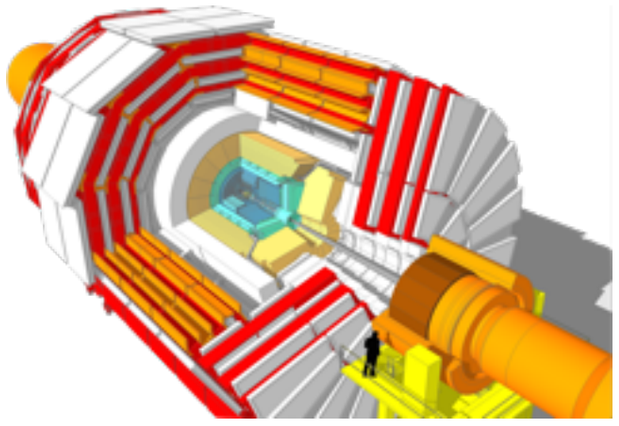
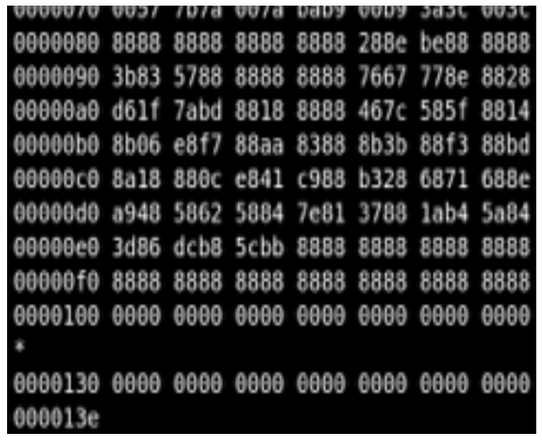
Data Preparation at CMS

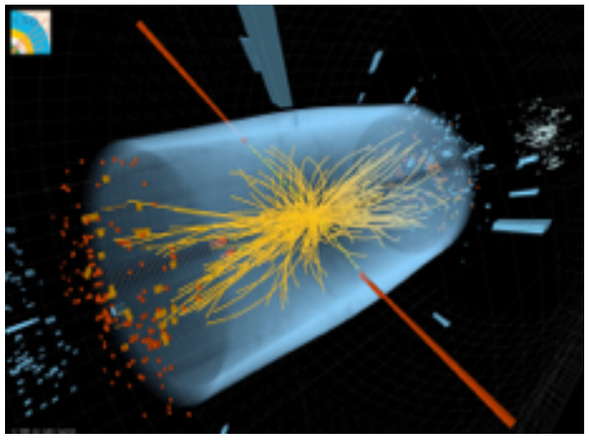
G. Franzoni (CERN)
17 February 2016
CMS Data Analysis School @Taipei

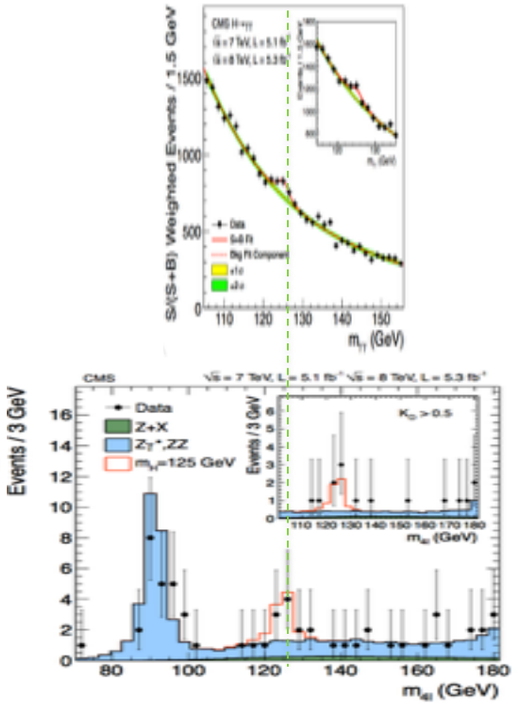
Acknowledgements: G. Cerminara (CERN)
N. Srimanobhas (U Chulalongkorn), M. Musich (U. Louvain)

Data Preparation in a nutshell









Outline

Datasets

- how they're made
- how you can find/use them

Validation

Data Quality & Certification

Alignment and Calibration

note: clickable references to documentation/contacts and exercises

Data Preparation & Coord. Areas

This talk focuses on the work of the CMS coord. area:

- **Physics Performance and Datasets** (PPD)
  - data quality & certification
  - alignment & calibrations
  - software validation
  - management of Monte Carlo requests
  - organization and configuration of datasets and data processing
- It also touches on many activities of the:

Offline & Computing

  - CMSSW software development, event reconstruction and simulation
  - data processing and Simulated events generation, events storage and management

Contact

Data Preparation & Coord. Areas

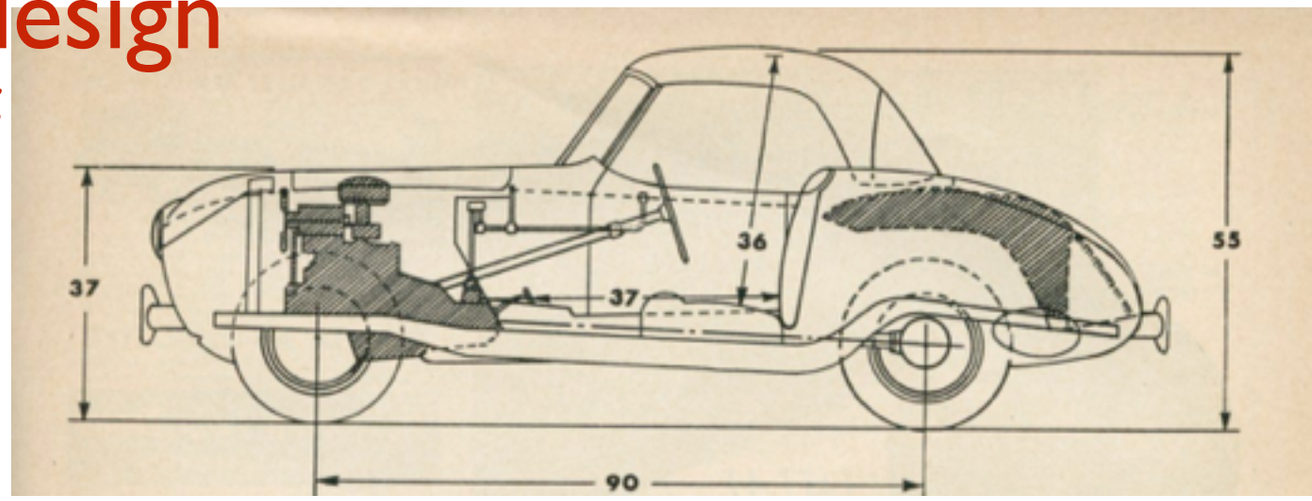
This talk focuses on the work of the CMS coord. area:

- **Physics Performance and Datasets** (PPD)
  - data quality & certification
  - alignment & calibrations
  - software validation
  - management of Monte Carlo requests
  - organization and configuration of datasets and data processing
- It also touches on activities of the:

Offline & Computing

  - CMSSW software development, event reconstruction and simulation
  - data processing and Simulated events generation, events storage and management

design



Contact

Data Preparation & Coord. Areas

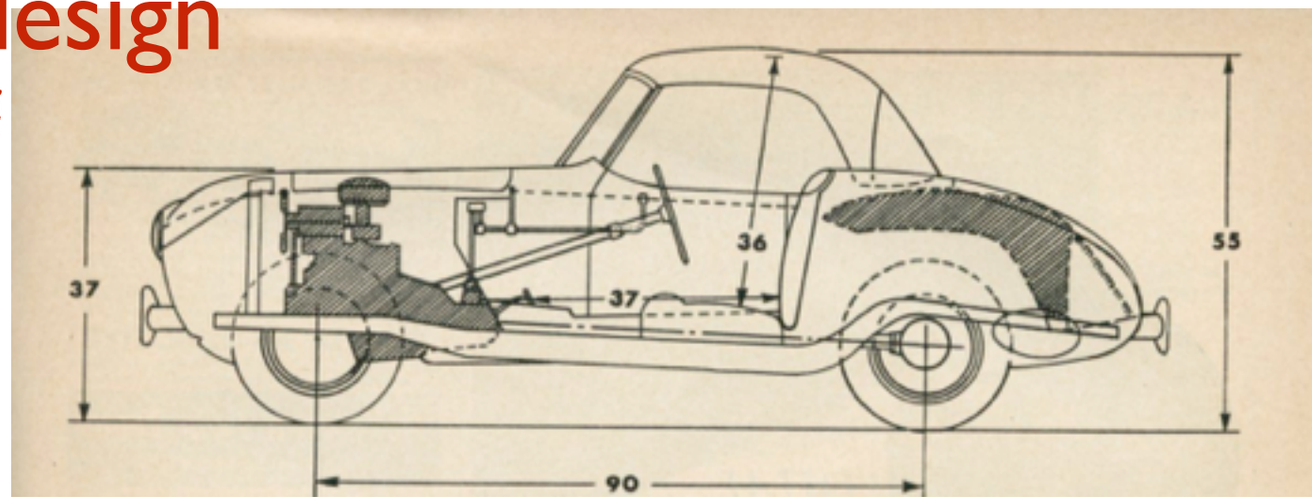
This talk focuses on the work of the CMS coord. area:

- **Physics Performance and Datasets** (PPD)
  - data quality & certification
  - alignment & calibrations
  - software validation
  - management of Monte Carlo requests
  - organization and configuration of datasets and data processing
- It also touches on activities of the:

Offline & Computing

  - CMSSW software development, event reconstruction and simulation
  - data processing and Simulated events generation, events storage and management

design



product industrialization

Contact

Data Preparation & Coord. Areas

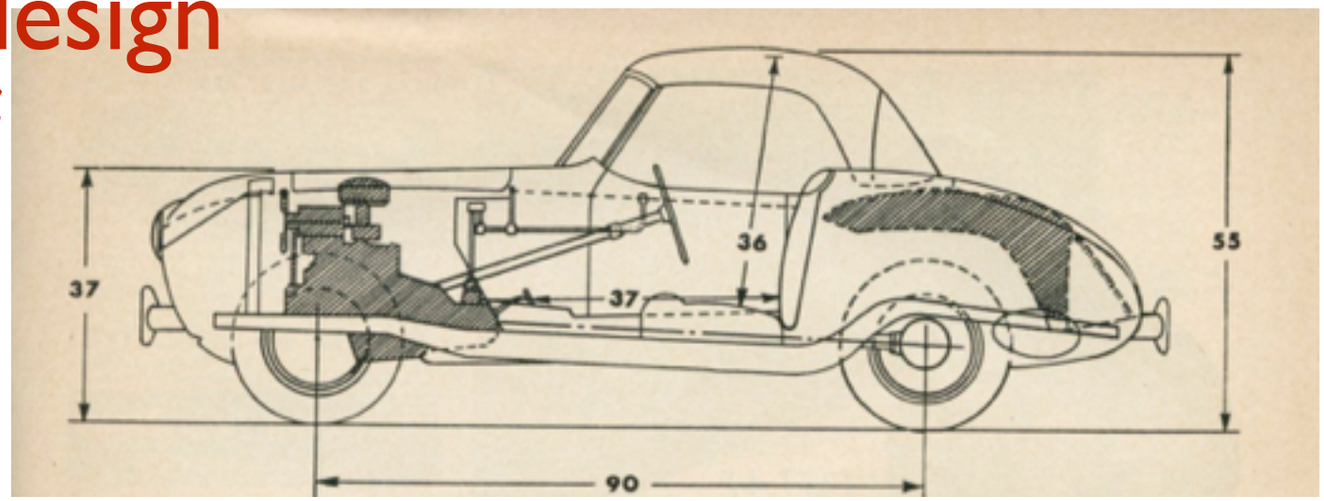
This talk focuses on the work of the CMS coord. area:

- **Physics Performance and Datasets** (PPD)
  - data quality & certification
  - alignment & calibrations
  - software validation
  - management of Monte Carlo requests
  - organization and configuration of datasets and data processing Contact&Doc
- It also touches on activities of the:

Offline & Computing

  - CMSSW software development, event reconstruction and simulation
  - data processing and Simulated events generation, events storage and management

design

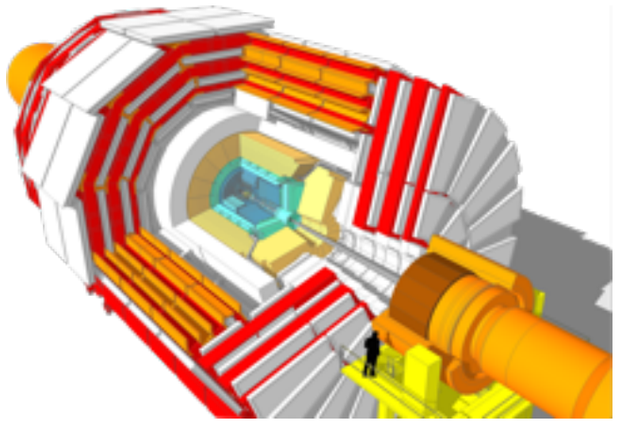


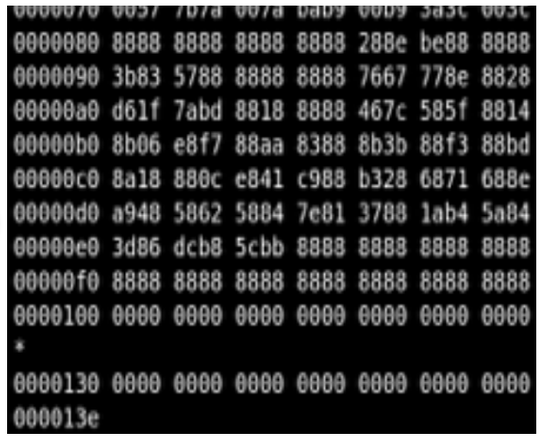
product industrialization

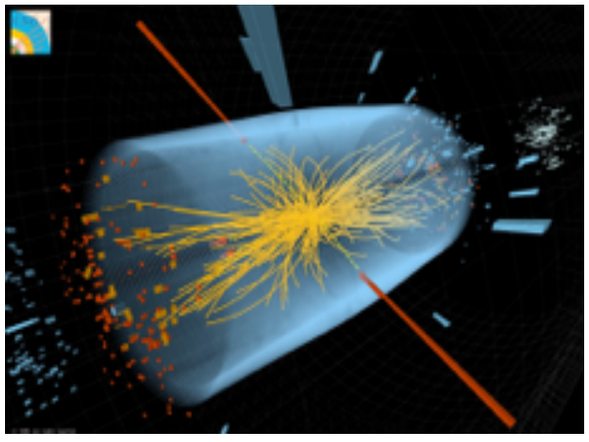
manufacturing

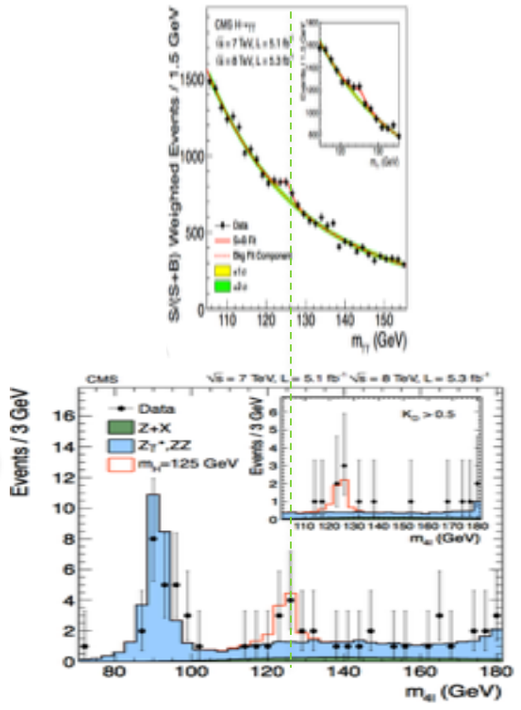
Contact

Data Preparation in a nutshell









Datasets

- how they're made
- how you can find/use them

From P5 to Offline

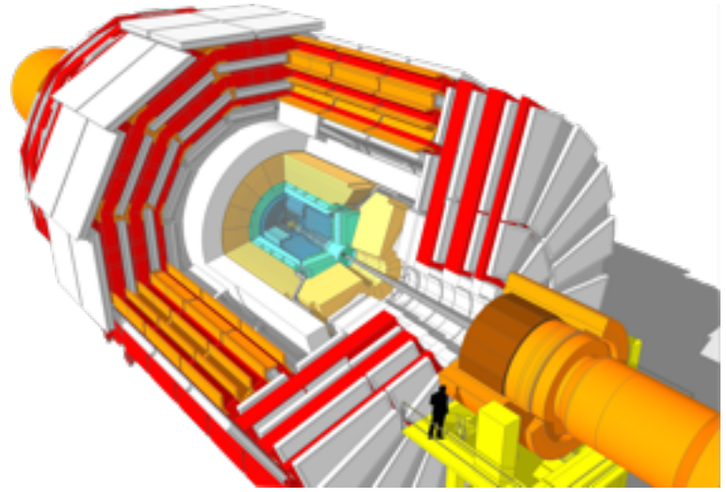
The events collected by CMS reach the Tier0 farm at CERN for tape archival, organization and processing

