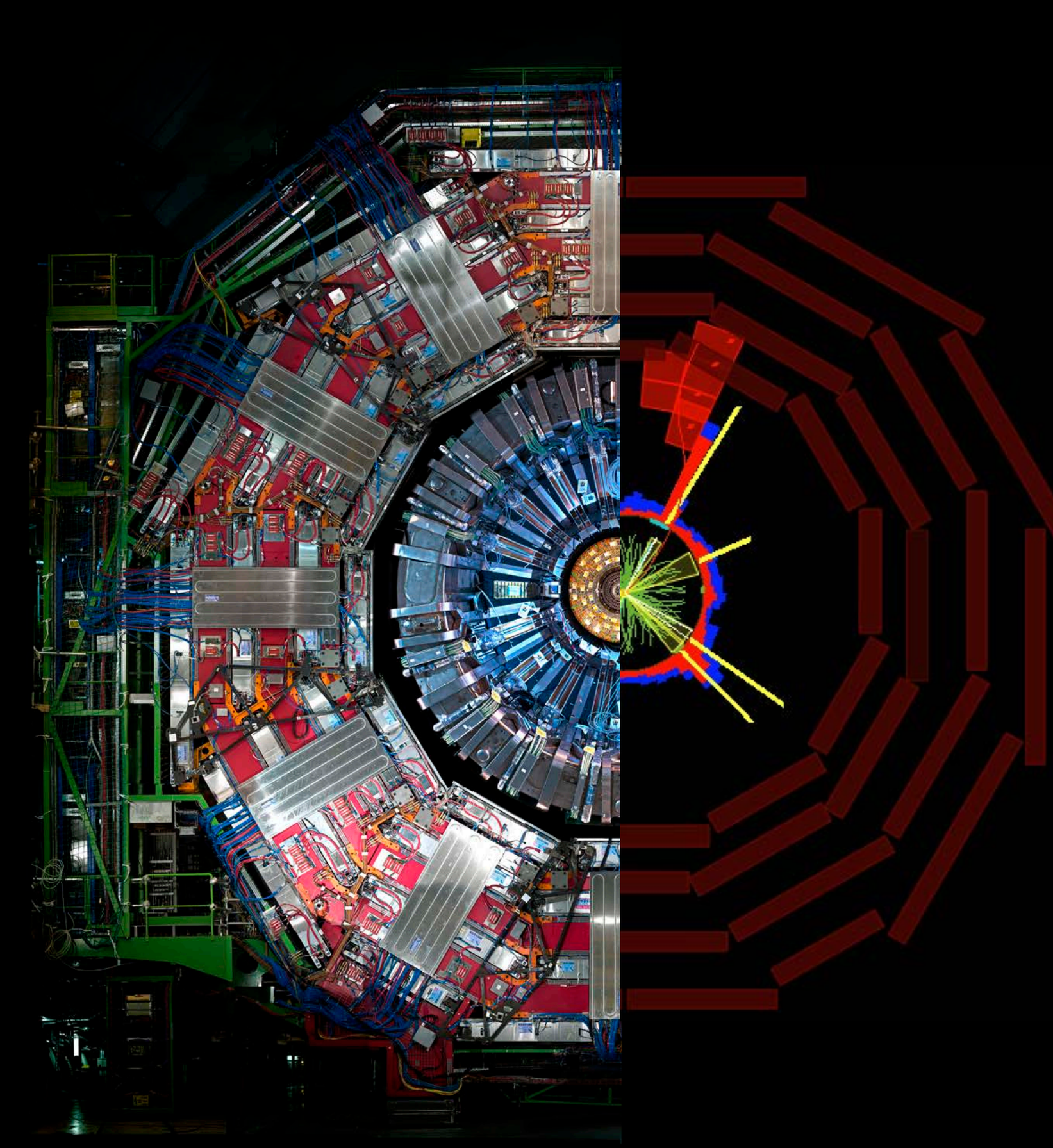
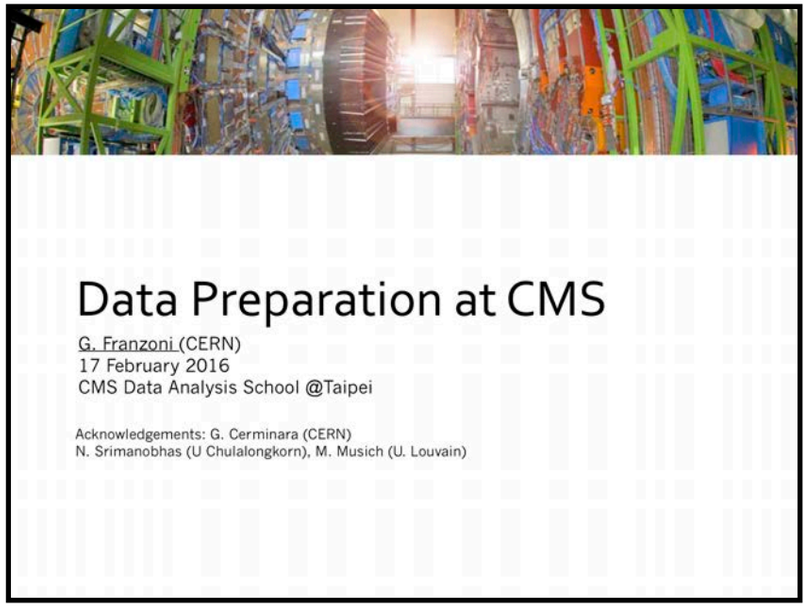


