce: make SAN the persistent version of the AOD.
  - AOD looks like SAN when opened directly in ROOT
  - Full AOD functionality preserved.
  - No more "ntuples"!
- Status:
- SAN prototype in 12.0.6... structured copy of the AOD.
- This 12-series SAN will be made in production & available through dq2.
- Provide a stepping stone to the full SAN/pAOD integration, which is slated for 13.

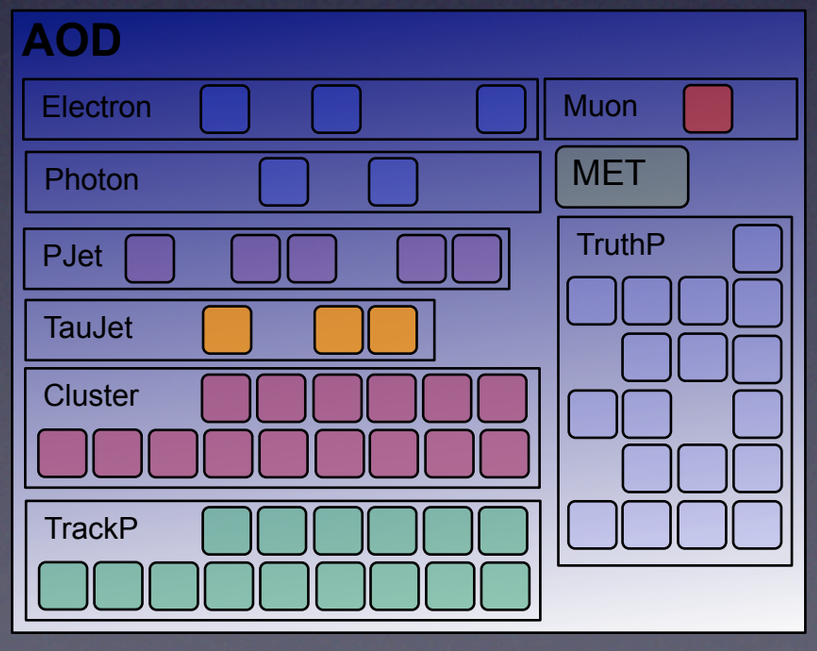


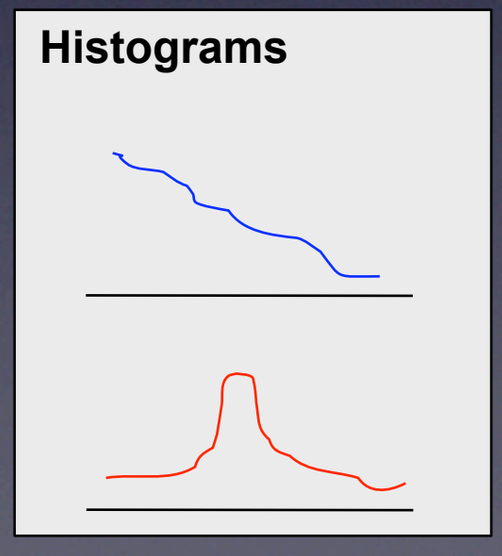
Goal: A Unified EDM



- All EDM levels are fundamentally the same type (ie written through POOL).
- EDM access speed near limit ⇒ less reason to leave Athena.
- Much of the EDM is directly readable in both Athena and ROOT.
- Easy to port code from ROOT to Athena... use the same code for AOD/ESD.
- Same Athena services and tools provide the Skimming, Thinning, Slimming, Adding UserData from ESD⇒AOD⇒DPD⇒DPD⇒...