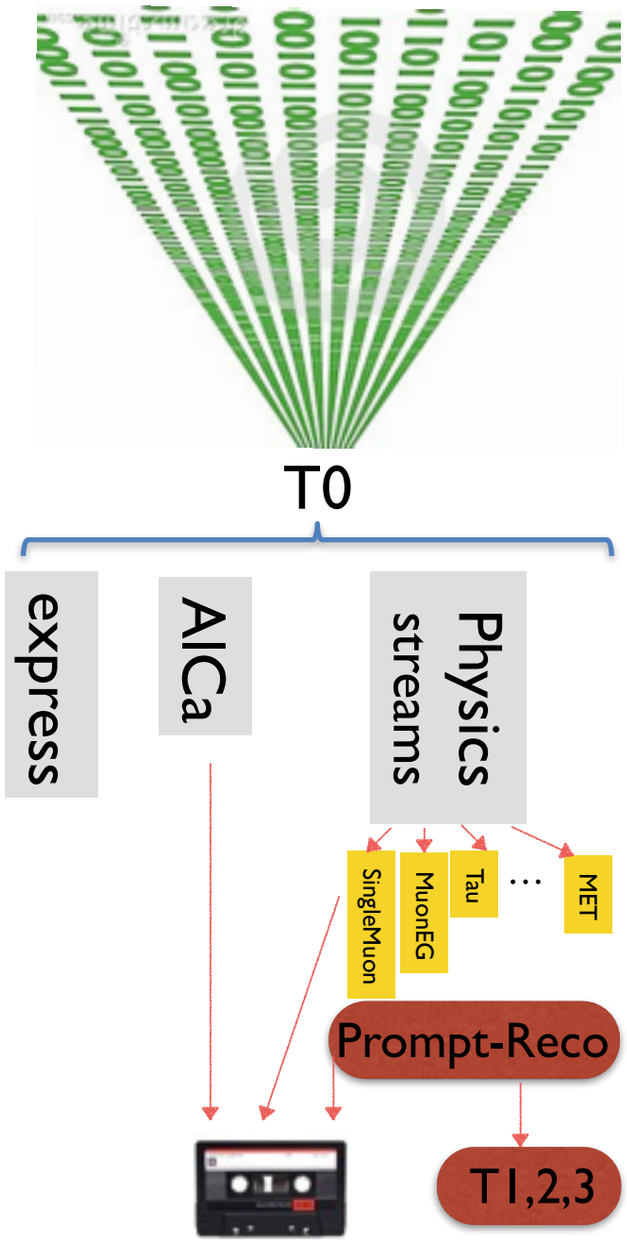
Different **streams** and workflows serve multiple **use-cases**

- **express**: available ~2h after data collection for prompt feedback & calibrations ; ~40Hz bandwidth shared by:
  - calibration (1⁄2) · detector (1⁄4) · physics (1⁄4) monitoring
- **Alignment&Calibration** (AlCa) streams
  - dedicated event selection & event content devised for calibration purposes
- **Physics Streams**: split into primary datasets and promptly reconstructed for physics analysis
  - delayed of 48h → allow the Prompt Calibration Loop
- other specialized streams (e.g “data parking”, “data scouting”,”hotline”)

RunI : 300Hz Prompt-Reco + 300-600Hz of parked data
= reconstructed only after the end of data taking

RunII : 1kHz of Prompt-Reco + high rate of scouting streams with reduced event content





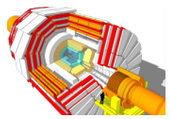
Prompt Reco and PCL at T0

- CMS aims at providing the analysts with reconstructed data within days of data collection!
- Data handling and measurement of calibrations must be robust, reliable and as automated as possible

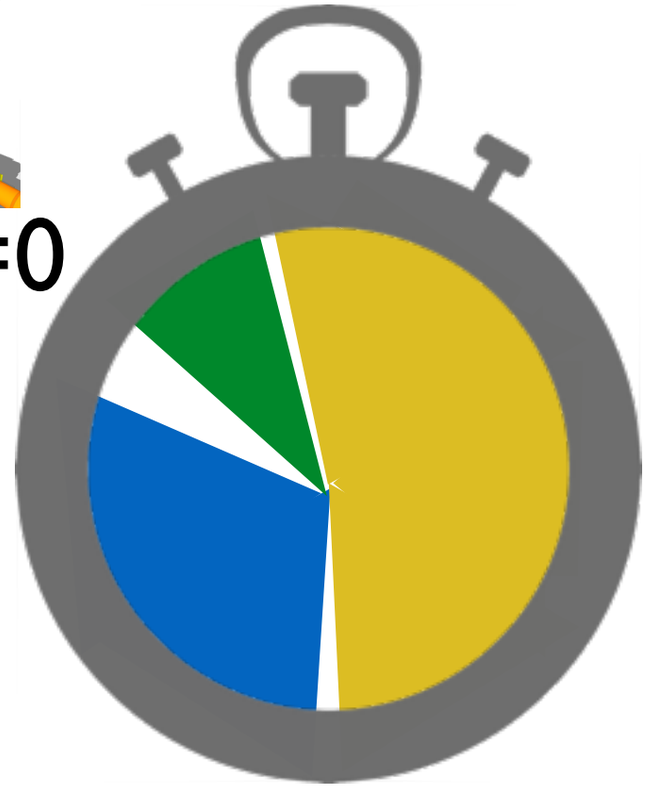
Express processing

Data reconstructed for:

- monitoring
- calibration



t=0



Prompt Reconstruction t=48h

Physics Streams events are reconstructed consuming calibrations computed by PCL
These are the datasets for analysis

Prompt Calibration Loop (PCL)

Express data are used as input to automated calibration workflows running @ Tier0 (or online)

- beam-spot LS by LS
- ECAL transparency corr.
- SiStrip bad-channels
- SiStrip gains
- SiPixel alignment

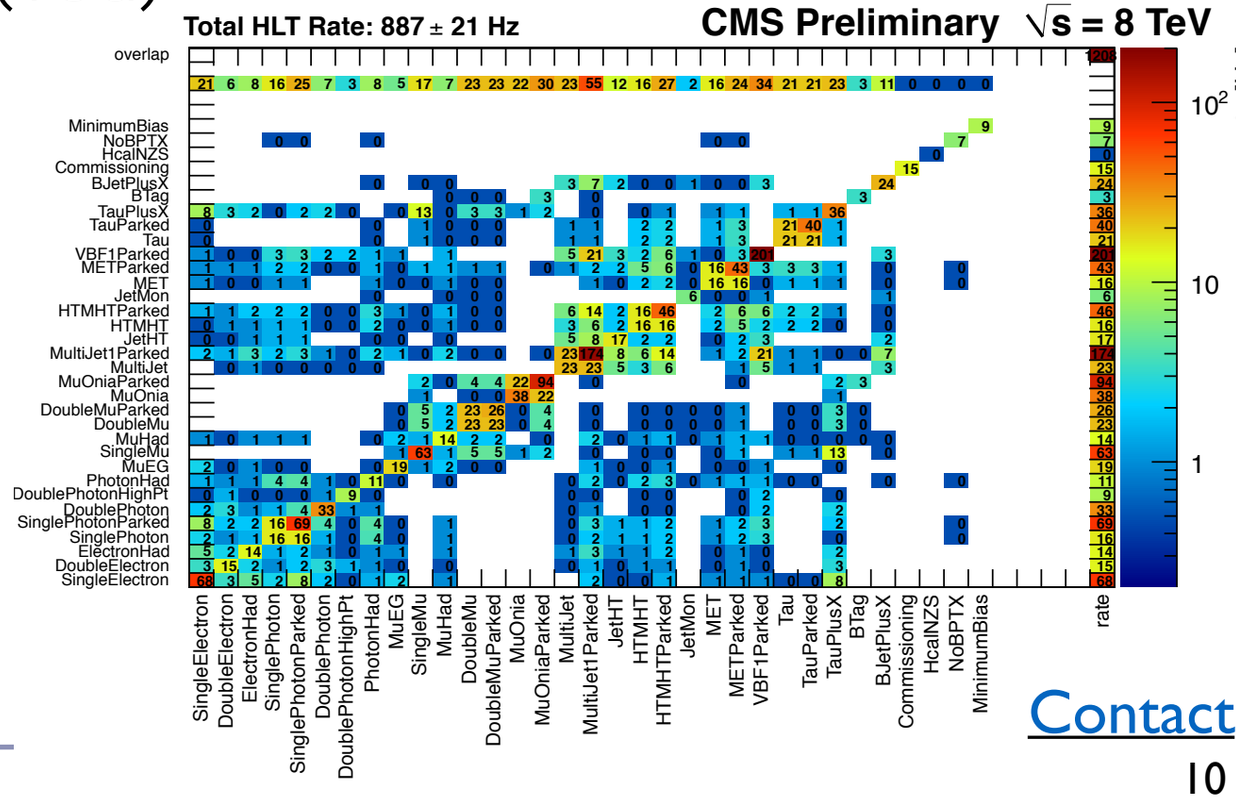
Primary Datasets

The **Physics Streams** from P5 are split in Primary Datasets (PD) on the basis of the HLT results in order to: **group events with related topology** in the same PD to ease consumption & **limit replication** of events (PD's overlap)

HLT conf

- Constraints from analysis:
  - definition centered on physics objects (e.g. SingleMuon, MuonEG, MET…)
- Constraints from processing and handling:
  - average event rate approximately uniform across different PDs, to ease distribution at the Tier2 centers
  - event rate > 10 Hz, to avoid small files & < 200 Hz

Dataset & skim definition, managed by the Dataset Definition Team (PdmV/DDT) in PPD together w/ Trigger Study Group (TSG)



Contact

Primary Datasets

The **Stream A** from P5 is split in Primary Datasets (PD) on the basis of the HLT results: **group events with related topology** in the same PD to ease consumption & **limit replication** of events (PD's overlap)

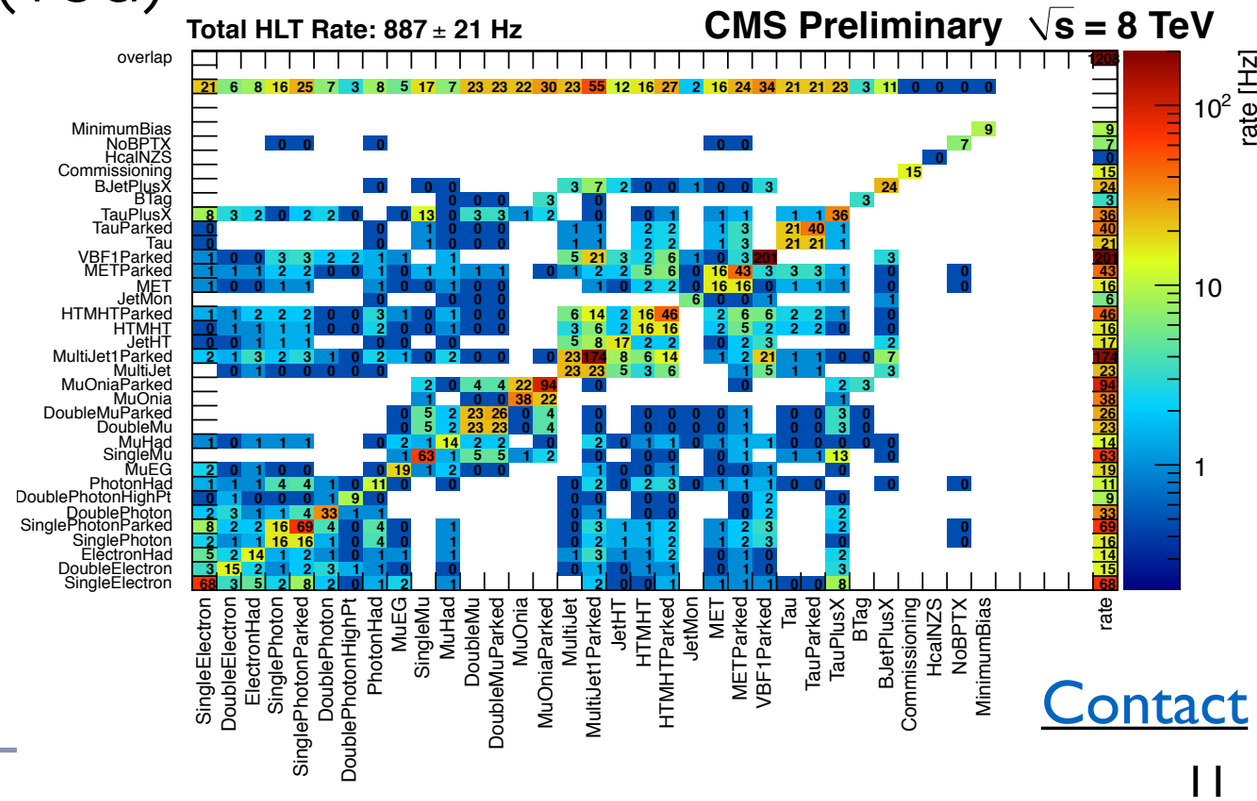
- Constraints from analysis:
  - definition centered on physics objects (e.g. SingleMuon, MuonEG, MET…)
- Constraints from processing and handling:
  - average event rate approximately uniform across different PDs, to ease distribution at the Tier2 centers
  - event rate > 10 Hz, to avoid small files & < 200 Hz