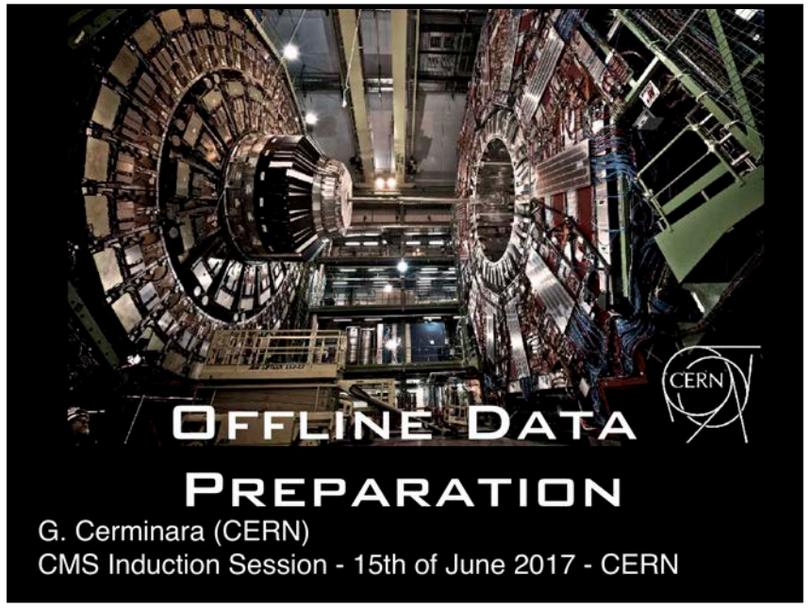
Exercise: Physics Performance and Datasets

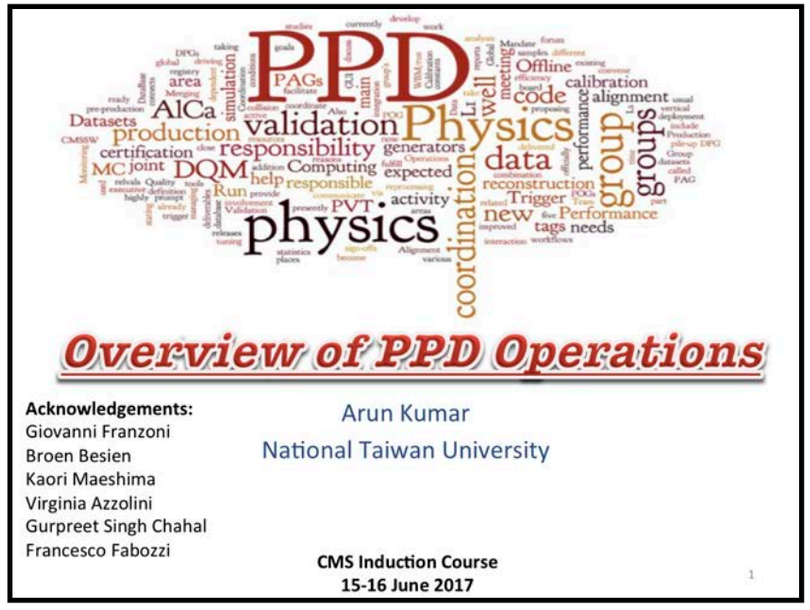
Norraphat SRIMANOBHAS
(Chulalongkorn U., CERN)
On behalf of PPD

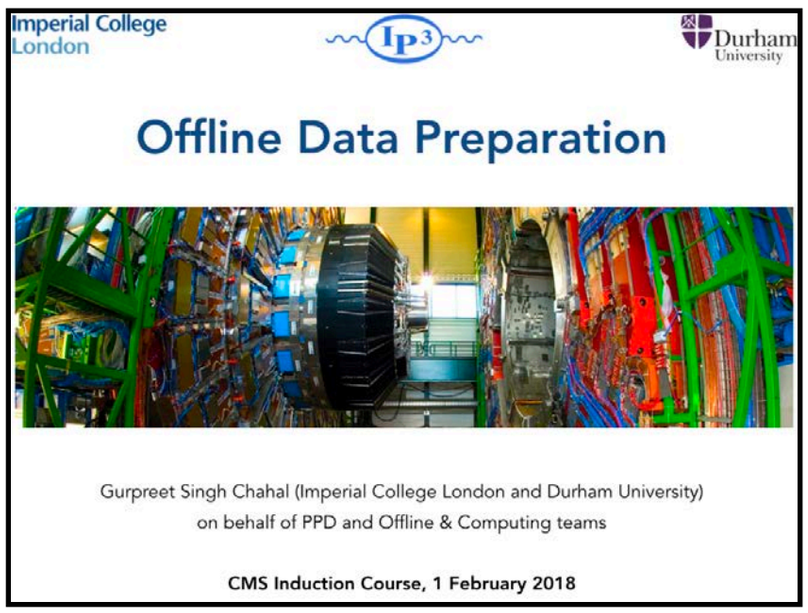
CMSDAS 2019 Pisa,
INFN Pisa,
28 Jan - 1 Feb 2019