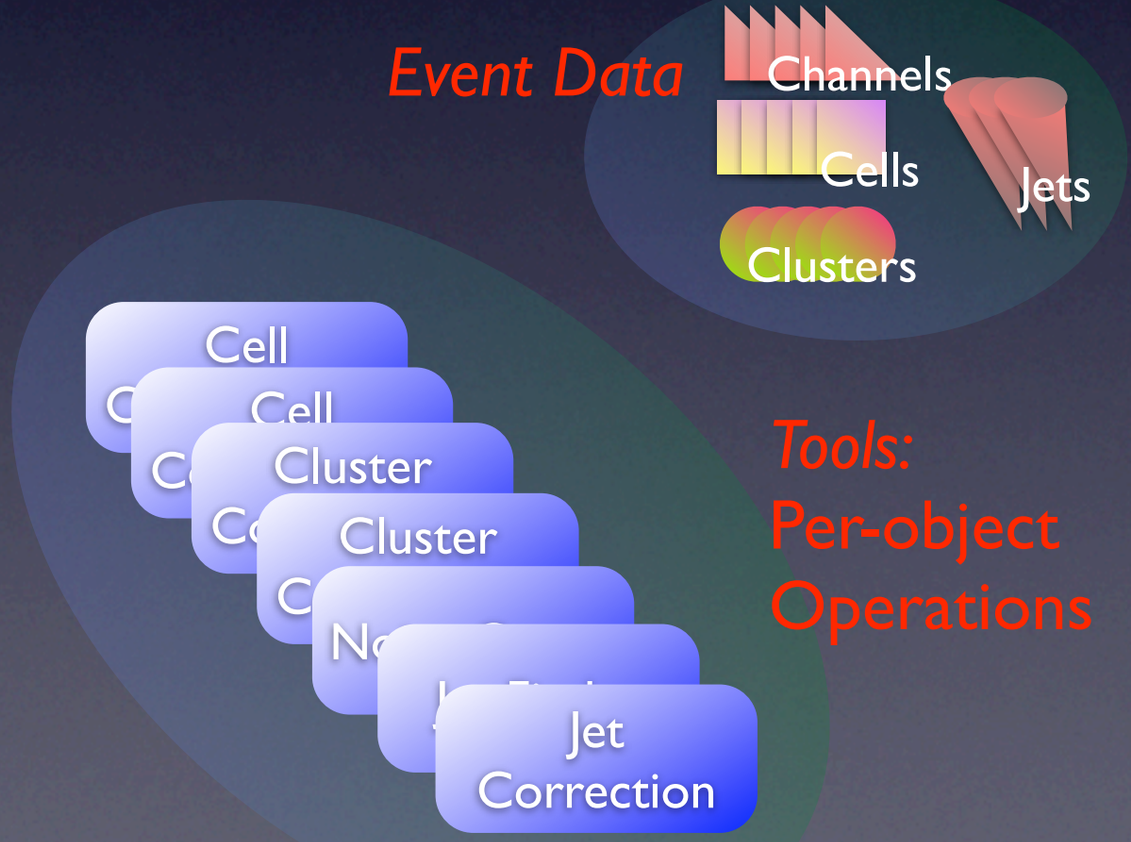




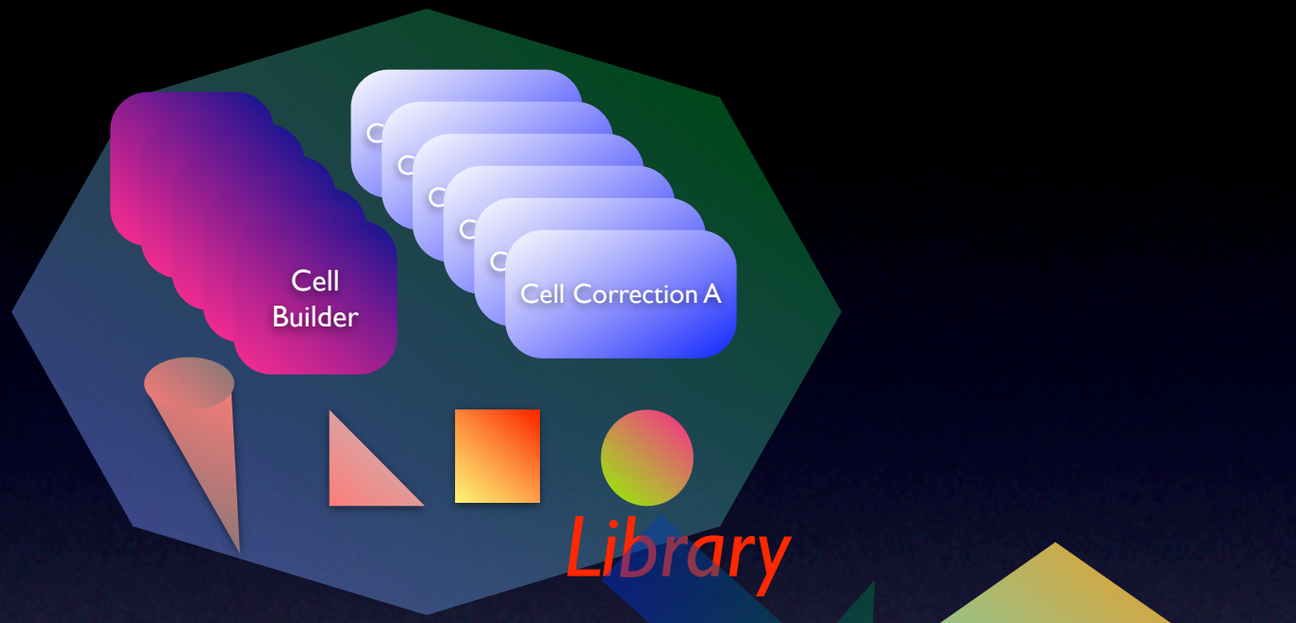
Collaborative Analysis

- Problem: how do you get 2000 physicists to
  - perform analysis in consistent ways
  - easily share & compare their work

- Same problem as reconstruction.
- The reconstruction software is simultaneously developed by 100's of people over many years.
- A common set of framework elements form the basic language of event processing.