Dataset & skim definition, managed by the Dataset Definition Team (PdmV/DDT) in PPD together w/ Trigger Study Group (TSG)

On the top of the primary datasets we can deploy "central skims" ⟶ customised event content + rate reduction using also RECO quantities

- used for Detector Studies (DPG) or Physics Analysis Groups (POG-PAG)

Dataset & skim definition, managed by the Dataset Definition Team (PdmV/DDT) in PPD together w/ Trigger Study Group (TSG)



Contact

Finding Datasets: anatomy

How do I look for a sample → Data Aggregation Service (DAS)

- list datasets and their properties (requestID, sites, run # and LS #....) aggregating information from various services

Anatomy of the dataset name:

- dataset = /PrimaryDataset/ProcessingVersion/DataTier

Examples:

- data (prompt reco): /SingleElectron/Run2015D-PromptReco-v3/AOD
- data (re-reco): /SingleElectron/Run2015D-16Dec2015-v1/MINIAOD
- MC (RunIIFall15DR76): /WprimeToMuNu_M-1600_TuneCUETP8M1_13TeV-pythia8/RunIISpring15MiniAODv2-74X_mcRun2_asymptotic_v2-v1/MINIAODSIM

Contact

Finding Datasets: anatomy

How do I look for a sample → Data Aggregation Service (DAS)

- list datasets and their properties (requestID, sites, run # and LS #....) aggregating information from various services

Anatomy of the dataset name:

- dataset = /PrimaryDataset/ProcessingVersion/DataTier

Examples:

- data (prompt reco): /SingleElectron/Run2015D-PromptReco-v3/AOD
- data (re-reco): /SingleElectron/Run2015D-16Dec2015-v1/MINIAOD
- MC (RunIIFall15DR76): /WprimeToMuNu_M-1600_TuneCUETP8M1_13TeV-pythia8/RunIISpring15MiniAODv2-74X_mcRun2_asymptotic_v2-v1/MINIAODSIM

Event Topology (data)
and physics process simulated (MC)
is indicated in the first segment of the dataset name

Contact&Doc

Finding Datasets: anatomy

How do I look for a sample → Data Aggregation Service (DAS)

- list datasets and their properties (requestID, sites, run # and LS #....) aggregating information from various services

Anatomy of the dataset name:

- dataset = /PrimaryDataset/ProcessingVersion/DataTier

Examples:

- data (prompt reco): /SingleElectron/Run2015D-PromptReco-v3/AOD
- data (re-reco): /SingleElectron/Run2015D-16Dec2015-v1/MINIAOD
- MC (RunIIFall15DR76): /WprimeToMuNu_M-1600_TuneCUETP8M1_13TeV-pythia8/RunIISpring15MiniAODv2-74X_mcRun2_asymptotic_v2-v1/MINIAODSIM

```
Acquisition Era + PromptReco/reprocessing (data)
Production campaign (MC)
Alignment and Calibration i.e. Global tag (MC)
Dataset Version
```

Contact&Doc

Finding Datasets: anatomy

How do I look for a sample → Data Aggregation Service (DAS)

- list datasets and their properties (requestID, sites, run # and LS #....) aggregating information from various services

Anatomy of the dataset name:

- dataset = /PrimaryDataset/ProcessingVersion/DataTier

Examples:

- data (prompt reco): /SingleElectron/Run2015D-PromptReco-v3/AOD
- data (re-reco): /SingleElectron/Run2015D-16Dec2015-v1/MINIAOD
- MC (RunIIFall15DR76): /WprimeToMuNu_M-1600_TuneCUETP8M1_13TeV-pythia8/RunIISpring15MiniAODv2-74X_mcRun2_asymptotic_v2-v1/MINIAODSIM

Data tier indicates the collections available at each event

Contact&Doc

ex I-2

Finding Datasets: campaigns

The **PdmV** (Physics Data Monte Carlo Validation, PPD L2) team organises of data reprocessing and simulated events production in "**Campaigns**"

- data reprocessing campaigns Doc
- Monte Carlo campaigns Doc

Datasets within a campaign share:

- CMSSW release
- version of Calibration and alignment
- for MC: centre of mass energy, beamspot, CMS era, pile up scenario

Monte Carlo campaigns are devised to match a certain version of the processed data. Chose the campaigns your datasets carefully and consistently

Organising MC Production

MC production matches data processing and targets specific conferences/events

- usual scale for 1 year of data to be analyzed is ~4 Gevents

Production organization is a complex task (PdmV L2 team in PPD):

- O(1000) requests from analyzers with different physics goals
- prioritization to deal with resource budget & conference timeline

Activity coordinated by PPD/PdmV with Request Managers, GEN conveners & Computing Operations

- each PAG has MC contacts taking care of requesting sample needed by the group (after consultation with conveners)
- requests are discussed in weekly Monte Carlo Coordination Meetings (MccM) on Wednesday 15-16
- tool for bookkeeping & monitoring of campaigns and single requests ⟶ McM

McM WebService

MccM Meeting
⟶ priority
⟶ monitoring
⟶ GENerators
⟶ McM

ex2

Computing "CompOps" team

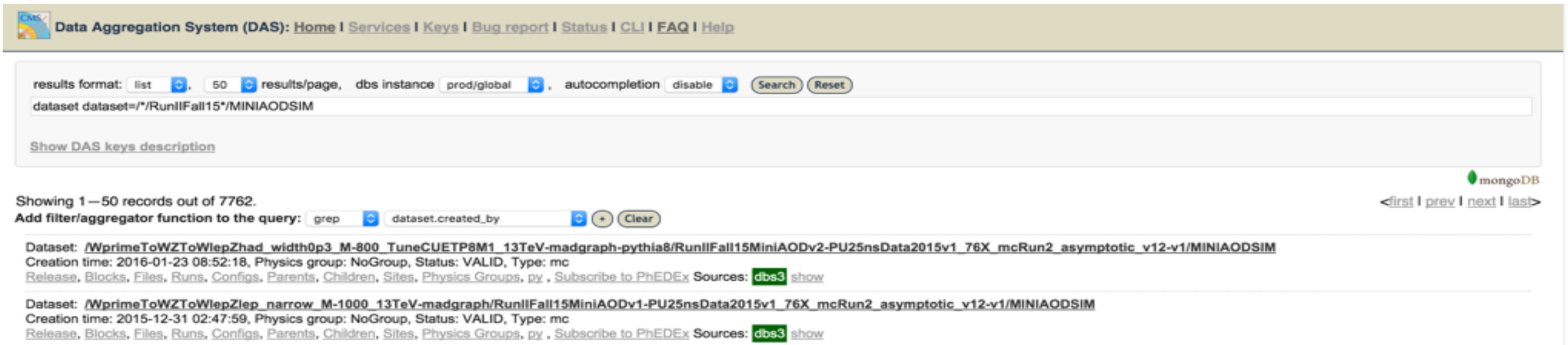
Contact

ex I-2

Finding Datasets

How do I look for a sample → Data Aggregation Service (DAS)

- list datasets and their properties (requestID, sites, run # and LS #) fetching information from various services

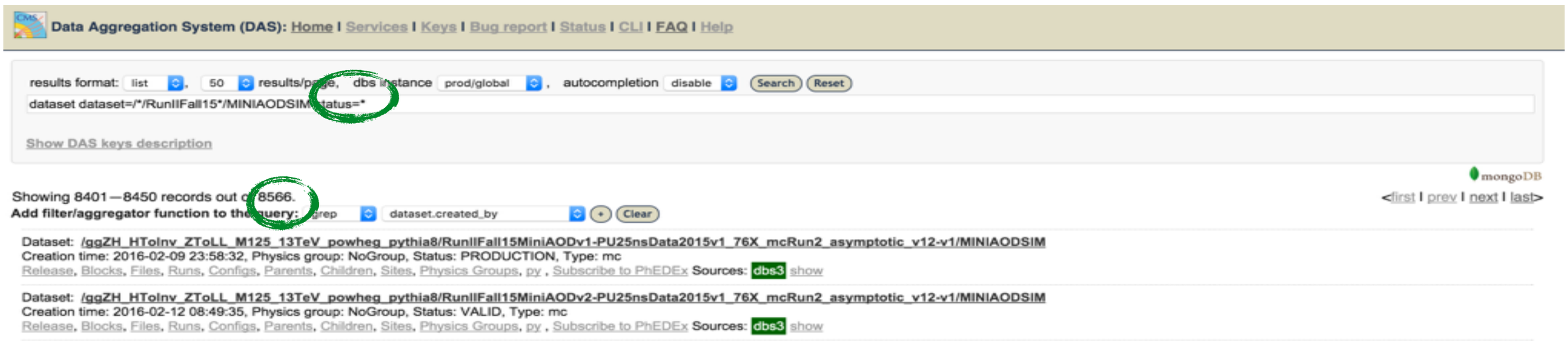


Finding Datasets while being produced

ex I-2

How do I look for a sample ⟶ Data Aggregation Service (DAS)

- list datasets and their properties (requestID, sites, run # and LS #) fetching information from various services
- If your sample has been recently submitted to production, and is not yet ready —> include the status "PRODUCTION" in your query

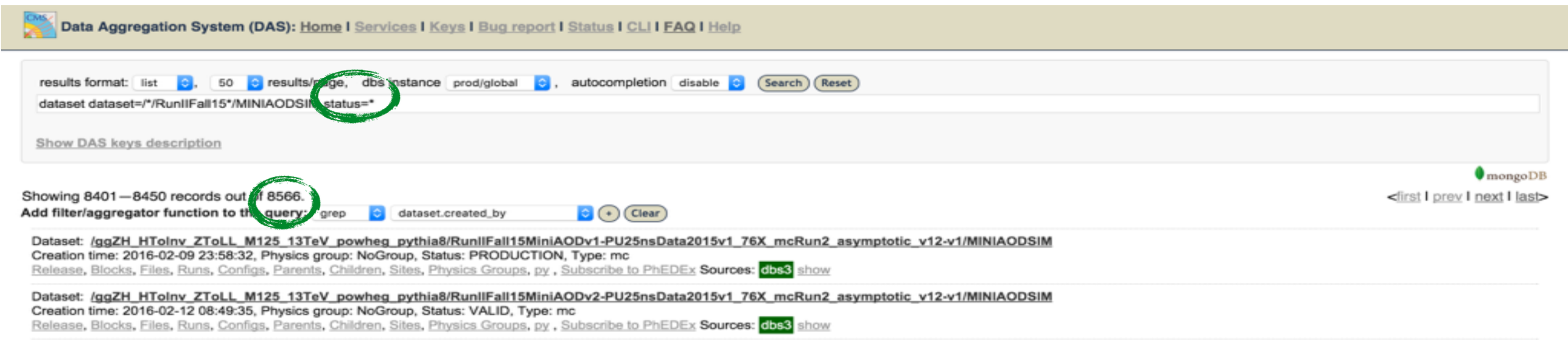


Finding Datasets while being produced

ex I -2

How do I look for a sample ⟶ Data Aggregation Service (DAS)

- list datasets and their properties (requestID, sites, run # and LS #) fetching information from various services
- If your sample has been recently submitted to production, and is not yet ready —> include the status "PRODUCTION" in your query



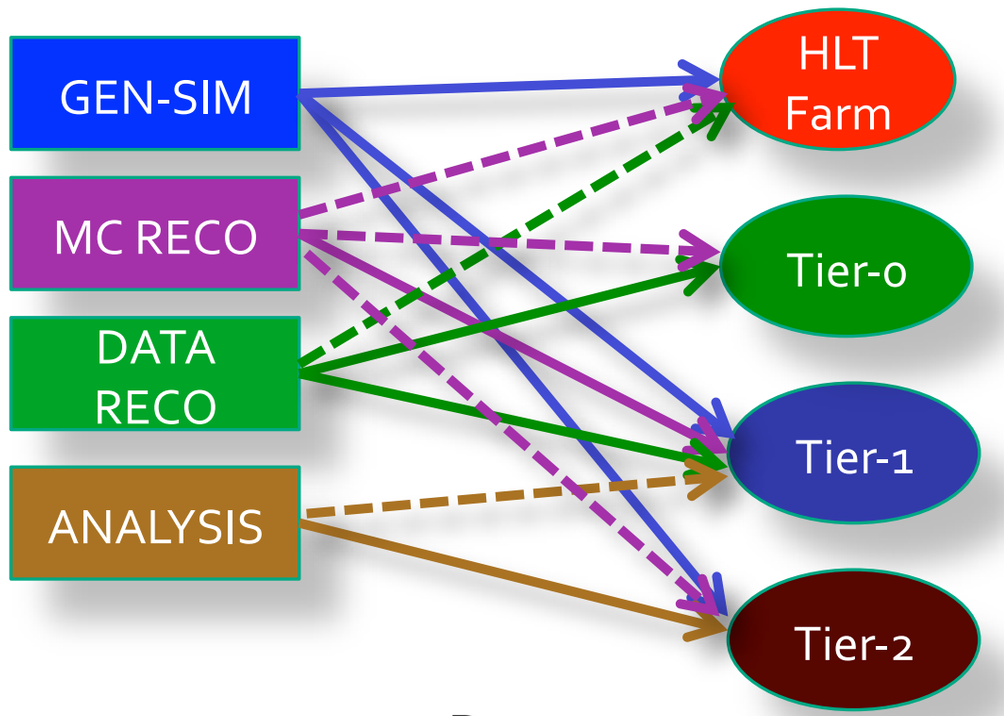
You can use the **Production Monitoring Platform** (PmP) to find out the how much has been produced of a specific request or of a production campaign. Get PrepId from das, search it in PmP

Producing datasets: distributed computing

Computing farms hosted by CMS institutions have different capacity of processing and storage

"Tiered" computing infrastructure: **different farms address distinct use cases**

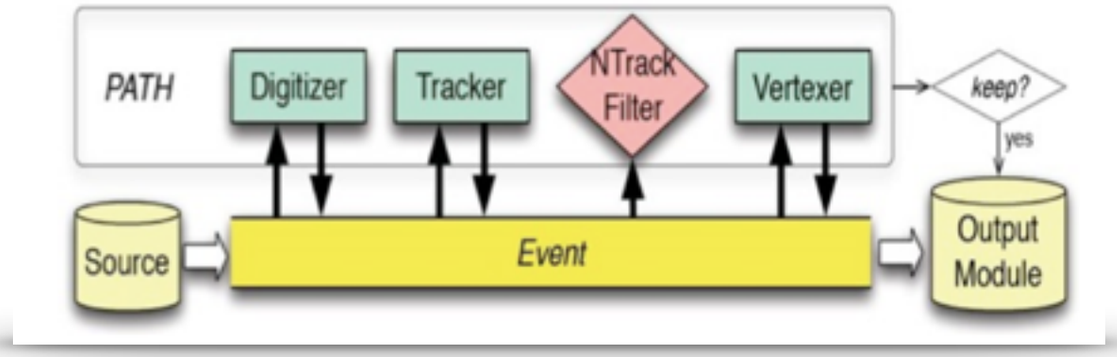
- **Tier0** (CERN): big farm @ CERN (~15k CPU cores)
  - dedicated to Prompt-Reco (used for other purposes during beam-off)
- **Tier1**:~10 big centers with large disk/tape cap act (tot. ~25k cores)
  - long term storage and central data processing (data and MC)
- **Tier2**: many smaller regional centers with disk "buffers" (tot. ~60k cores)
  - analysis jobs and MC simulations (or central jobs with low IO requirements)
- **Tier3**: even smaller centers for data analysis (no "central" jobs, usually "institute level")
- **CAF**: Central Analysis Facility: fast turn-around analyses or prompt feedback/calibrations
- **Opportunistic resource**:
  - big computer centers for limited periods (e.g. Amazon)
  - HLT farm when LHC not running (~15k cores)