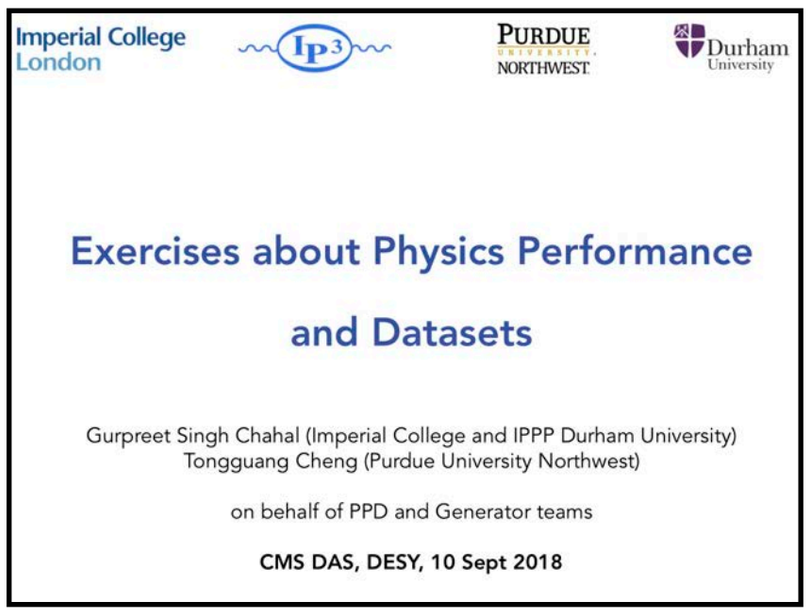
Credits



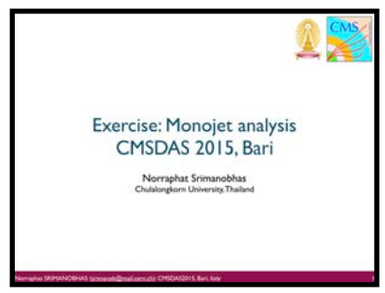


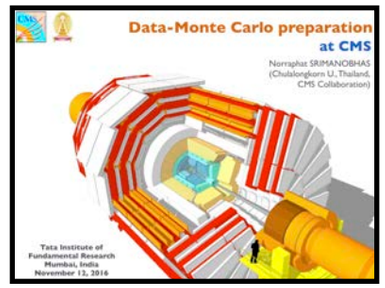






and all feedbacks to my previous talks …





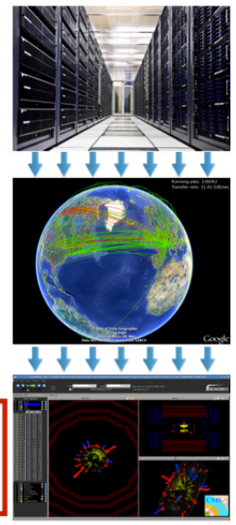
Outline

Outline: PPD + O&C

Real data

- From P5 to offline analysis
- Trigger (another lecture)
- Prompt-Reco & PCL at T0
  - Alignment and calibration
- CMS Primary Datasets
  - Scouting and Parking
- Event contents & data tiers
- Data quality monitoring
  - Data certification

Detector/Data operations are not far from you. If we don't have qualify data, we will not have publications.



Monte-Carlo data

- From simulation to offline analysis
  - Simulation flow
  - Pileup mixing
  - Trigger (another lecture)
  - Reconstruction flow
  - CMSSW
  - Productions
    - Distributed computing
    - Search, Access, Transfer
    - Analysis

Outline: Exercises

In this set of exercises, we have

- Exercise 0: Starting your RunII analysis
- Exercise 1: Exploring dataset using DAS : searching 2018 EGamma dataset as an example
- Exercise 2: Explore information for a Monte Carlo miniAODSIM sample from DAS, McM and pMp
- Exercise 3: Miscellaneous details about datasets
- Exercise 4: Generating MC NanoAOD events from scratch
- Exercise 5: Explore GlobalTag through cmsDBbrowser
- Exercise 6: Compute the integrated luminosity collected by CMS in RunII

Introduction

- Real data flow
- Monte-Carlo flow
- Analysis

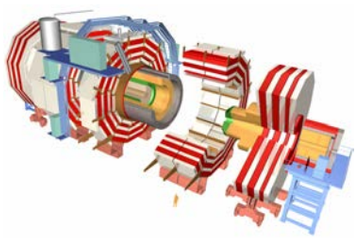
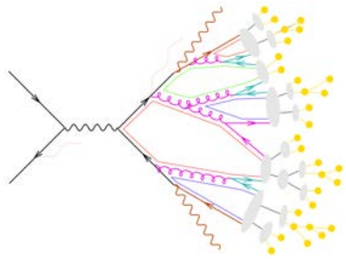
Exercises

Our target in CMSDAS

- Exploring CMS datasets
- MC production from scratch
- Conditions
- Compute integrated luminosities used in your analysis

Outline: PPD + O&C

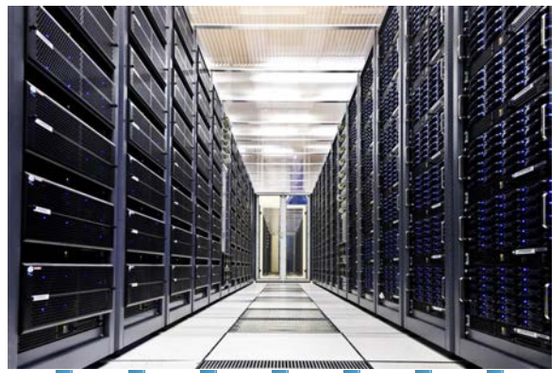
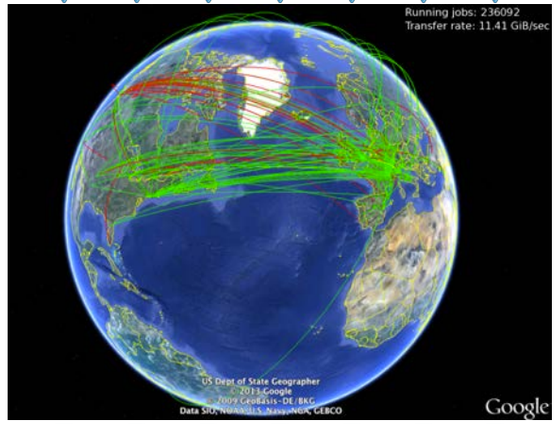


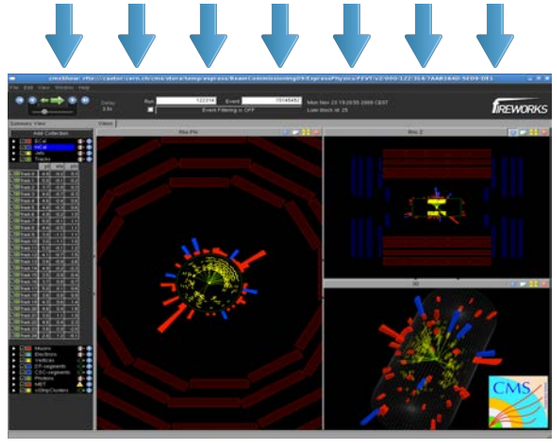



Real data

- From P5 to offline analysis
- Trigger (another lecture)
- Prompt-Reco & PCL at T0
  - Alignment and calibration
- CMS Primary Datasets
  - Scouting and Parking
- Event contents & data tiers
- Data quality monitoring
  - Data certification

Detector/Data operations are not far from you. If we don't have qualify data, we will not have publications.