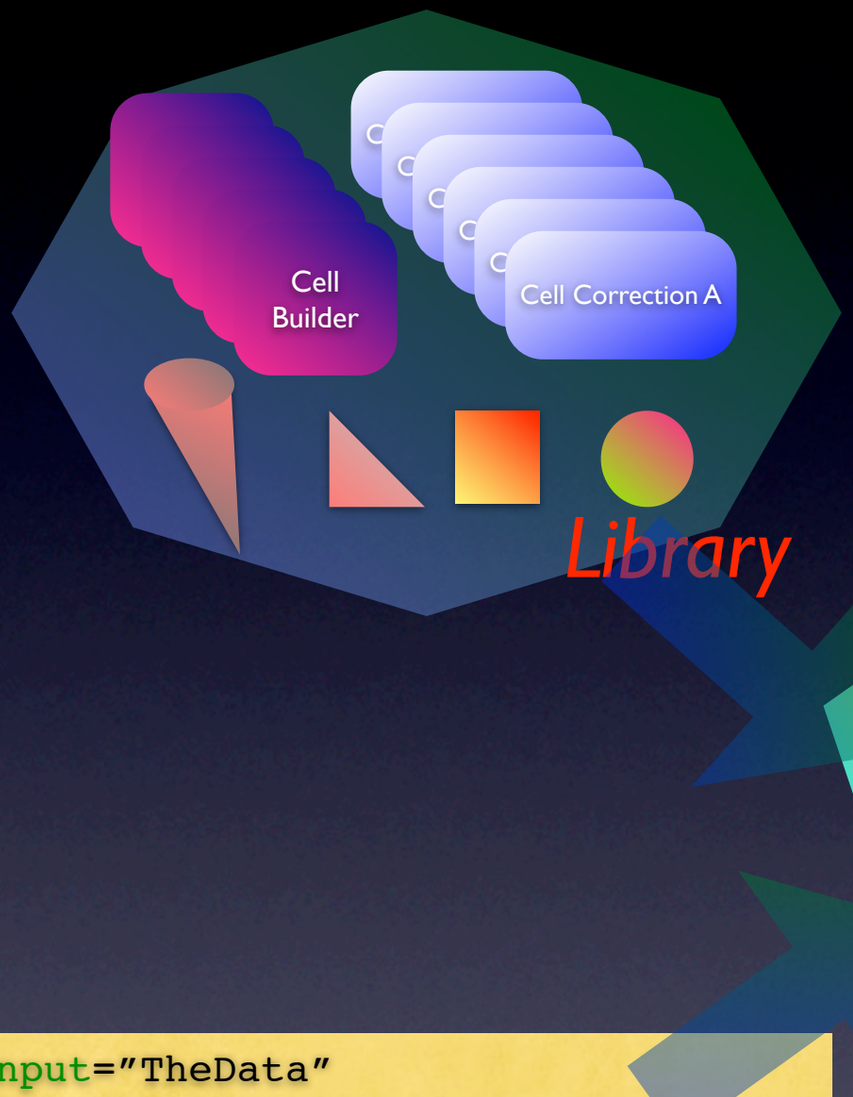
- The reconstruction application is a specific configuration of a library of framework elements.




```
Input=”TheData”
Algorithms+=CellBuilder
(In=“LArgChannels”,Out=”Cells1”)
Algorithms+=CellCalibrator
(In=“Cells1”,Out=”Cells2”)
CellCalibrator+=CellCorrectionA()
CellCalibrator+=CellCorrectionB()
Algorithms+=ClusterBuilder
(In=“Cells2”,Out=”Clusters1”,MinEne
rgy=10*GeV)
....
```