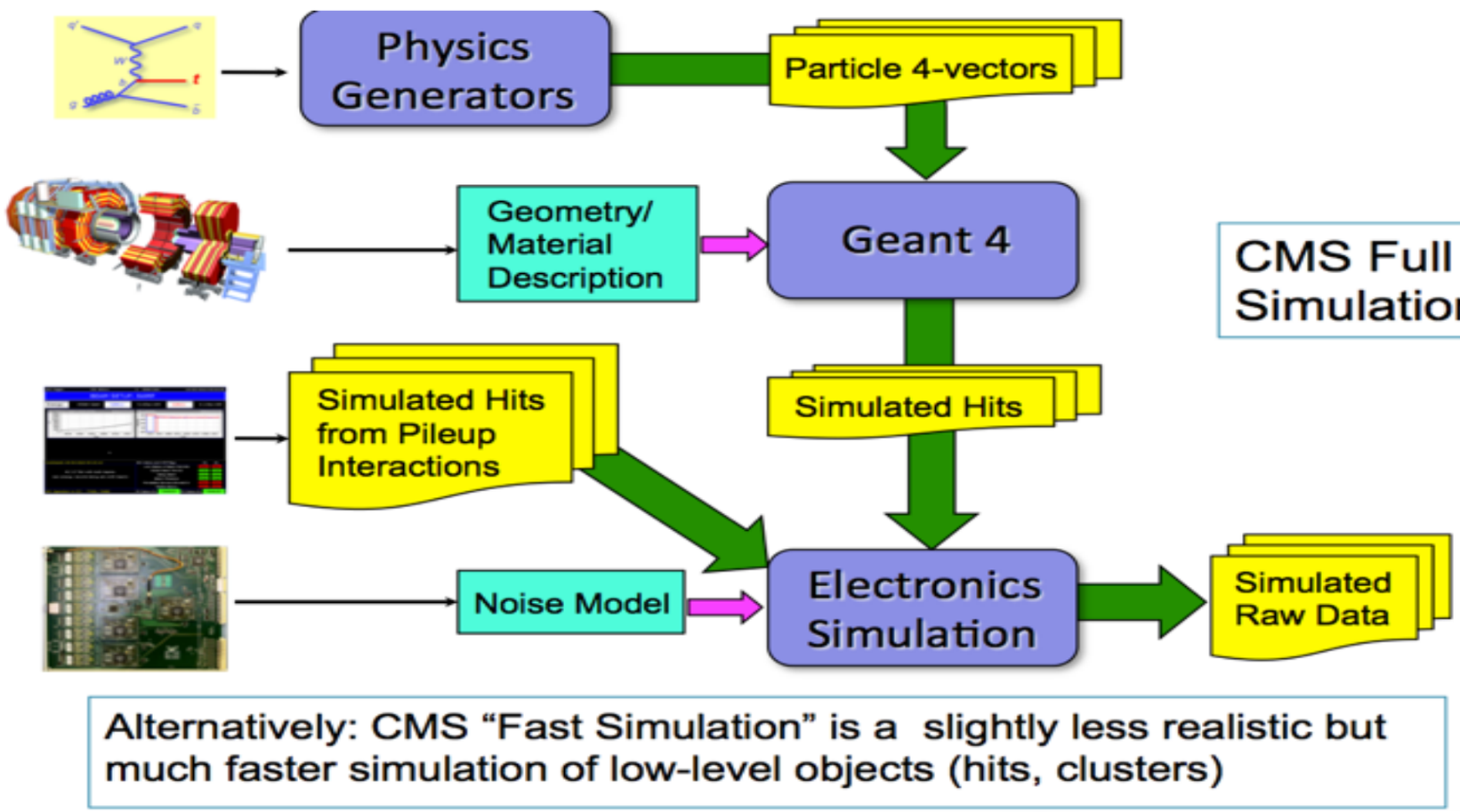


Simulation Flow

CMSSW simulation sequence:

- event simulation algorithms are implemented as “modules” communicating via the “Event”
- The Simulation sequence aims at producing MC truth + RAW data as if it came from CMS@P5





Data tier

GEN, LHE

CMS Full Simulation

GEN-SIM

GEN-SIM-RAW

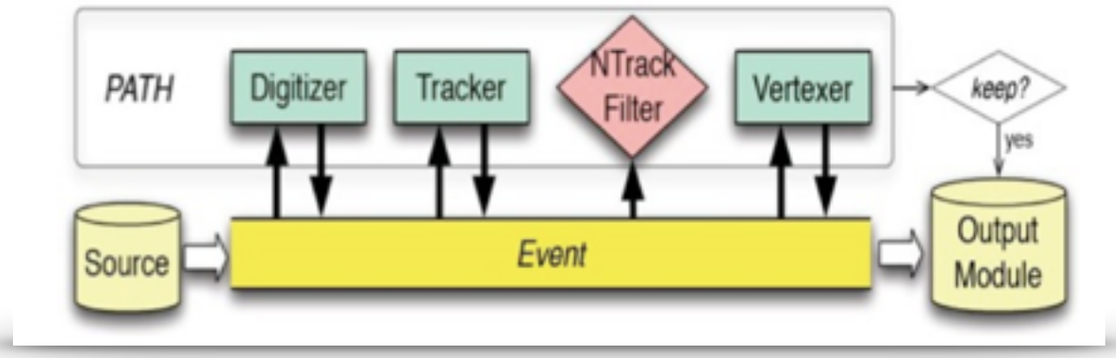
Alternatively: CMS “Fast Simulation” is a slightly less realistic but much faster simulation of low-level objects (hits, clusters)

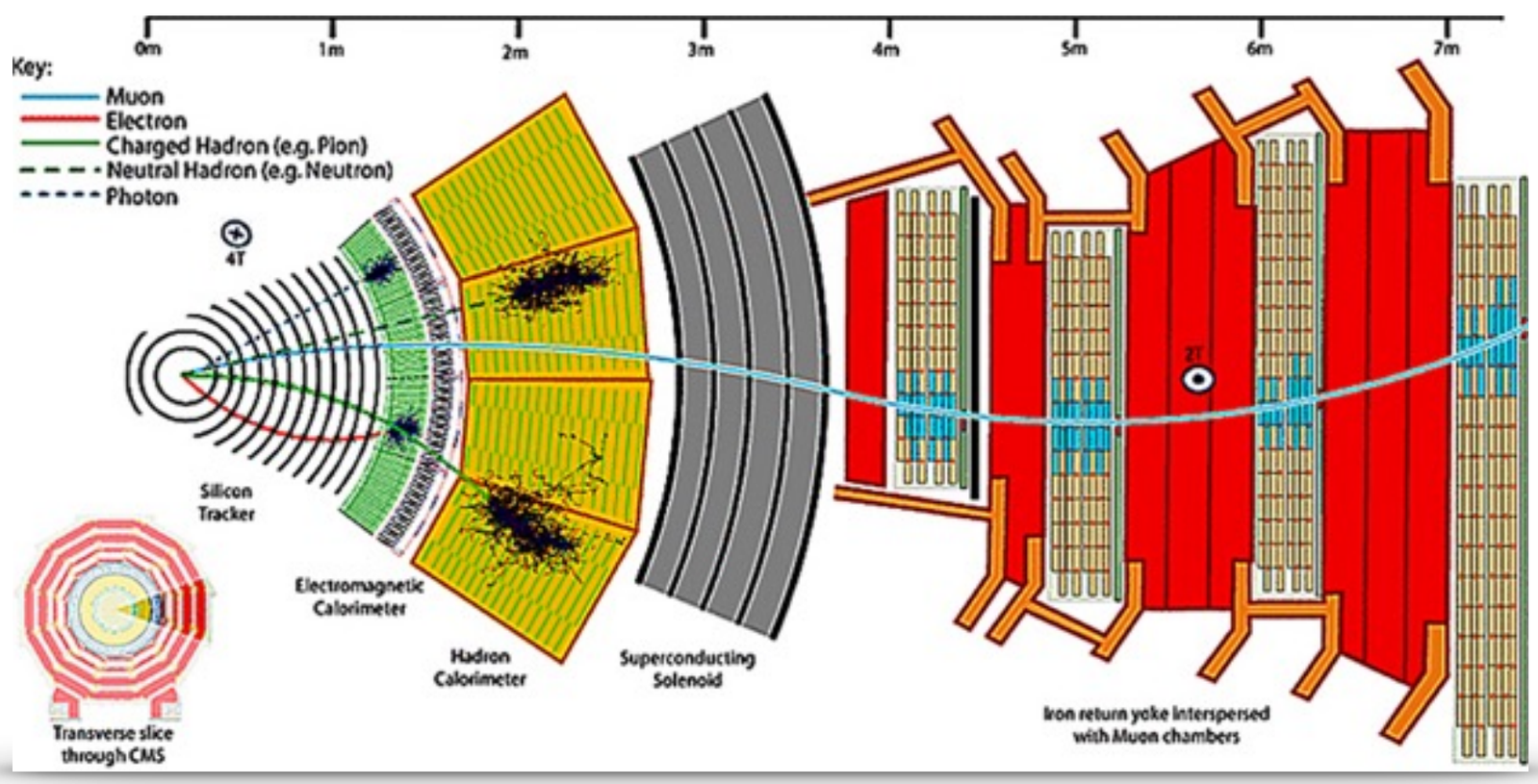
Contact

Reconstruction Flow

CMSSW reconstruction sequence:

- event reconstruction algorithms are implemented as “modules” communicating via the “Event”
- The Reconstruction sequence turns the binary output (RAW) from CMS/DIGI into **physically interpretable quantities** ready for data analysis
- Hits in the detector are aggregated in cluster and tracks, which in turn are matched to create **particle candidates** (PFAlgo): Tracks, muons, electrons, photons, jets …





Data tier

RECO, AOD

MINIAOD

MINIAODSIM

Contact

More on Data Tiers

Data Tiers: define the event content for the dataset.

The most common ones:

- RAW, RECO, **AOD, AODSIM, MINIAOD, MINIAODSIM**, USER, GEN, FEVT...

RAW ⟶ is what is repacked in root format @ Tier0
serves as input for any further processing step

- new release can always read old RAW files

RECO vs AOD:

- RECO is needed for dedicated studies and detector commissioning;
- its size per event 2-3 times the size of RAW
  ⟶ kept only for ~6 months after prompt-reco
  ⟶ not produced by default in reprocessing
- AOD subset (~40%) of RECO content meant for analysis

In runII one step further in data reduction: MiniAOD
(~10-15% of AOD)

- more than just a subset ⟶ designed for analysis

More on Data Tiers: miniAOD

Lightweight analysis data tier (30-50 kb/ev)
⟶ serve the needs of ~90% of the CMS analyses

- High level physics objects (leptons, photons, jets) with detailed information
- Compressed information for all PF candidates, to allow re-clustering jets, re-computing isolation
- Usable both in full CMSSW and in FWLite/Python

Centrally produced both for MC and Data

- can be re-produced from AOD to include newer high-level calibrations and improvements

Contact

Analysing Datasets

distributed data analysis is performed with crab

crab

Datasets need to be on disk @ Tier2/3 to be accessible by your analysis

Thanks to the GRID, "where" exactly they are doesn't matter; crab will match your analysis task to one of the suitable sites

- **AnalysisOps** subscribes the most common samples to "central" space ⟶ no action needed

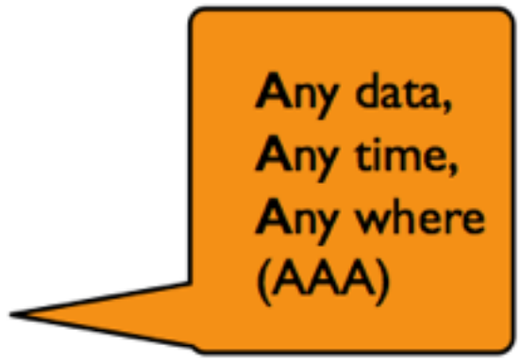
What if you need something more "exotic", which is not yet on disk? You can ask the transfer @ T2s using the Phedex service

Phedex

- **users** can request using the interface (complete datasets or "blocks" of files)
  ⟶ requests are associated to "groups" and assigned to a given "site"
- **data manager** approves/rejects the request, evaluating the available quota at the destination site
  ⟶ "standard" analysis use cases addressed by AnalysisOps group, each PAG/POG group has a person taking care of data-management
- **Dynamic Data Placement** tools
  ⟶ optimal data replication and quota control

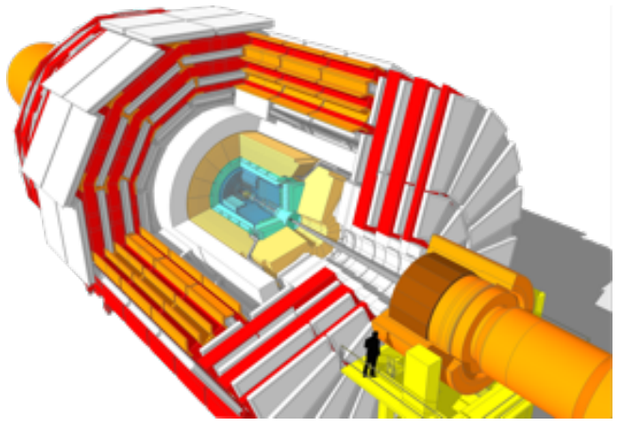
before submitting large analysis task (w/ crab) run on local test job on a single file
⟶ how do I do it

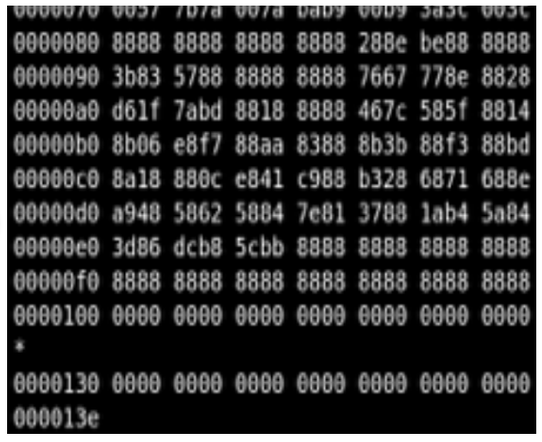
- most datasets are accessible remotely via XROOTD protocol
  ⟶ e.g. can run @ CERN reading files @ FNAL