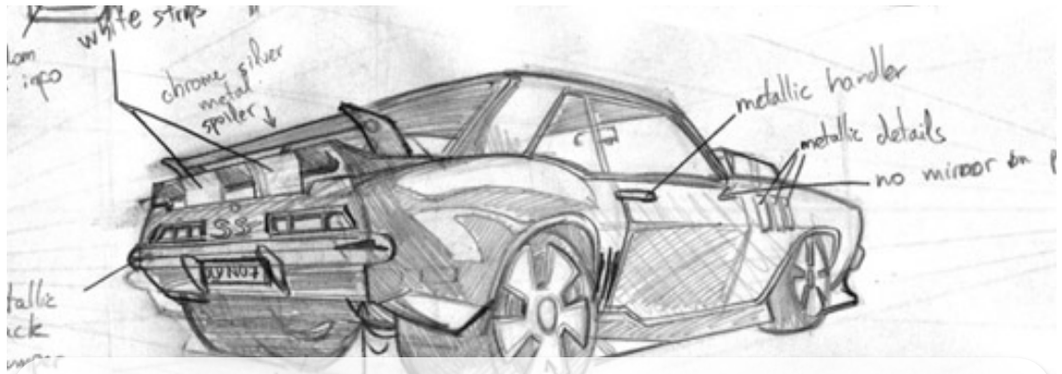
Monte-Carlo data

- From simulation to offline analysis
  - Simulation flow
  - Pileup mixing
- Trigger (another lecture)
- Reconstruction flow
- CMSSW
- Productions
  - Distributed computing
  - Search, Access, Transfer
  - Analysis

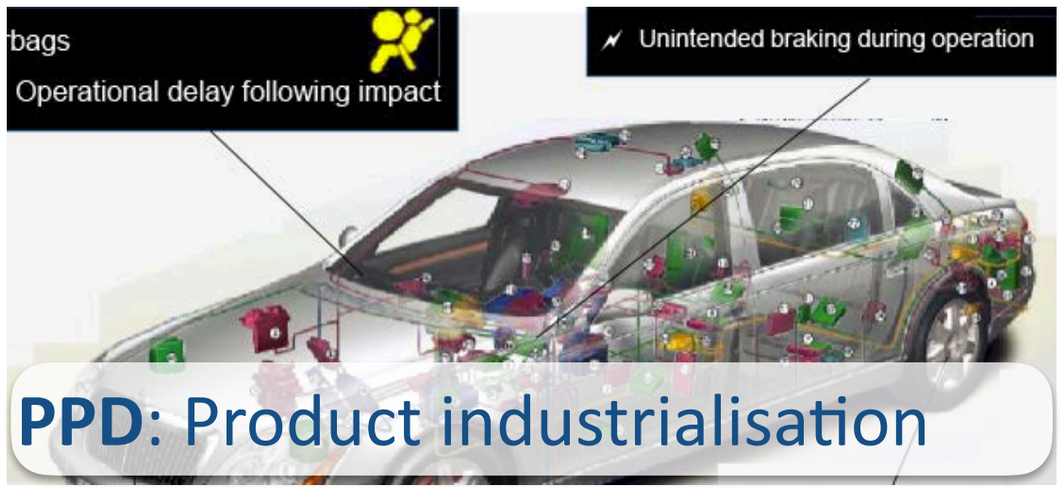
Data preparation and coordination areas

In this talk we focus on 2 main CMS coordination areas:

- **Offline & Computing (O&C)**
  - CMSSW software development, event reconstruction and simulation
  - data processing and Simulated events generation, events storage and management
- **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



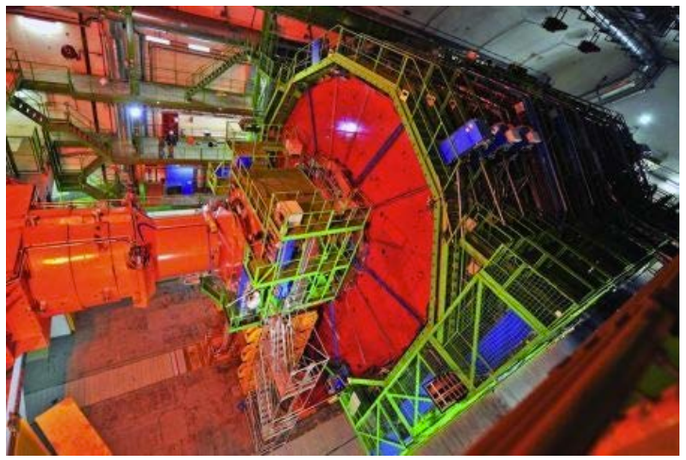
Offline: Design



PPD: Product industrialisation

Computing: Manufacturing

From P5 to offline



Events collected by CMS reach the Tier-0 at CERN for tape archival, organization, and processing.