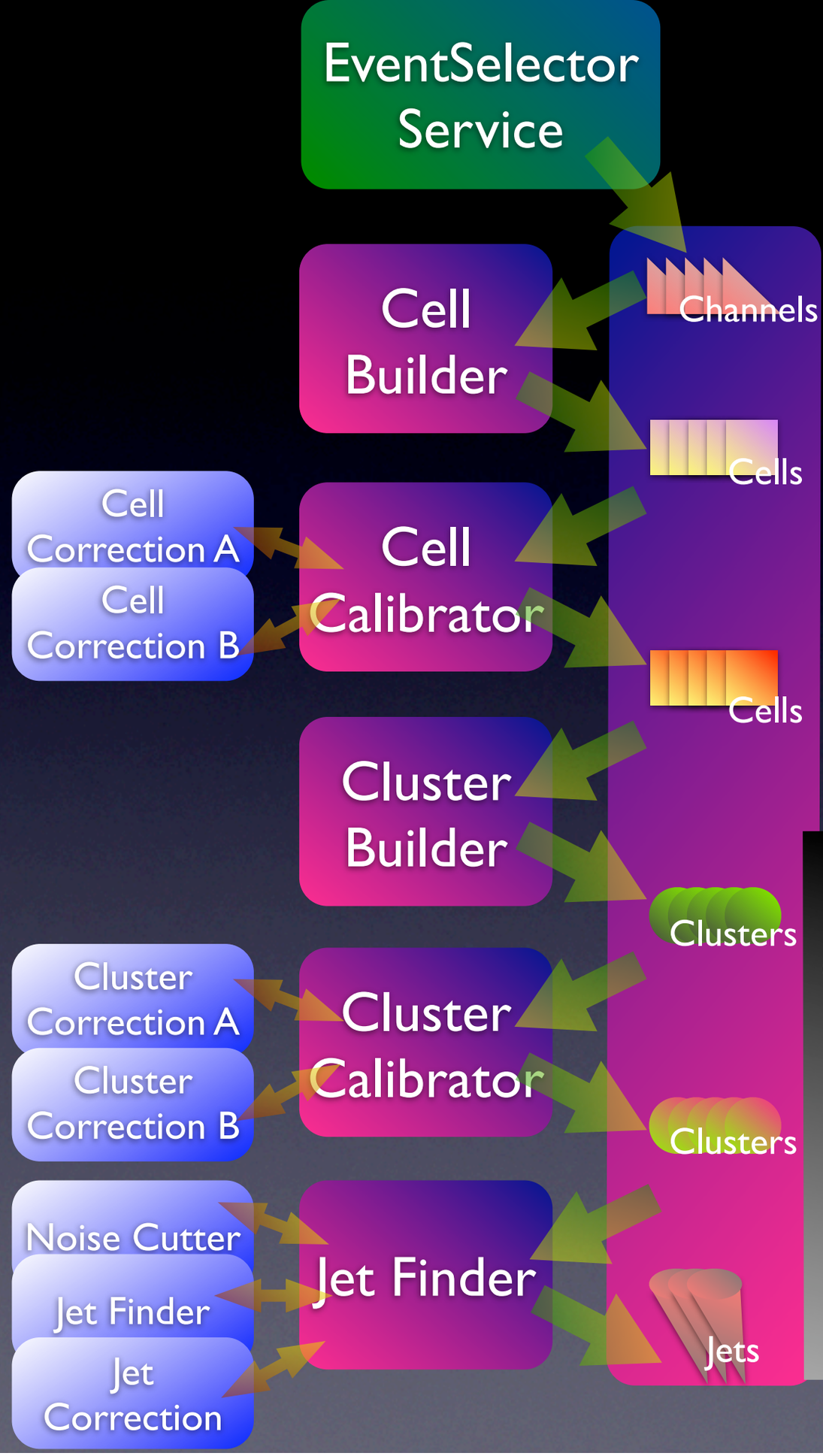
A Configuration

- The reconstruction application is a specific configuration of a library of framework elements.

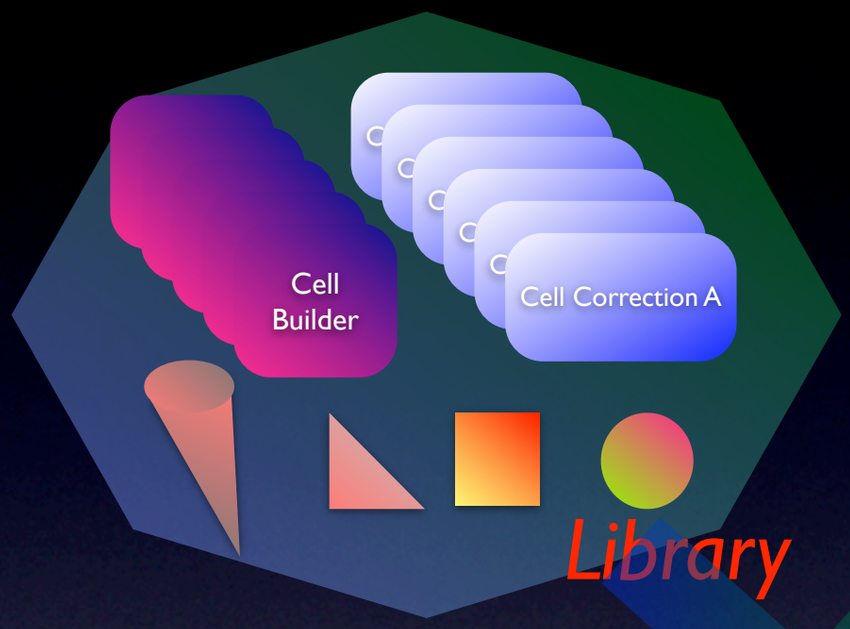



```
Input="TheData"
Algorithms+=CellBuilder
(In="LArgChannels",Out="Cells1")
Algorithms+=CellCalibrator
(In="Cells1",Out="Cells2")
CellCalibrator+=CellCorrectionA()
CellCalibrator+=CellCorrectionB()
Algorithms+=ClusterBuilder
(In="Cells2",Out="Clusters1",MinEne
rgy=10*GeV)
....
```