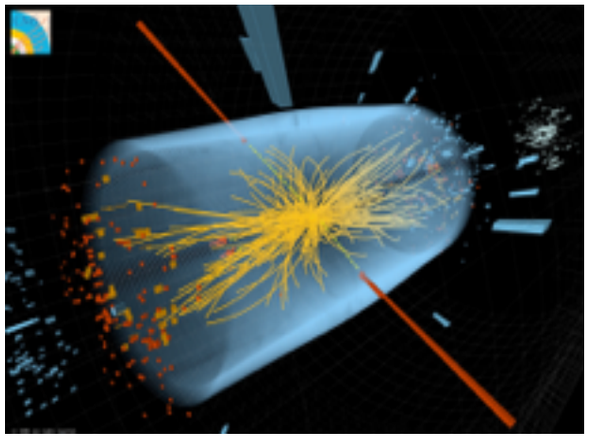


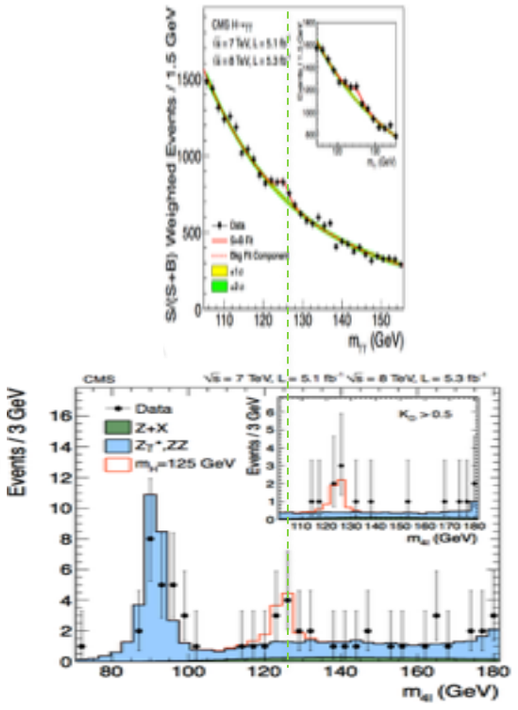
Contact

Data Preparation in a nutshell









Release Integration and Validation

CMS Software

CMSSW: one software project for all key event processing workflows:

- **GEN**erator, **SIM**ulation, **DIGI**tization, **RECO**nstruction **ANALYSIS**

The **algorithms** (a.k,a. producers) are in **C++**

The configuration of producers and their assembly in complex sequences are based on Python

git+github are used for code development, versioning and integration

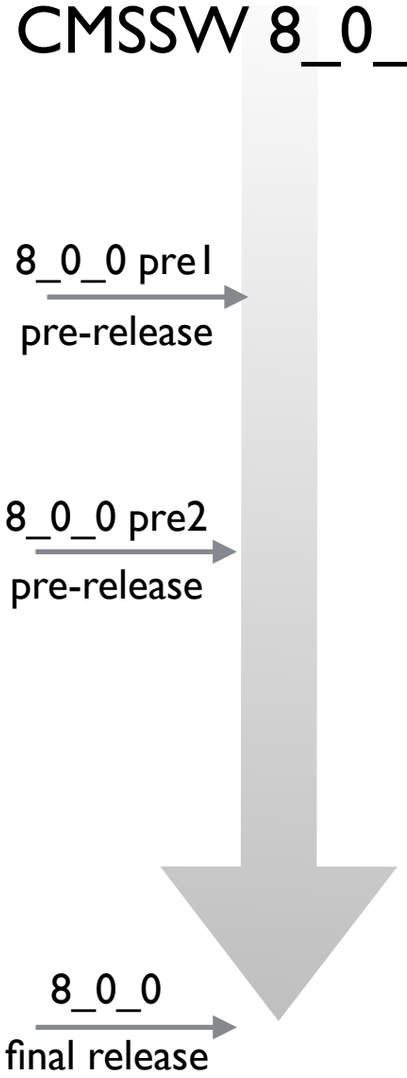
The release schedule follows a “train model”: << Dear Developer, catch this train or wait for the next>>

- 1 release every ~6 months, with a set of key milestones
  - tuned to major conferences & physics needs
  - e.g. : 7_6 for 2015 data rereco, 8_0 for 2016 pp data taking & MC
- bi-weely pre-releases during the development
  - frequent snapshots to ease integration and physics validation

cmssw @github

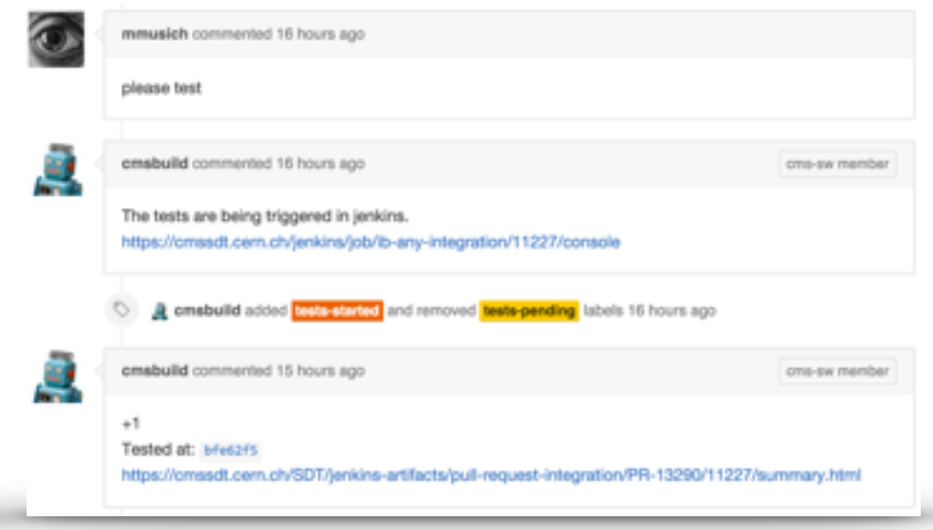
Contact

CMSSW: development and release integration



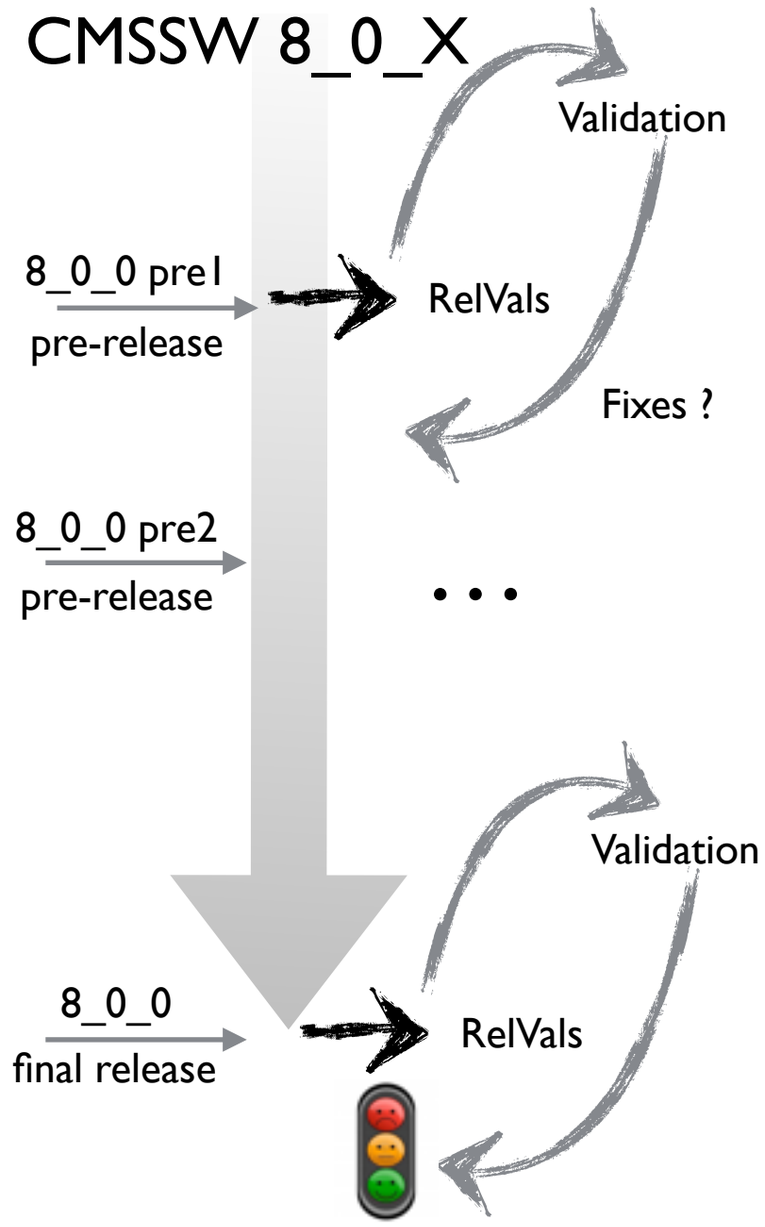
Integration of new code developments in CMSSW relies on:

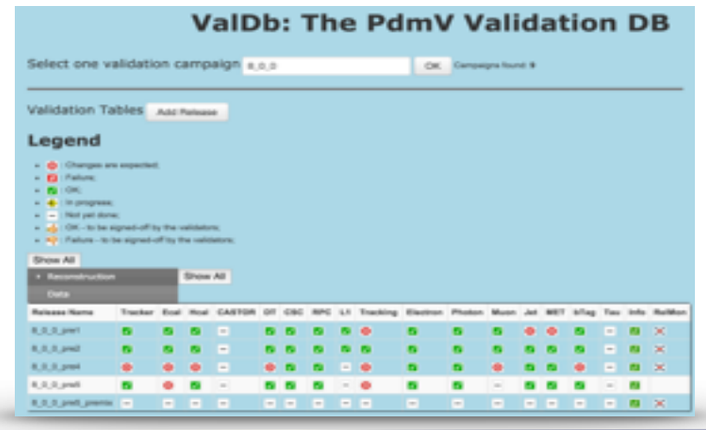
- automated Quality Assurance Tests producing plots any time a developer puts forward new code (pull request)
- scrutiny of each piece of new code proposed by code reviewers appointed among the L2 convenors/experts



Contact

CMSSW: release validation





Validation: iterative process performed all along release cycle, at each pre-release

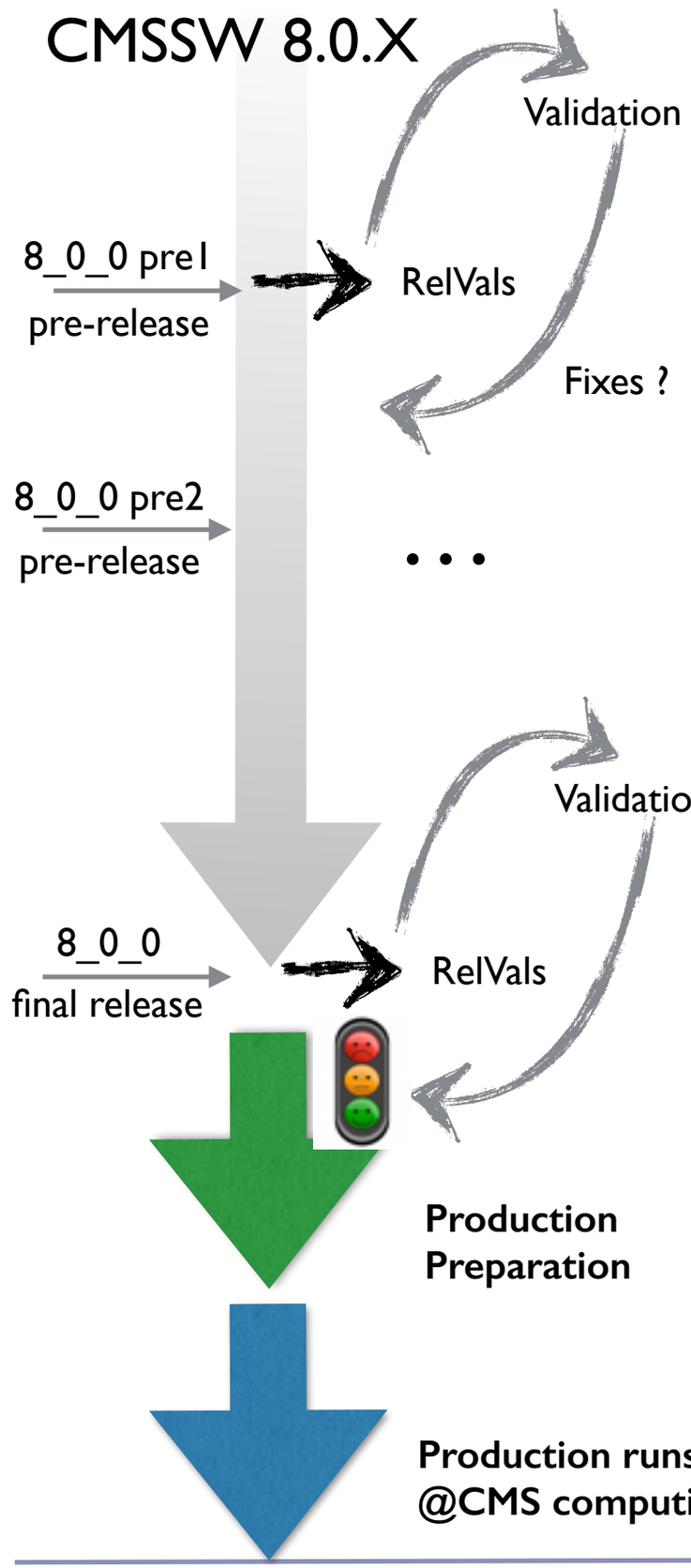
1. production of small scale datasets: **Rel**ease **Val**idation Test (**RelVal**) including DQM plots
2. DPG, POG and PAG validation experts check the plots
3. PPD/PdmV (L2 team in PPD) group coordinates the validation campaigns
4. sign-off on quality of release and calibrations

differential validation: compare plots of each release w.r.t the previous one

- each cycle takes 4-5 days for the RelVals production + 1 week for feedback from the validators

vadlb
Contact

CMSSW: ready for Physics Production



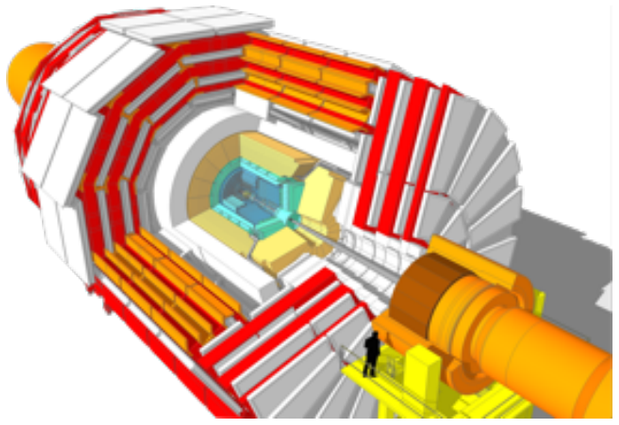
While a major release (e.g. 8_0_0) gets scrutinised for green-light
⟶ start preparation of the campaign, data reprocessing or MC production

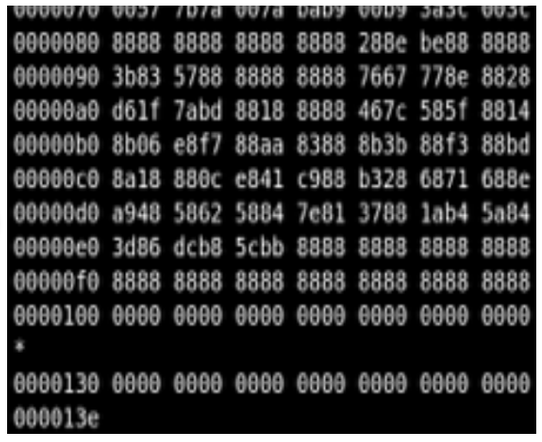
1. finalization of the alignment and calibration conditions, and their validation
2. finalization of the parameters for the Pile-Up overlay (PU scenario)
3. preparation of the injection machinery for the central processing by computing