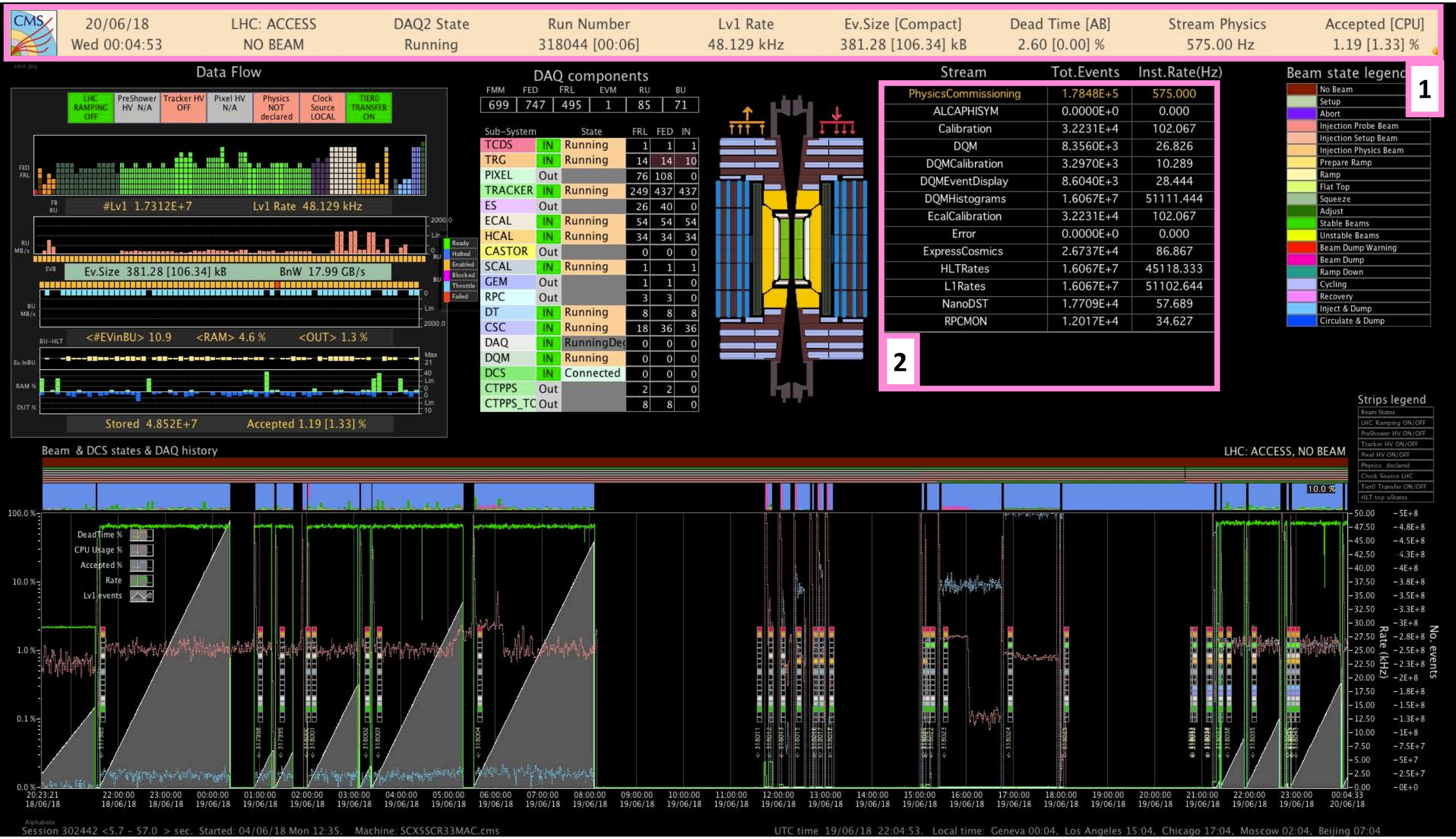
Data steams

- **Express**: Available ~2h after data collection for prompt feedback and calibration. About 100 Hz bandwidth shared by Calibrations, Detector, Physics monitoring.
- **Alignment & Calibration**: Dedicated event selection & event content devised for calibration process.
- **Physics**: Split into primary datasets and promptly reconstructed for physics analysis.
- **Other specialized streams**: Scouting, Parking.

Data rates

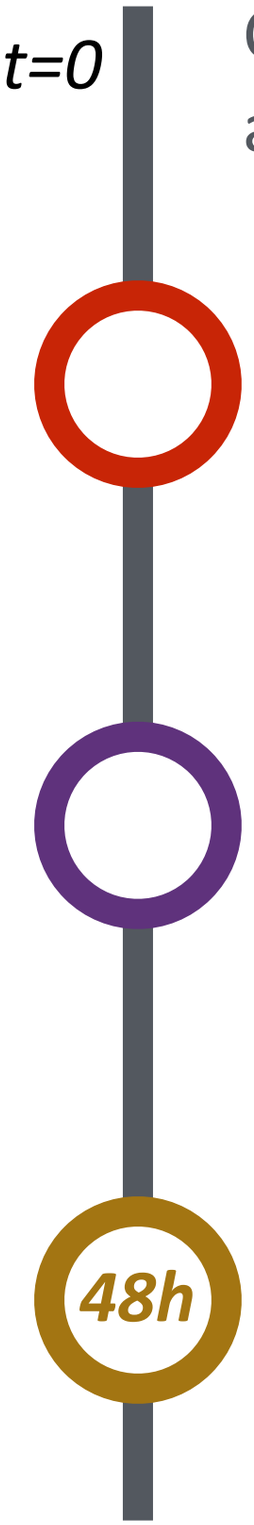
- **Run I**: 300 Hz Prompt-Reco + 300-600 Hz of parked data
- **Run II**: 1 kHz of Prompt-Reco + high rate of scouting data with reduced event content + parking

From P5 to offline



cmsonline.cern.ch (1) Basic DAQ, (2) Data streams and their rates

Prompt reconstruction and PCL at Tier-0



CMS aims at providing analysts with reconstructed data within days after data collections.