A Configuration



- The reconstruction application is a specific configuration of a library of framework elements.



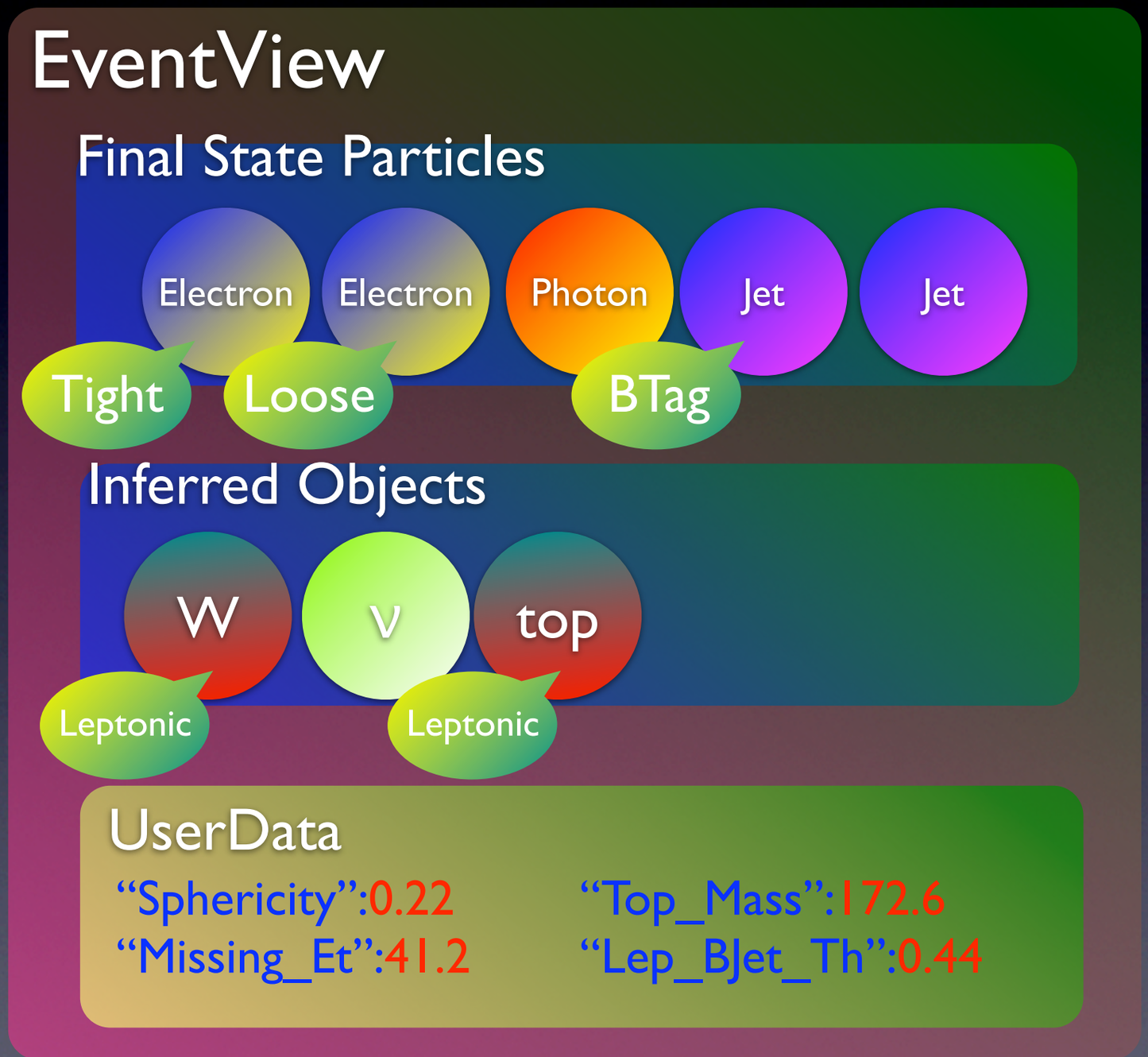
- Apply same concept to analysis: The EventView Framework

```
Input="TheData"
Algorithms+=CellBuilder
(In="LArgChannels",Out="Cells1")
Algorithms+=CellCalibrator
(In="Cells1",Out="Cells2")
CellCalibrator+=CellCorrectionA()
CellCalibrator+=CellCorrectionB()
Algorithms+=ClusterBuilder
(In="Cells2",Out="Clusters1",MinEne
rgy=10*GeV)
....
```