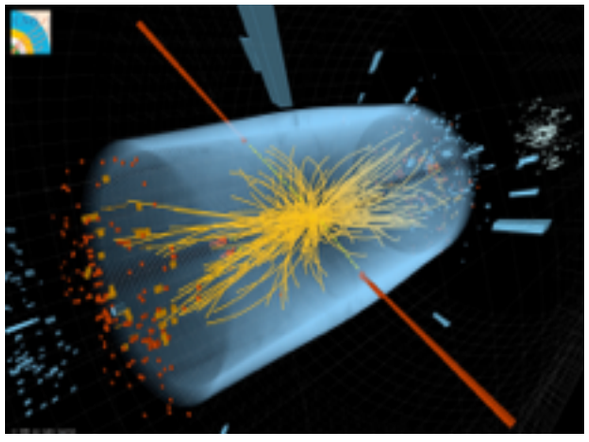
At the start of production, the Computing Operation team receives a and puts to work the production requests

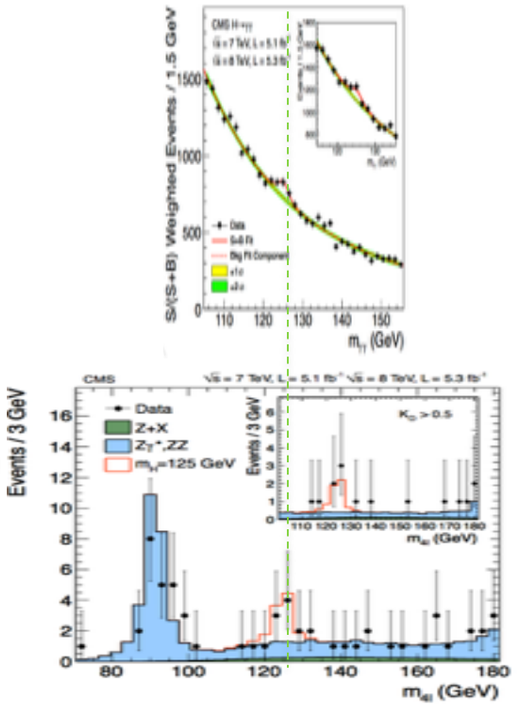
Contact

Data Preparation in a nutshell









Data Quality & Certification

Data Quality Monitoring

DQM (L2 team in PPD) provides tool and infrastructure to produce and distribute **plots** for **any CMSSW workflow** (SIM, RECO, miniAOD)

DQM-DC

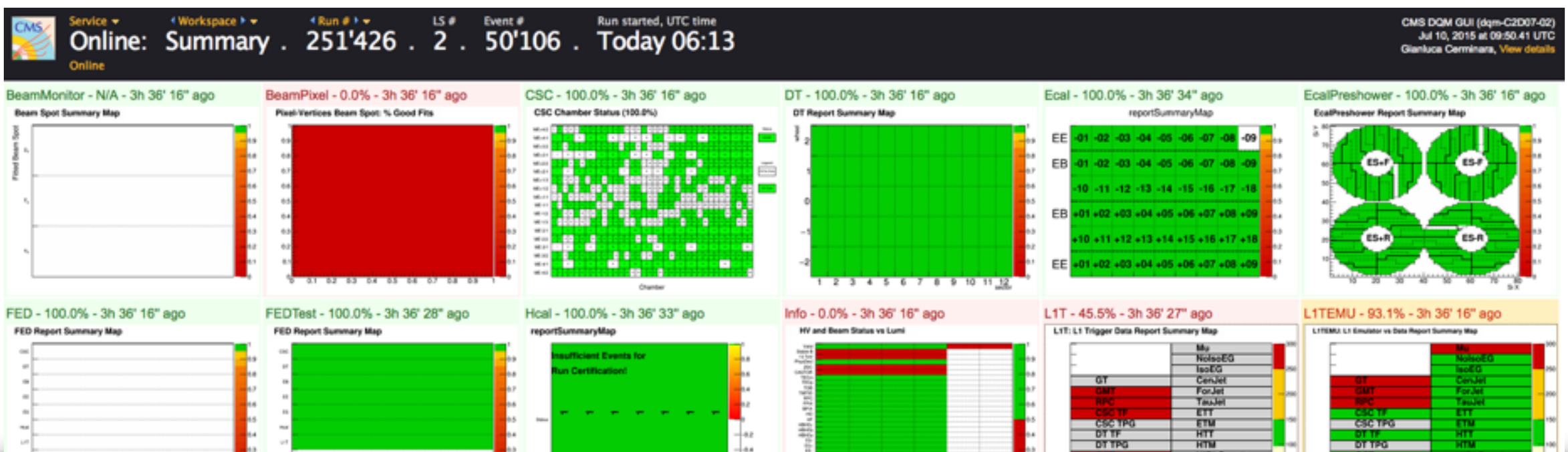
Contact

2 main areas of application:

- **online**: process events selected by HLT to display variables in the control room with very low latency
  - *live monitoring of detector* performance during data taking
- **offline**: process all events while they are simulated or reconstructed and fill diagnostic plots for detailed monitoring of the performance;
  - *data certification:* select collisions good for physics analysis
  - *validation+verification:* assess CMSSW and condition performance (PdmV, see later)

DQM Graphical User Interface: front-end web service to browse histograms for a given dataset/run, both data and MC

DQMgui



Data Certification

ex3

The complexity of the CMS detector and offline processing requires continuous monitor of data quality

Team of Detector and Physics Object experts check DQM plots for each run and select those usable for analysis

- look for unexpected effects that could affect analysis level quantities
  - e.g noise spikes, dead areas of detector, problematic calibrations

During data taking ⟶ continuous certification of prompt reco datasets

  - incremental updates every week

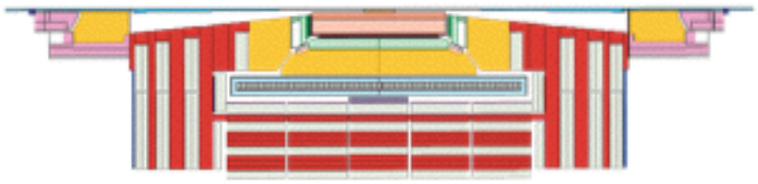
PPD/DQM-DC team → coordinates the validation activity

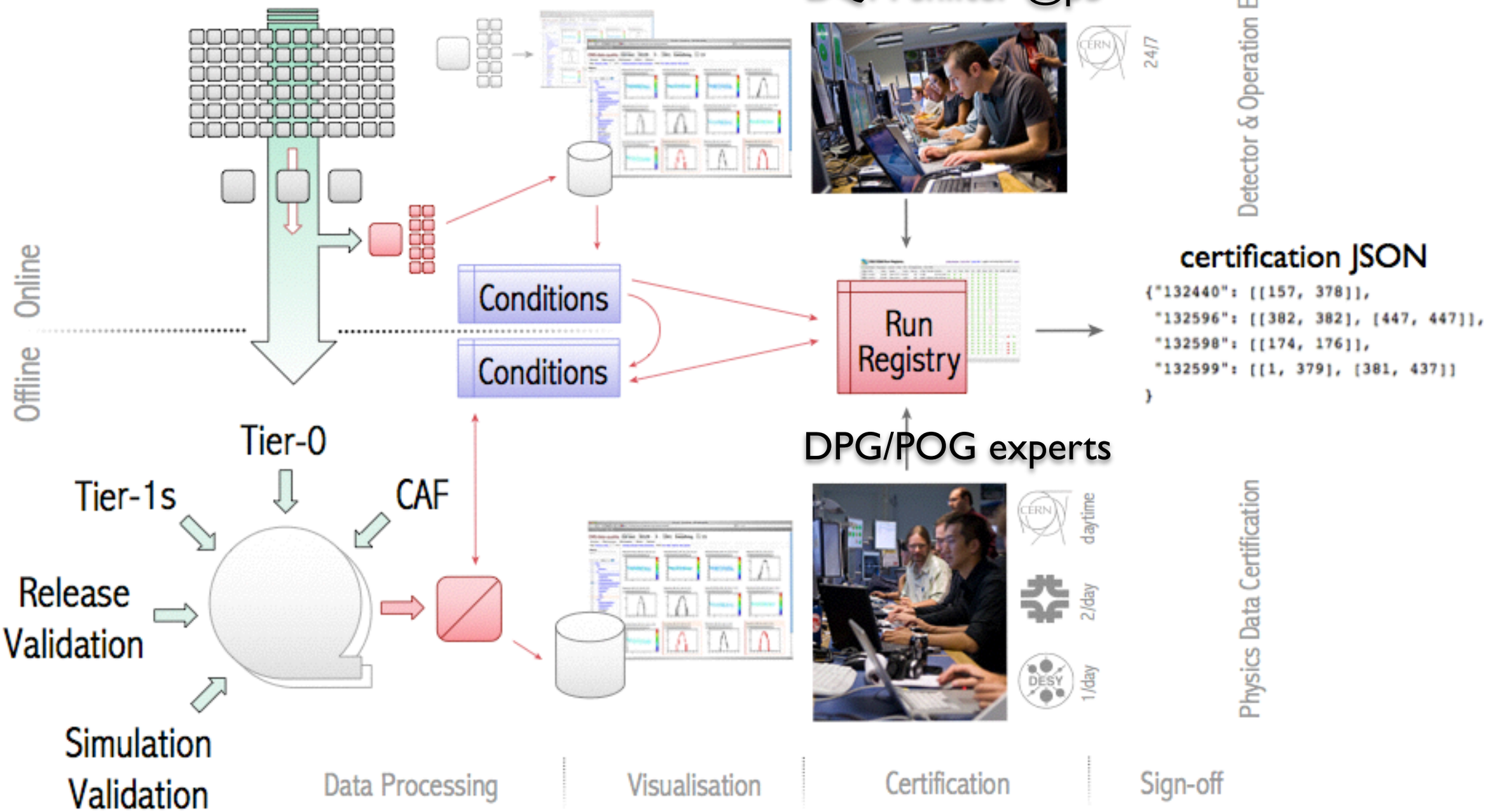
- Reports @ PPD General Meeting on Thursday 14:00 -16:00

Contact

Data Certification: flow

ex3





Certification ⟶ Analysis

ex3

Selection of LumiSections (LS) (≃ 23s of run) considered GOOD for physics

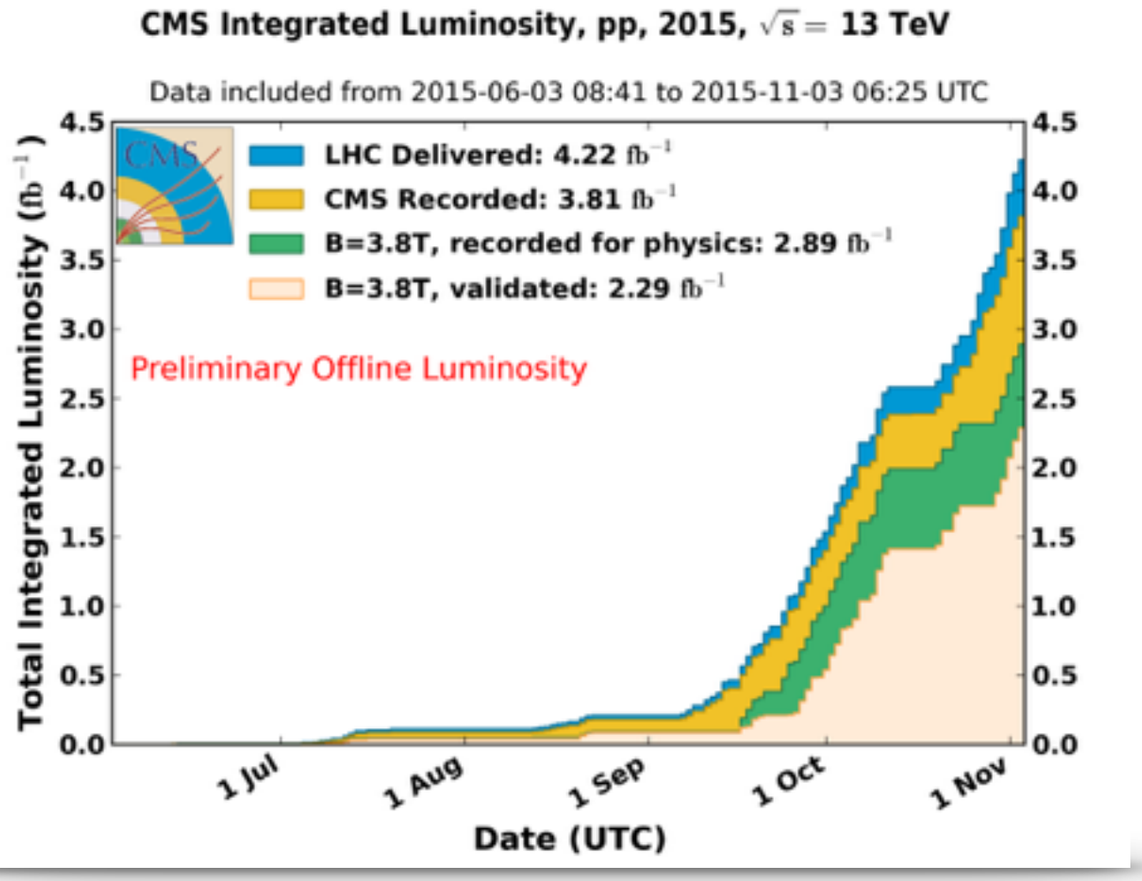
- distributed in JSON format
  ⟶ weekly for PromptReco
  ⟶ after each major reprocessing

Several "flavors" of the JSON file, to cater for analysis with different requirements:

- **golden** ⟶ require all sub-detectors/POGs to be "GOOD"
- **silver** ⟶ in 2015, require Tracker and muon detectors ncluding runs affected by low occupnacy issue in HF
- **muon-only** ⟶ no requirements on calorimeters
- **DCS-only** ⟶ require only Tracker to be powered

How do I use the JSON file:

- to be used in CRAB to run only on CERTIFIED LS of your dataset



CMS preliminary results: June 3rd-November 3rd 2015								
Tracker		Calorimeters			Muon Spectrometer			Operational Issue
Pixel	SST	ECAL	ES	HCAL	CSC	DT	RPC	Tracker HV ramp
99.5	99.8	99.8	99.8	81.2	100	98.2	98.8	99.5
All good for physics: 79%								

Luminosity weighted fraction (in %) of data certified as good for physics analysis relative to 2.8/fb of data recorded by the CMS experiment during 2015 proton-proton collisions at 13 TeV with magnet at 3.8T.

json-repo

Contact

Integrated Luminosity: HowTo

ex3

For your analysis, you need to quote the luminosity of ALL the LumiSections you run on, and ONLY those

- Limit the processing done by crab to the certification json required by your analysis
- CRAB reports the LS successfully processed by your jobs in the same JSON format used for certification

certification {JSON}


```
{ "251027": [[1, 60], [61, 61]]
  "251028": [[1, 18], [19, 21], [141, 141]],
  "251143": [[1, 16], [17, 19]],
  "251147": [[1, 24], [29, 29], [30, 187]]
}
```