Express processing: Data reconstructed for

- Monitoring
- Calibration
- ECAL pedestal
- Pixel counting for luminosity

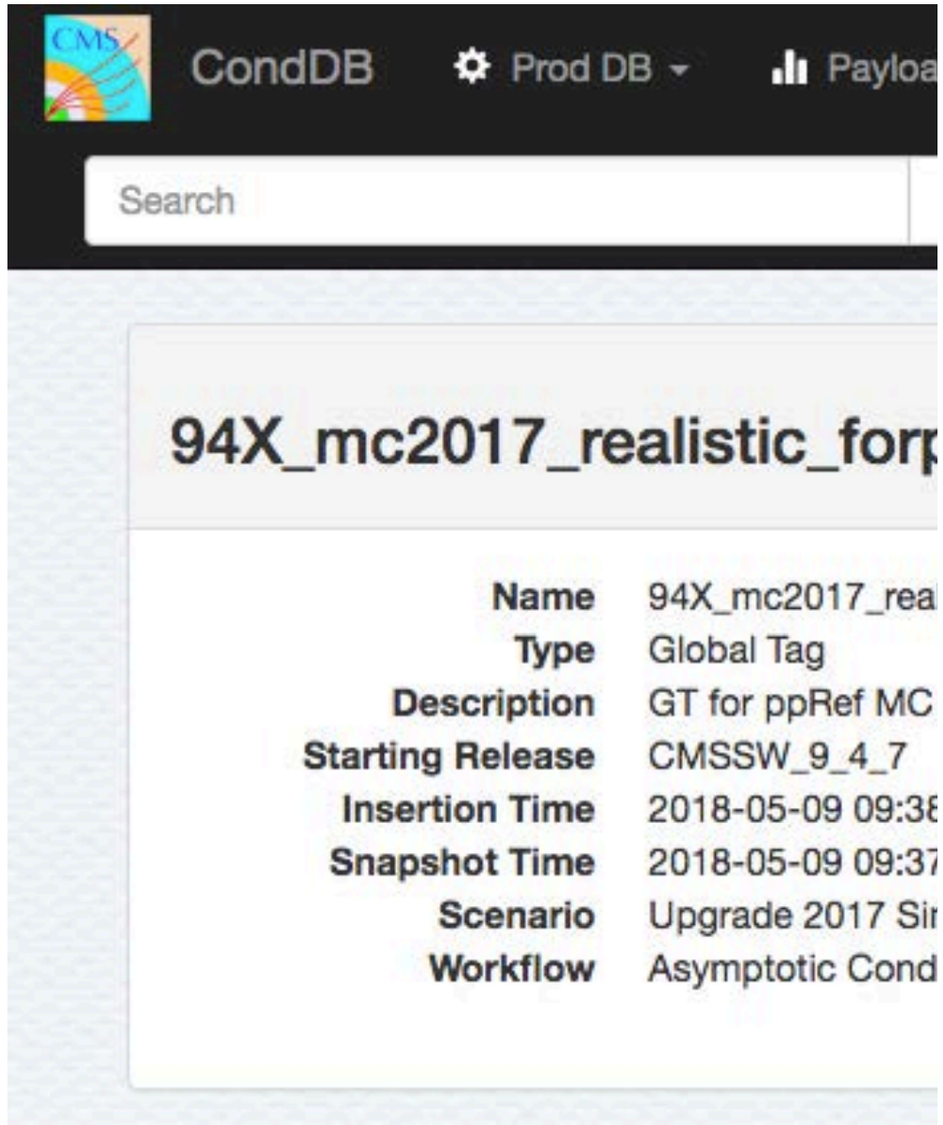
Prompt Calibration Loop (PCL): Express data is used as input to automated calibration workflows running at Tier-0 (or online)

- beamspot (LS by LS)
- ECAL transparency corr
- SiStrip bad-channels/gains
- SiPixel alignment

Prompt Reconstruction: Physics streams are reconstructed consuming calibrations computed by PCL. These are datasets for analysis. We normally start prompt reconstruction within 48 hour.

Calibration workflow

- Providing the most up-to-date alca conditions at all stages of the data and Monte Carlo processing is a major challenge involving all DPGs, POGs, AlCa team in PPD
- Critical to sustain the quick peace of analysis turn-around
  - Prompt-reconstruction must be of near-perfect quality;
- Few level of workflows depending on the time scale of updates
  - **Quasi-online calibrations** for HLT and express processing
  - **Prompt calibrations**: monitor and update conditions expected to vary run-by-run, or even more frequently - essential to guarantee performance of prompt reach
  - **Offline workflows** for data reprocessing and analysis level conditions



Primary datasets

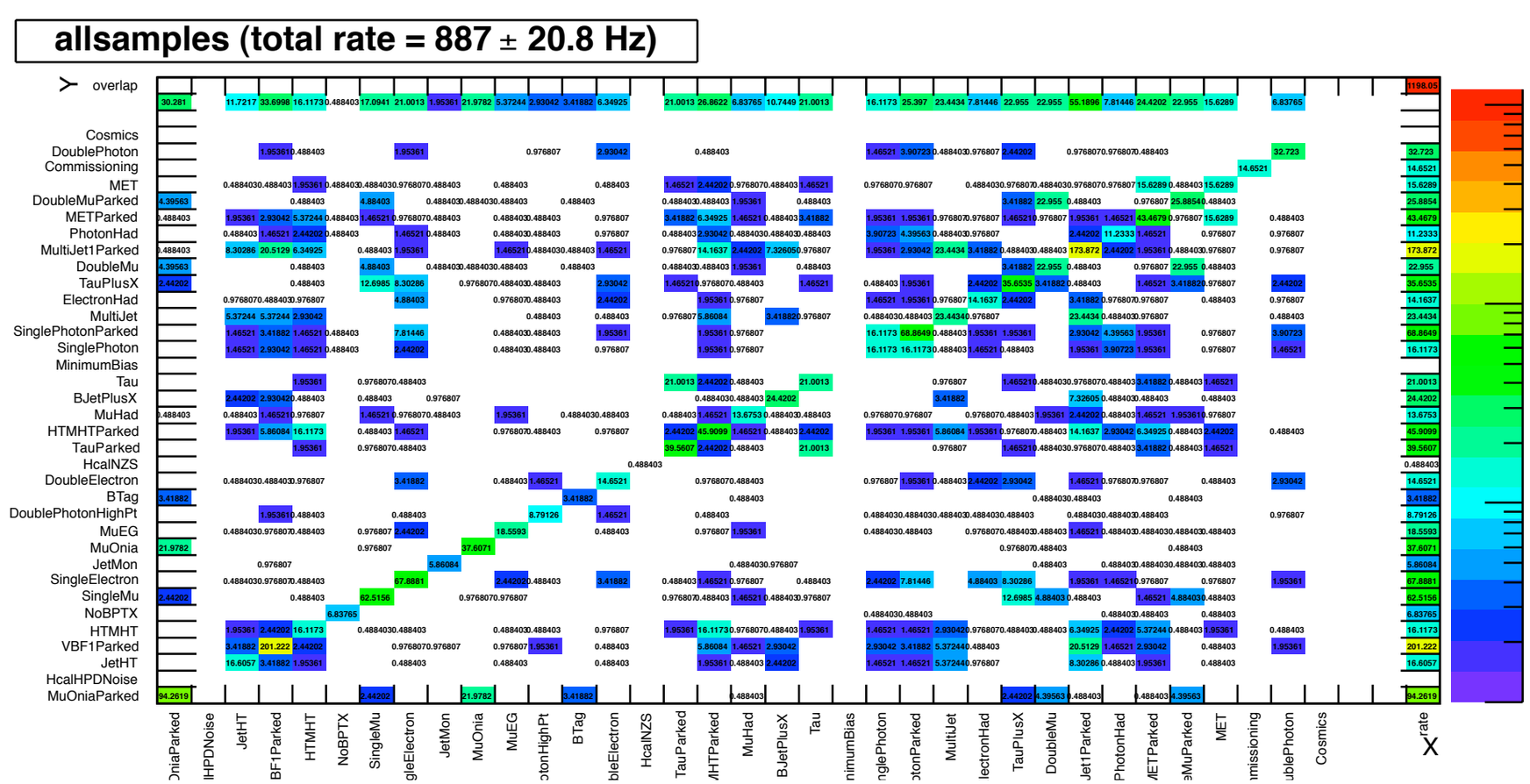
The physics steams from P5 are split to Primary Datasets (PD) on the basis of HLT results in order to group events with related topology and limit replication of events (PD's overlap).