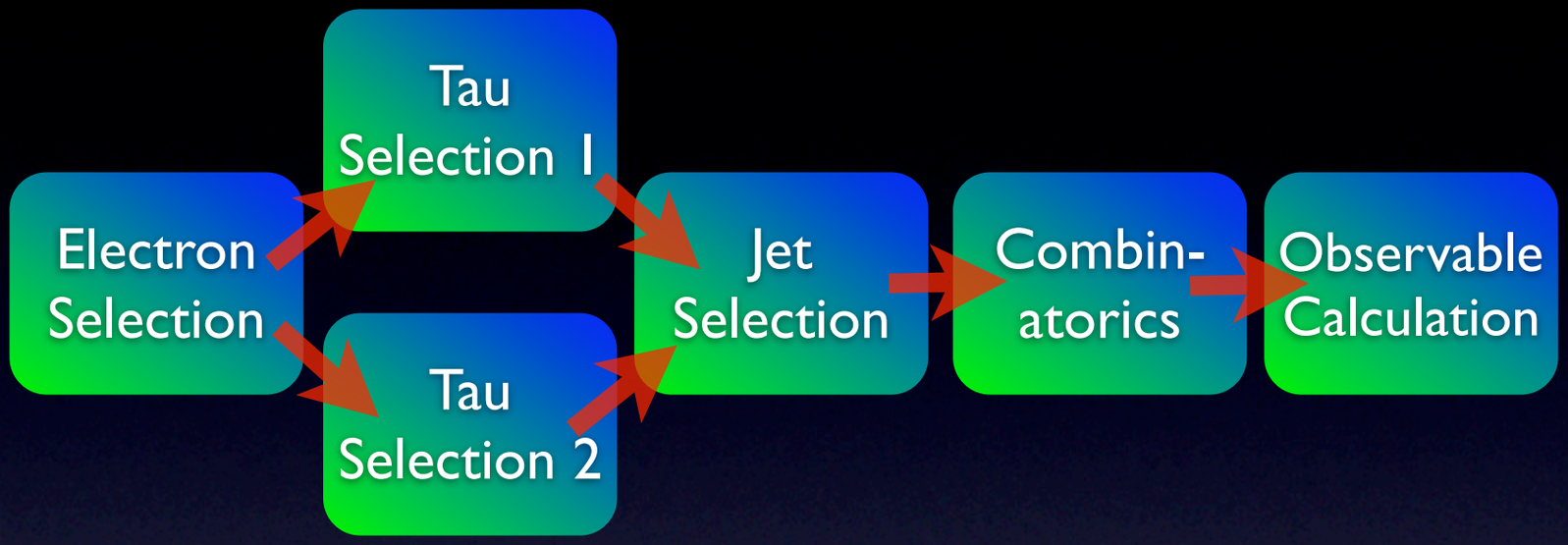
A Configuration

The EventView

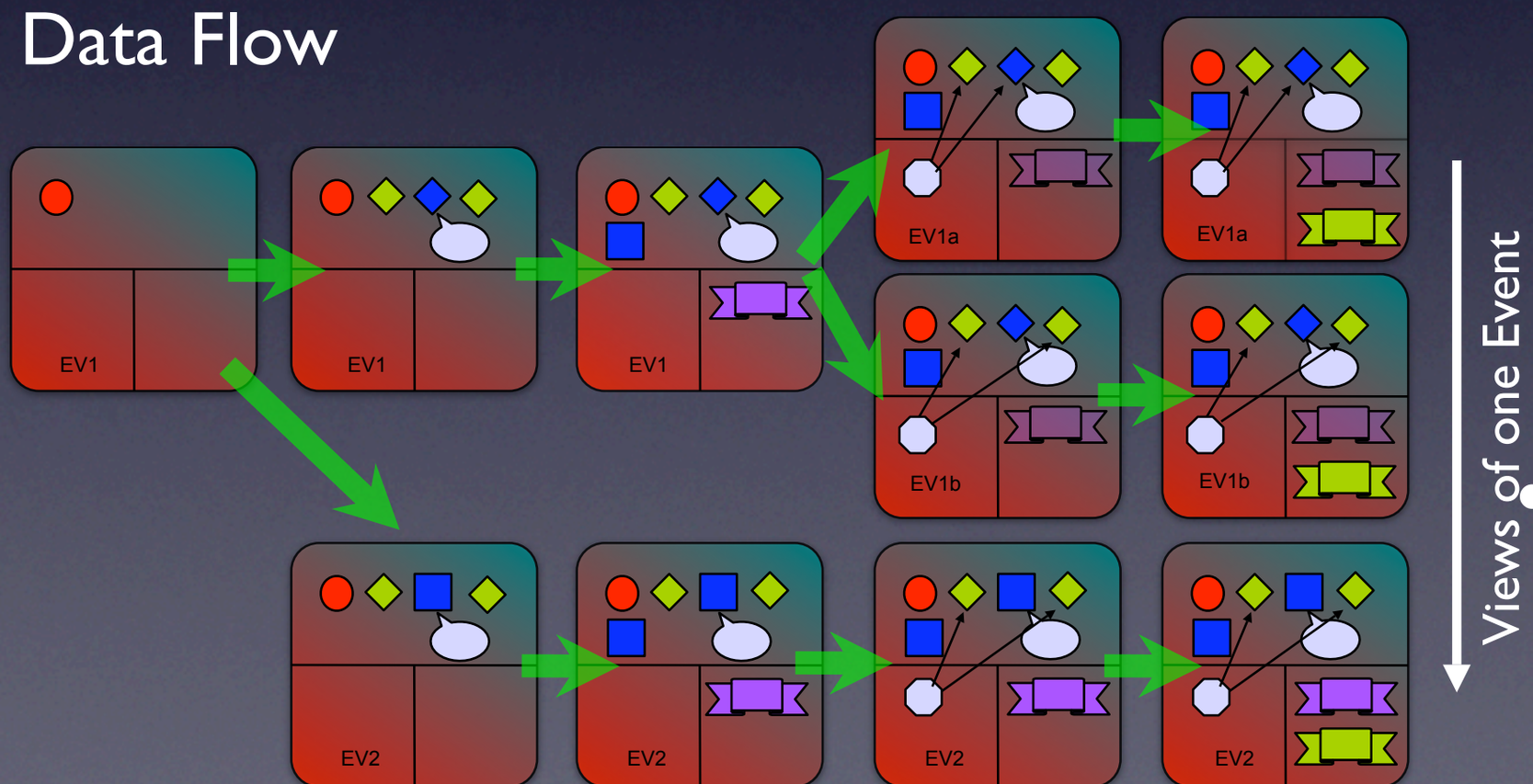
- Holds the "state" of an analysis.
  - Objects in the AOD + Labels.
  - Objects created in the coarse of analysis + Labels.
  - UserData: Anything other data generated during analysis.
- Can be written/read from file and shared (even with a theorist!)
- Convention: each EventView holds *one* interpretation of an event... very natural book keeping tool.