CRAB output {JSON}


```
{ "251027": [[1, 60],
  "251028": [[1, 18], [19, 21], [141, 141]],
  "251143": [[1, 16], [17, 19]],
  "251147": [[1, 24], [29, 29], [30, 187]]
}
```


The Lumi POG and Bril DPG provide tools to compute luminosity starting from a JSON file:

- `brilCalc.py`: reports the online measurements of the luminosity (by run or by lumi-section):
  - delivered by the LHC,
  - recorded CMS
- Options are available to select measurement from a specific luminometer or their best combination

NOTE: using directly the certification JSON to compute integrated luminosity can be problematic for several reasons:

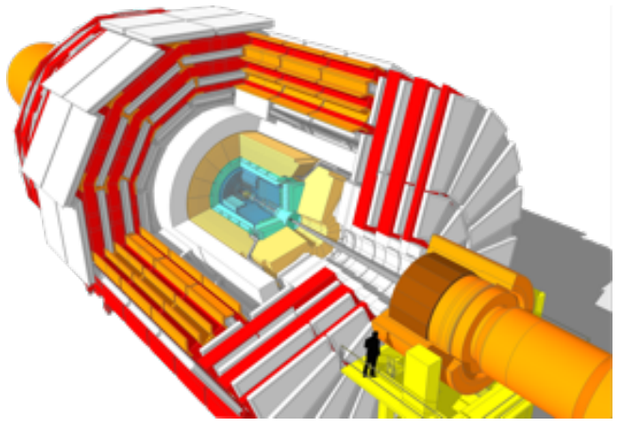
bril

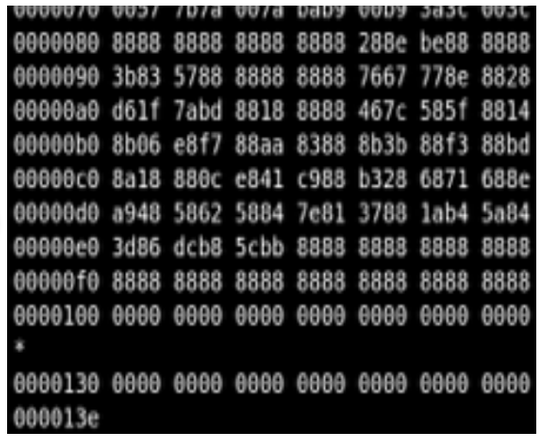
lumi

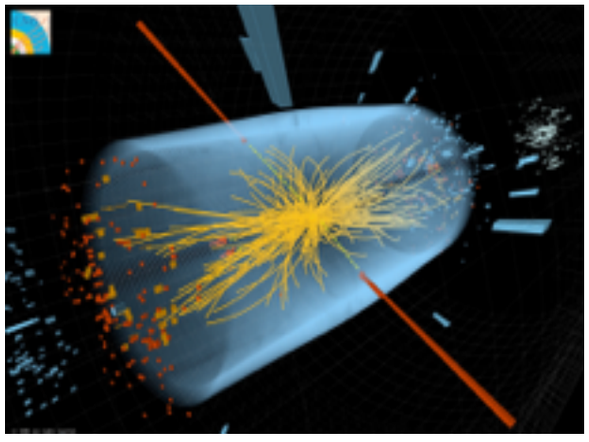
Contact

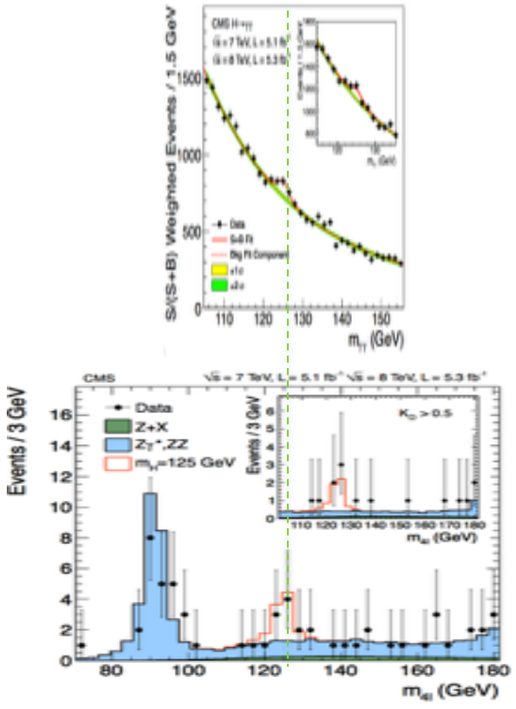
- LS that failed prompt reco or re-reco will appear in the cert. JSON
- LS that failed in your analysis jobs will appear in the cert. JSON

Data Preparation in a nutshell









Alignment and Calibration

ex4

Alignment and Calibration Workflows

Must provide most up-to-date conditions @ all stages of the data processing (AlCa/DB L2 team in PPD)

AlCaDb

Critical to sustain the quick peace of analysis turn-around
Different workflows depending on the time scale of updates:

- **quasi-online calibrations** for HLT and express (e.g. beam-spot used by HLT)
- **prompt calibrations**: monitor/update conditions expected to vary run-by-run (or even more frequently) ⟶ exploit 48h delay of prompt-reco
- **offline workflows** for re-reco passes and analysis level conditions (e.g Jet Energy Corrections, Muon Momentum Scale)

A consistent set of conditions for a given task (i.e. HLT, prompt-reco, analysis... etc) is provided by a **GlobalTag** (GT)

- non trivial interdependencies between calibrations
- ~300 sets of parameters are needed to calibrate and align CMS

Dedicated GTs for analysis are made available by AlCaDB team
⟶ to be used with consistent dataset

Contact

ex4

Alignment and Calibration Workflows

Providing the most up-to-date **conditions** at all stages of the **data and Monte Carlo** processing is a major challenge involving:

AlCaDb

- all DPGs and several POGs
- AlCa L2 team in PPD

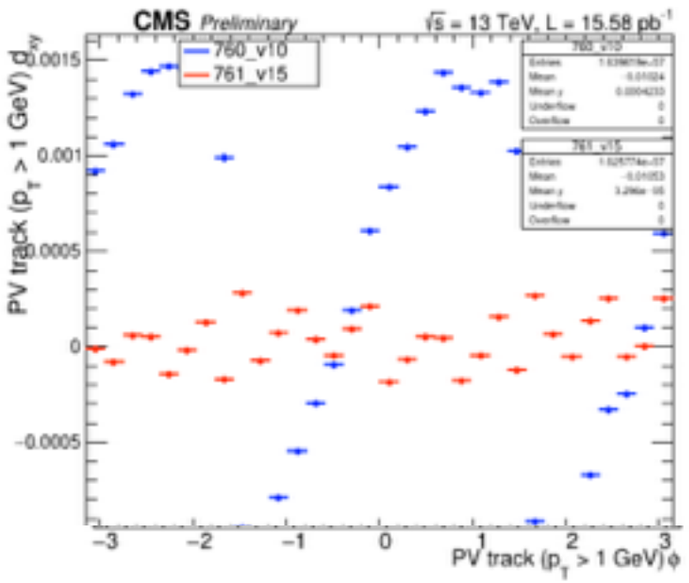
Critical to sustain the quick peace of analysis turn-around

- Prompt-reconstruction must be of near-perfect quality; CMS can only afford ~1 data reprocessing per year

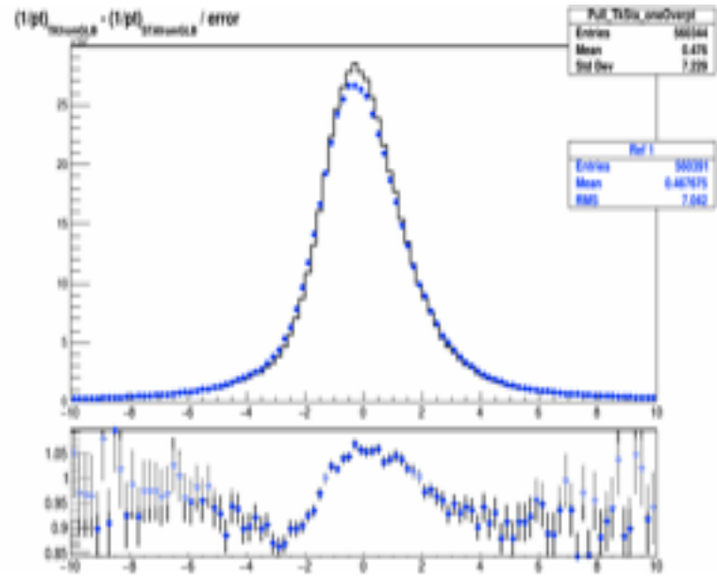
Different workflows depending on the time scale of updates:

- **quasi-online calibrations** for HLT and express processing
  - e.g. beam-spot used by HLT
- **prompt calibrations**: monitor and update conditions expected to vary run-by-run, or even more frequently - essential to guarantee performance of prompt reach
  - e.g. : offline beamspot, ECAL transparency,
- **offline workflows** for data reprocessing and analysis level conditions
  - e.g: Jet Energy Corrections, Muon Momentum Scale

prompt15
updated AlCa



prompt15
updated AlCa



Contact

ex4

Conditions consumption: Global Tag

Any CMSSW application, running in an online workflow at HLT/DQM or offline in data/MC processing, requires 200-400 sets of AlCa parameters ⟶ a.k.a. tag

tag

- **set of AlCa parameters** measured by calibration experts in DPG/POGs, released to dedicated database
- can be **time dependen**t, comprising multiple Intervals of Validity (IOVs)
- is in general be **dependent on several other** sets e.g.
  - tracker alignment ⟶ beamspot, muon alignment
  - calorimeters calibration ⟶ jet energy corrections
- e.g.
  - TrackerAlignment_run2Asymptotic_v1_mc, HcalChannelQuality_v2.10_mc, EcalIntercalibConstantsMC_digi_2011_V3_Bon_mc

GlobalTag (GT)

- **consistent and complete sets of tags** to run a CMSSW process
- a single entry point to retrieve all conditions consumed by a given workflow
- identified by a string
- e.g.
  - 76X_mcRun2_asymptotic_2016EcalTune_15fb_v0, 76X_mcRun2_asymptotic_v14

Dedicated GTs for analysis are made available containing also the high level calibrations which depend on all other ones (e.g. jet energy corrections, BTV discriminant calibrations)
⟶ to be used with consistent dataset at analysis level

Contact

Operation of AlCaDB

During data taking, AlCa conditions must be updated in:

- L1 and HLT trigger
- Prompt-Reco

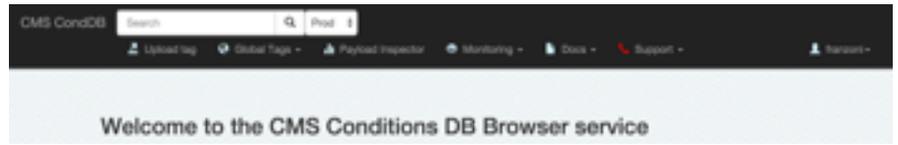
due to multiple reasons:

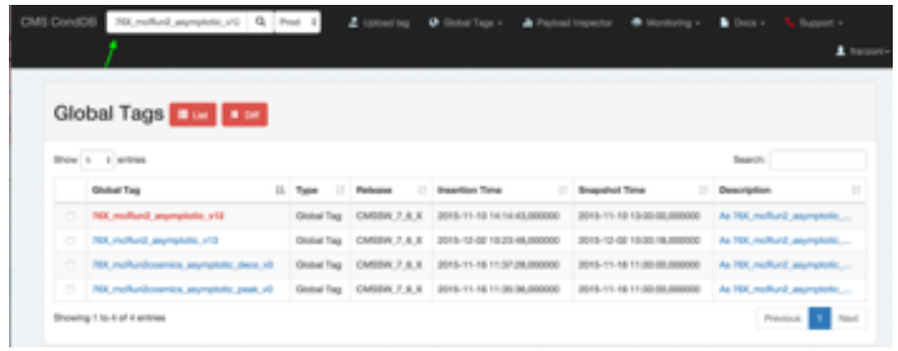
- tracking detectors move, e.g. with magnetic cycles
- configuration of data taking at P5 evolves (L1 menu, problematic channels)
- radiation-induced miscalibration

A limited set of "emergency" calibration can be deployed as soon as ready, if vital

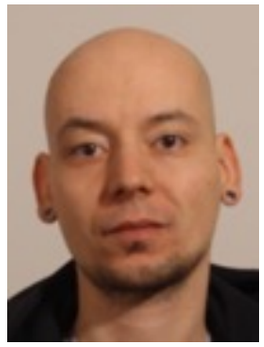
To recover performance, AlCa conditions undergo a close scrutiny - in 2015

- conditions needing update are aggregated every Monday
- relVal and DQM used by DPG/POG experts to sign-off conditions before deployment





PPD core (dream) team



Physics Performance & Datasets (PPD) organisation

PPD Coordinators
Gianluca Cerminara
Maurizio Pierini

PPD Resource Manager
Kirsti Aspola

Physics Data & MC Validation (PdmV)
Francesco Fabozzi;
Hengne Li;
Norraphat Srimanobhas;

Data Quality Monitoring & Data Certification (DQM-DC)
Federico De Guio;
Broen Van Besien;
TBD;

Alignment Calibration & Database (AlCaDB)
Giacomo Govi;
Marco Musich;
Giovanni Franzoni;

DPG and POG Validation
Mengqing Wu;

PAG Validation
Patricia Rebello Teles;

Monte Carlo Request Management
Hongbo Liao;
Denys Lontkovskyi;

Dataset Definition Team
Dustin Anderson
Kai-Feng Chen;

Online Development
Broen Van Besien;
TBD

Offline Development
Dmitrijus Bugelskis;

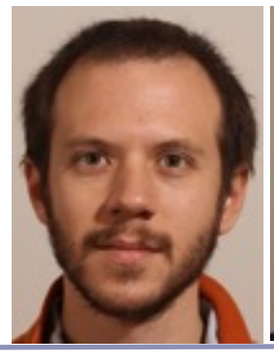
Data Certification
Laura Borrello;
TBD

Software Coordinator
Gregor Mittag

AlCa-HLT Contact
Thiago Tomei
Tongguang Cheng

DPG, POG & PAG contacts and validators





Conclusion and Outlook

Data Preparation is a complex endeavour

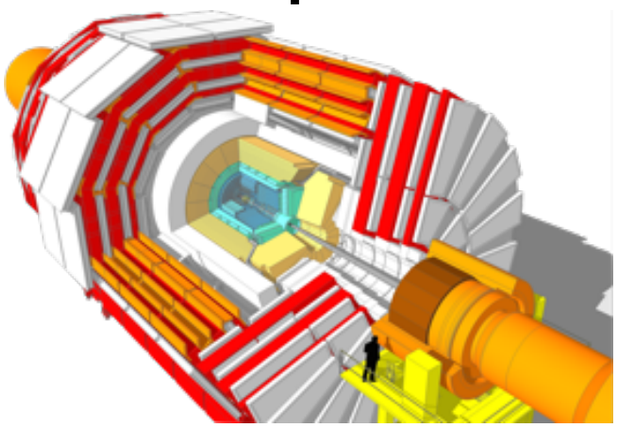
- this lecture hopefully unveils some of it to ease your life as data analyst

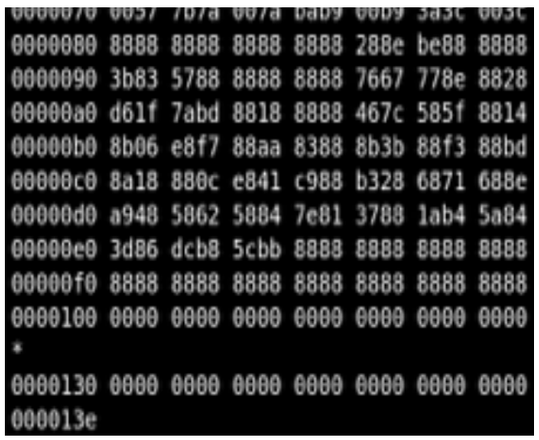
Delivering consistent & high quality ingredients for analysis at CMS entails plenty of technology, infrastructure, operations and clever&energetic collaborators

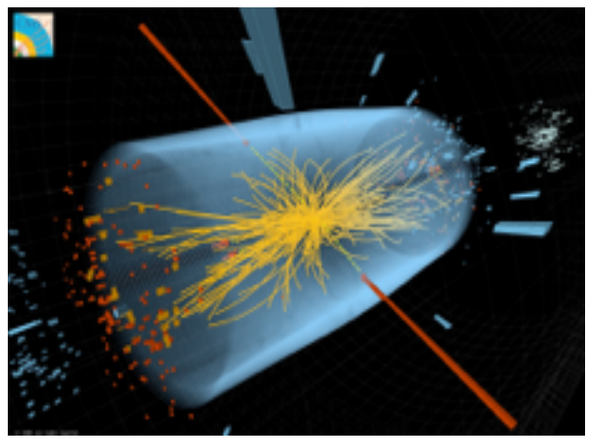
Plenty of opportunities for exciting work in these fields:

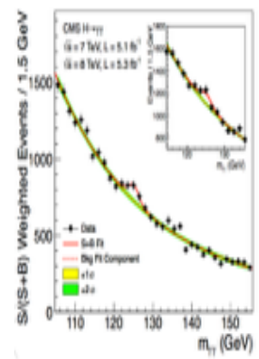
- a single person can actually have a large impact (and have fun!)