Constraints

- Physics: Definition centered on physics objects (i.e. SingleElectron, JetMET)
- Processing & Handling: Proper event rates for each PD, to be able to handle by Tier-2.

On top of the primary datasets we can deploy "central skims" to

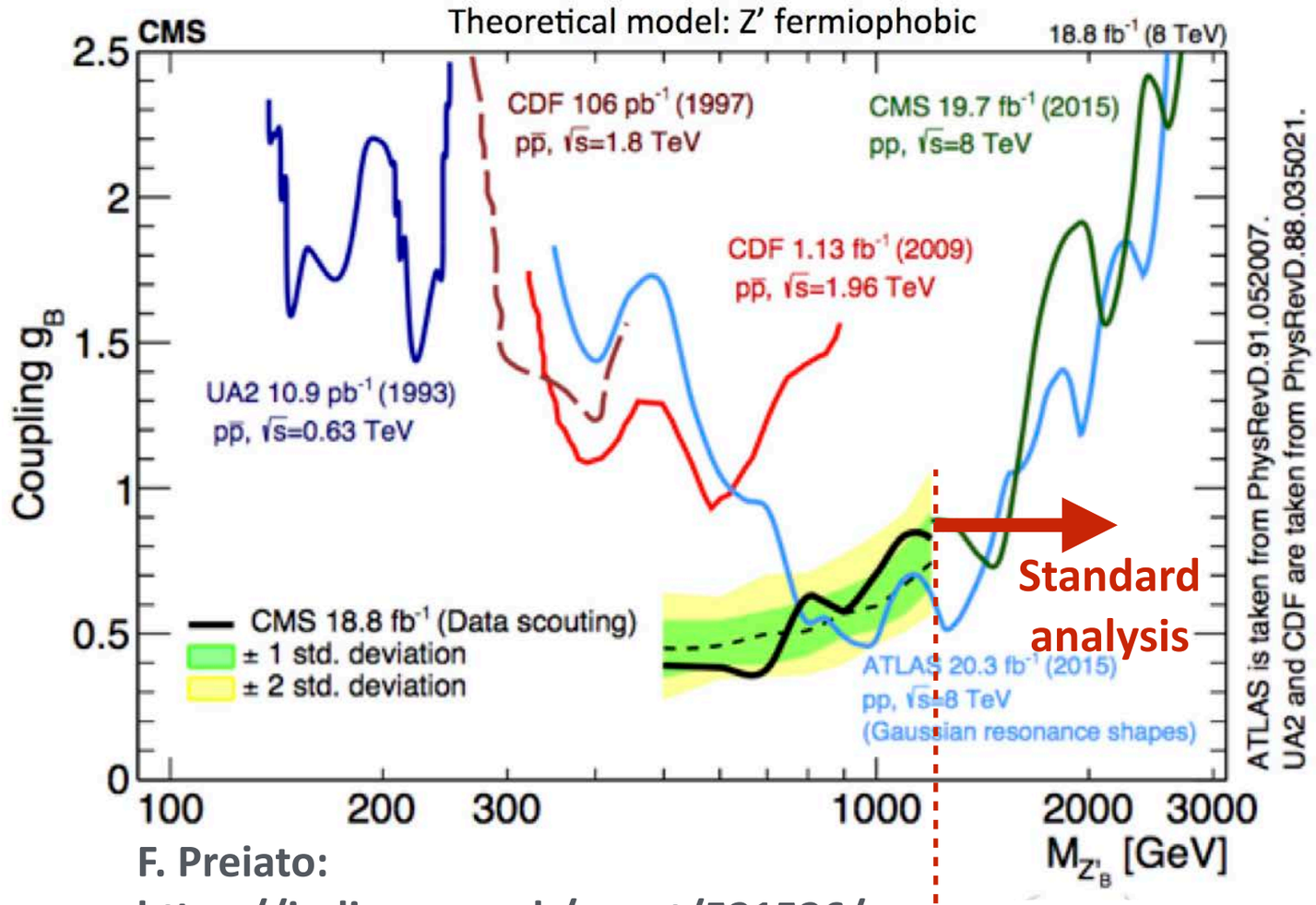
- customise event content
- Reduce rate using also RECO quantities.



Data scouting and data parking

As LHC luminosity rises, increasing trigger thresholds pose a challenge for analyses in CMS