EventView Framework



Analysis Flow

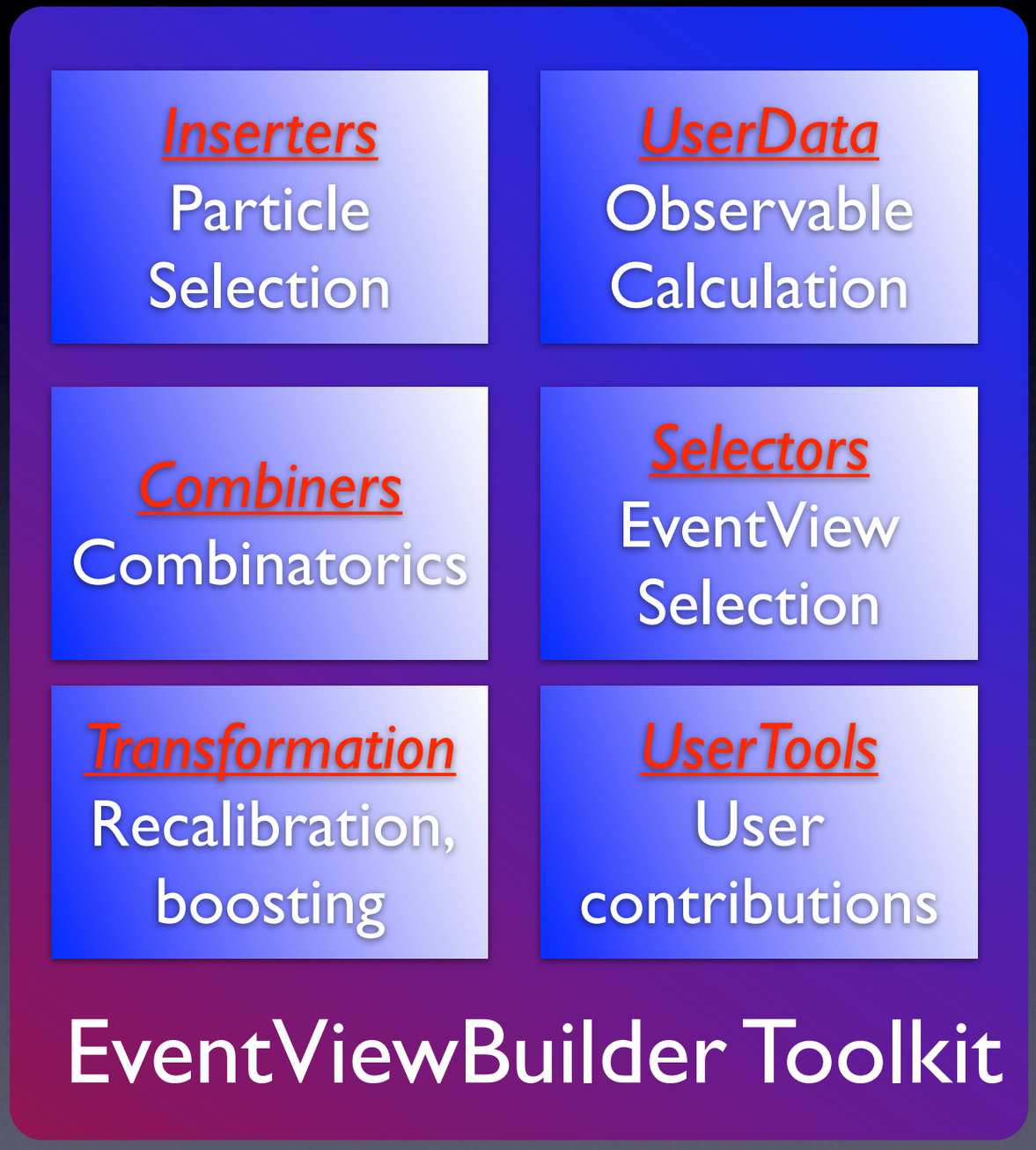


- Analysis is a series of EventView Tools executed in a particular order.
- Framework generates multiple Views of an event representing
  - Different analysis paths
  - Different combinatorics choices
  - Different input (eg: generator, full reconstruction, fast simulation)

Everything consistent within one EventView ⇒ Framework handles bookkeeping.

EventView Toolkit

- 100's of generalized tools which can be configured to perform specific tasks.
- Tools instantiated/configured in python... users can perform complicated analyses w/ o any C++.
- Provide the language for basic analysis concepts: "inserter", "looper", "associator", "calculator", "combiner", "transformer".
- Tools explicitly designed to be extended by users (when necessary).
  - Complicated Athena stuff in base classes.
  - Users only need to implement "the physics".
  - Users now routinely contribute new tools.



"View" Packages

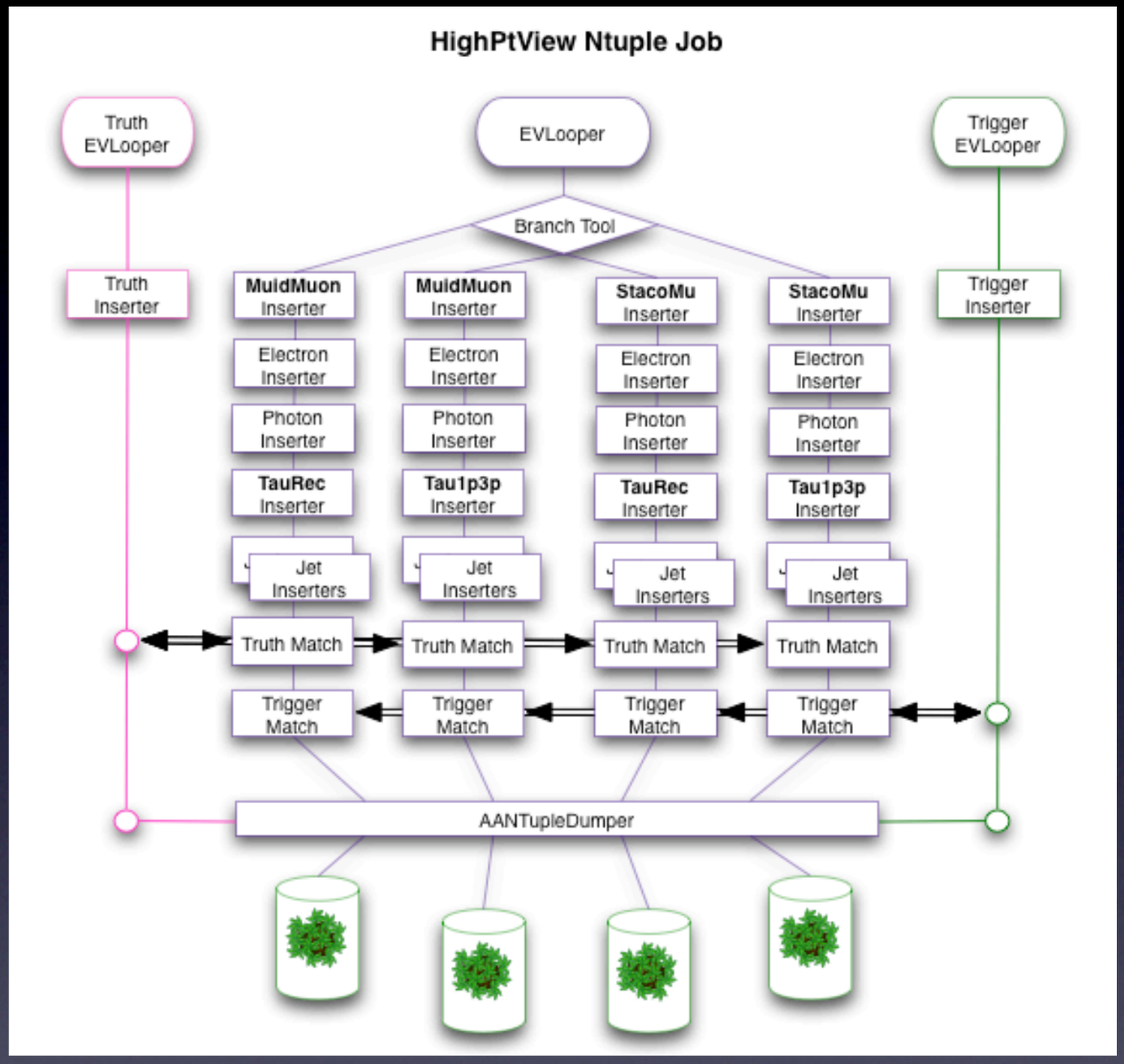
- EventView popularity:
  - In top 11 most visited Atlas web page for past 3 months
  - #9/125 HyperNews forum in # of subscribers
  - #2/125 Hypernews forum in postings...