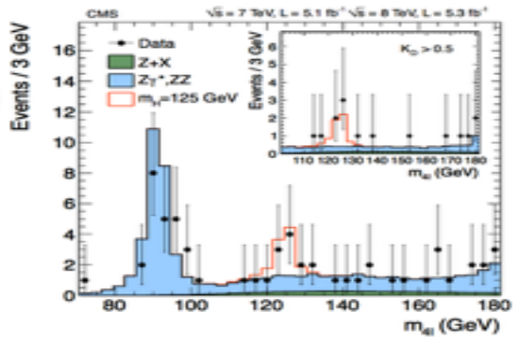
Data Preparation in a nutshell





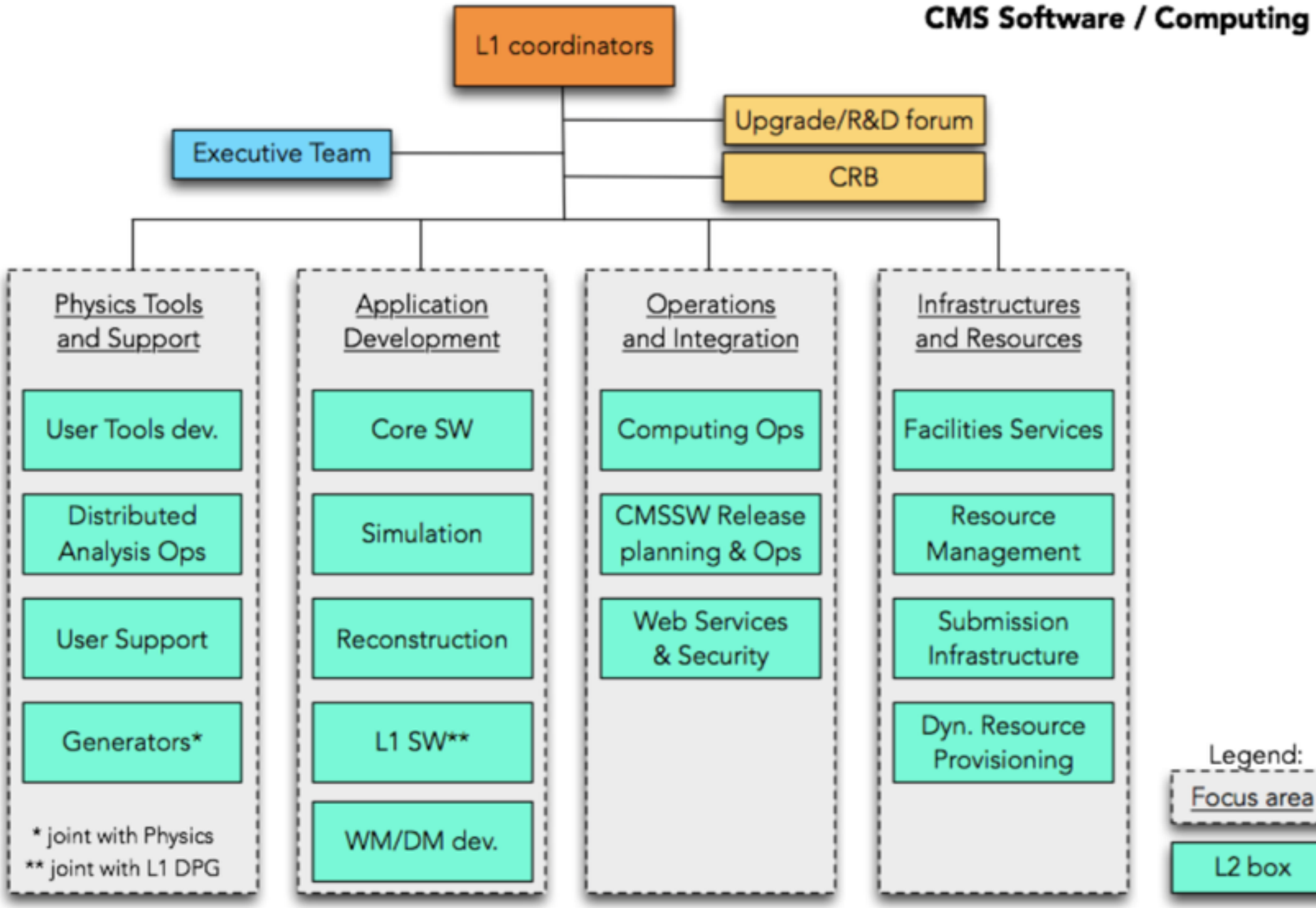






additional material

Offline & Computing
Project reorganisation



Off&Computing Organization @2015

Offline and Computing Co-Coordinators
M. Girone, D. Lange

Offline & Computing Management Board
D. Bonacorsi, P. Elmer, O. Gutsche, L. Sexton-Kennedy

Project Office
T. Boccali, I. Fisk

R&D
D. Bonacorsi, P. Elmer

Physics Support
S. Belforte, J. Lett, S. Malik, J. Hernandez

Data Management Development
N. Magini, T. Wildish

Workflow Management Development
E. Vaandering

Dynamic Resource Provisioning
D. Colling, C. Grandi, D. Hufnagel

Operations
C. Wissing, C. Pauss

Security
D. Da Silva, M. Altuna

Core Software
C. Jones, S. Muzaffar

Event Generators
J. Bendavid, R. Covarelli

Simulation
M. Hildreth, V. Ivantchenko, S. Sekmen, N. Vanelderen

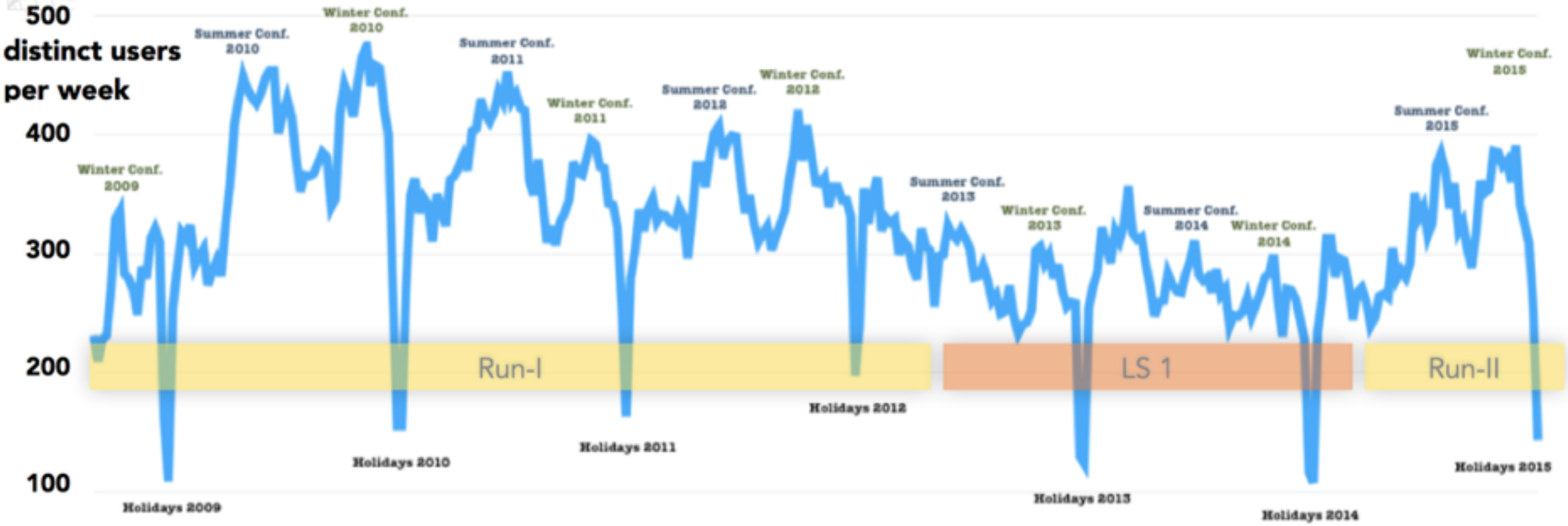
Reconstruction
S. Krutelyov, C. Vuosalo

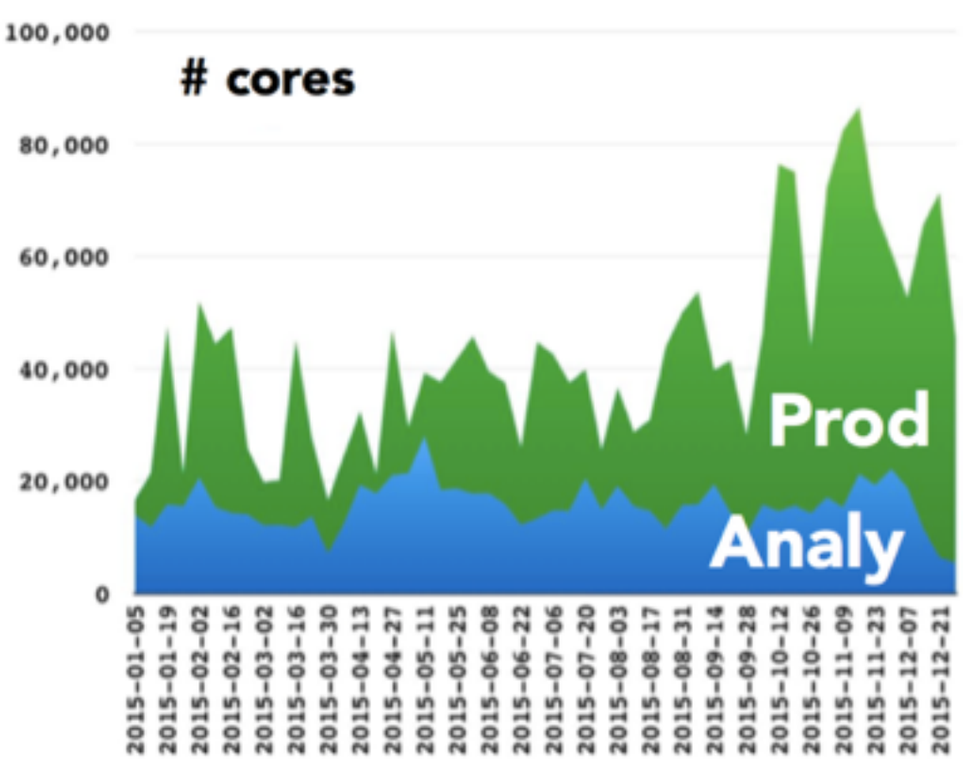
Analysis Tools
T. Jeong Kim

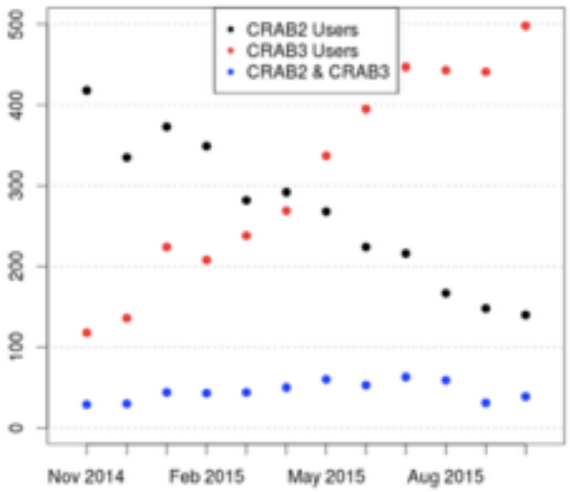
Upgrade Software
D. Elvira, M. Grimm

L1 Software
M. Mulhearn

Distributed analysis







CRAB3 migration done.

In 2015:

- **560** CRAB submitters per month (+8% w.r.t. 2014)
- **18k** analysis job slots used for analysis at T2s
- **~1.3M** analysis jobs / week (**190k** /day) (+26% w.r.t. 2014)

CMS Physics week - 8-12 Feb 2016 D. Bonacorsi for Sw/Comp 11

References

Page	Contacts	Documentation
3	cms-ppd-coordinator@cern.ch cms-offcomp-coordinator@cern.ch	PPD Main Twiki Offline Main Twiki Computing Main Twiki
10	hn-cms-dataset-definition@cern.ch	DDT Twiki
12	hn-cms-computing-tools@cern.ch	DAS
1731	hn-cms-prep-ops@cern.ch	PdmV Twiki
21		Computing Model Workbook
22,23,28,29	hn-cms-offlineAnnounce@cern.ch hn-cms-relAnnounce@cern.ch	Offline Workbook SW Guide
25	hn-cms-physTools@cern.ch	MiniAOD Workbook
26	hn-cms-phedex@cern.ch	XROOTD doc Phedex - Phedex Workbook
30	hn-cms-relval@cern.ch hn-cms-physics-validation@cern.ch	PdmV Twiki
33	hn-cms-evfdqmannounce@cern.ch	DQM Twiki
34	hn-cms-data-certification@cern.ch	DQM-DC Twiki
36	hn-cms-data-certification@cern.ch	JSON File Twiki
37	hn-cms-luminosity@cern.ch cms-dpg-conveners-bril@cern.ch cms-pog-conveners-lum@cern.ch	brilcalc Doc bril dpg lumi pog
39-41	hn-cms-alca@cern.ch	AlCaDB Twiki GlobalTag Twiki