- Analysis packages are mostly configurations of standard tools... minimal new C++.
- HighPtView: Generic Analysis package running in production ⇒ Standard:
  - Particle selections
  - Truth/Trigger Match
  - Output

  ➡ Serves as benchmark/starting point for analyses
- Many physics groups customizing HighPtView for specific analyses ⇒

  SUSYView, TopView, ...
- Performance packages also coming: egammaView, JetView, MuonView

More on HighPtView



- Full integration of HPTV with 12-series SAN is already in 12.0.6, but not tested yet (help needed).
  - HPTV ntuple can make references to SAN objects instead of making flat ntuple.

- Provides 4 parallel branches of analysis
  - Outputs 4 ntuples
  - Each reflects choice of muon and tau reconstruction algorithms + All jet algorithms.
- Default HPTV will run in Production
  - At Tag + AOD Merge + SAN stage
  - Ntuples available via DQ2
- SAN vs HPTV: SAN is a data format. HPTV is an analysis package. *My* recommendation:
  - 12-series SAN for low-level performance studies.
  - HPTV for high-level analysis (eg Top, SUSY, etc) or assessing global performance (eg compare trigger vs offline).

Customizing HPTV

- HPTV designed as a starting point... users expected to customize for their specific analyses.
- HPTV customization possible at various levels:
  - Easily change HPTV defaults on the command line (or through addition configuration file)
    - Override selections
    - Change Ntuple content
  - Simple mechanism to insert user analysis inside HPTV. This is how the new SUSYView works.
  - Easy to build new packages using components of HPTV. This is how the new TopView will work.
- EventView/HPTV are not meant to be black boxes..

Customizing HPTV

- HPTV designed as a starting point... users expected to customize for their specific analyses.
- HPTV customization possible at various levels:
  - Easily change HPTV defaults on the command line (or through addition configuration file)
  - Override selections
  - Change Ntuple content
  - Simple mechanism to insert user analysis inside HPTV. This is how the new SUSYView works.

- With EventView ATLAS has an analysis framework which
- Makes building complex analyses easier.
- Allows sharing and comparing ideas, code, and results.
- Provides a common language (and tools) across wide array of analyses.
- Is deployed in production, adopted by physics working groups, and widely used by the physics community.

- Easy to build new packages using components of HPTV. This is how the new TopView will work.

- EventView/HPTV are not meant to be black boxes..

Other Analysis Software

- General Multivariate discriminant framework: TMVA.
  - Easily build and compare various discriminants... eg Fisher, Neural Network, boosted decision tree, ...
- Interactive Analysis Frameworks (SPyRoot, sFrame, ...):
  - Simultaneously manipulate multiple datasets ⇒ make plots and perform studies.
  - Share "macros"... collaboratively build analyses.
- General Statistics Framework (for LHC).
  - RooStats... based on RooFit... under development now.
  - Build models of data ⇒ fits, "toy" Monte Carlos, calculate significance... share models/data.
  - Provide standard (and correct) calculation of significance and handling of (systematic) errors.
  - Compare different techniques/calculations.

Final Remarks

- We cannot predict the details of the physics, detector, computing, or software challenges that will confront ATLAS or LHC.
- But the LHC will deliver sufficient 14 TeV data in 2008 to allow discovery of SUSY signatures into the ~1 TeV range... not to mention other interesting physics.
- LHC will also produce lots of Z, W, and tops to help understand the detector and the SM at 14 TeV.... more control samples than at comparable stage at the Tevatron.
- Actually making such measurements early is matter of organization and preparation within the experiments...
- So we are hoping to be as prepared as possible by is feverishly:
  - Installing and commissioning the detector.
  - Building software which anticipates the fundamental issues.
  - Running increasingly realistic Data Challenges to test and explore calibration and analysis strategies.
- A good Analysis Model is vital to our success... so please help us.
- If we don't deal with these issues now, we won't have the chance after we get data.