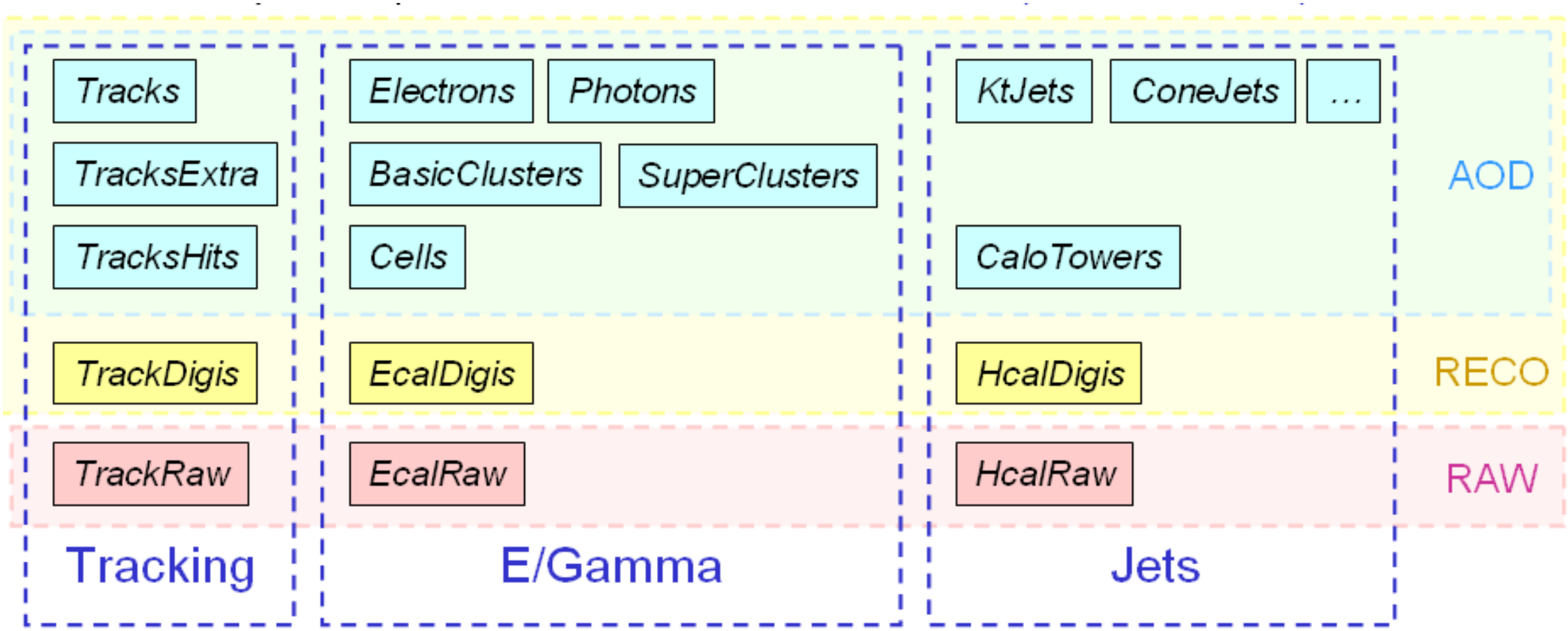
- Trigger rates are constrained by the CMS prompt reconstruction system, which cannot process much more than 1 kHz of events.
- Getting around this limit:
  - ☑ **Data parking**: send events from the HLT to tape without reconstructing them
  - ☑ **Data scouting**: save only a small subset of the event content (e.g., only the HLT-level jet objects)



F. Preiato: https://indico.cern.ch/event/521526/

Event contents and data tiers

- Event information from each step in the simulation and reconstruction chain is logically grouped into what we call a ***data tier***.
- Examples of what are in RAW/RECO/AOD

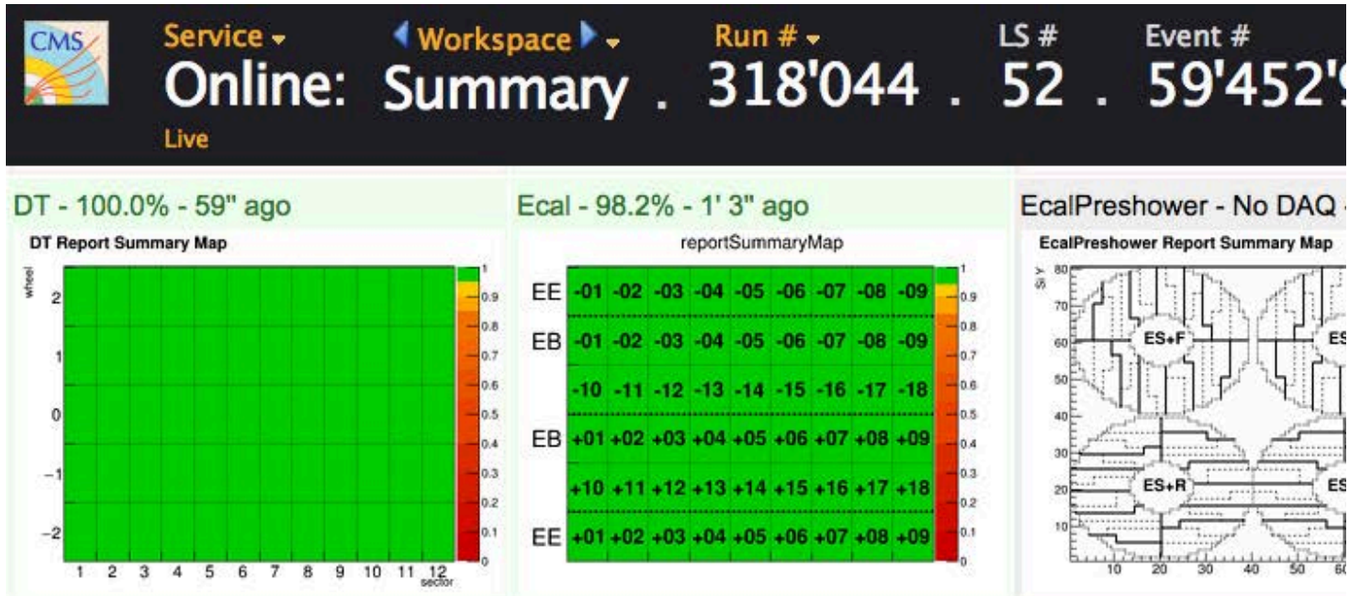


- Currently, two step further in data reduction:
  - ☑ **MiniAOD:** ~10-15% of AOD size, designed for 90% analyses at CMS.
  - ☑ **NanoAOD**: design for analysis with bare root, cover 30-50% of analyses.

Data Quality Monitoring (DQM)

DQM is the tool to produce plots while running RECO (or any CMSSW workflow). There are two main areas of DQM 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
  - ☑ Validation+verification



https://cmsweb.cern.ch/dqm/online
https://cmsweb.cern.ch/dqm/offline

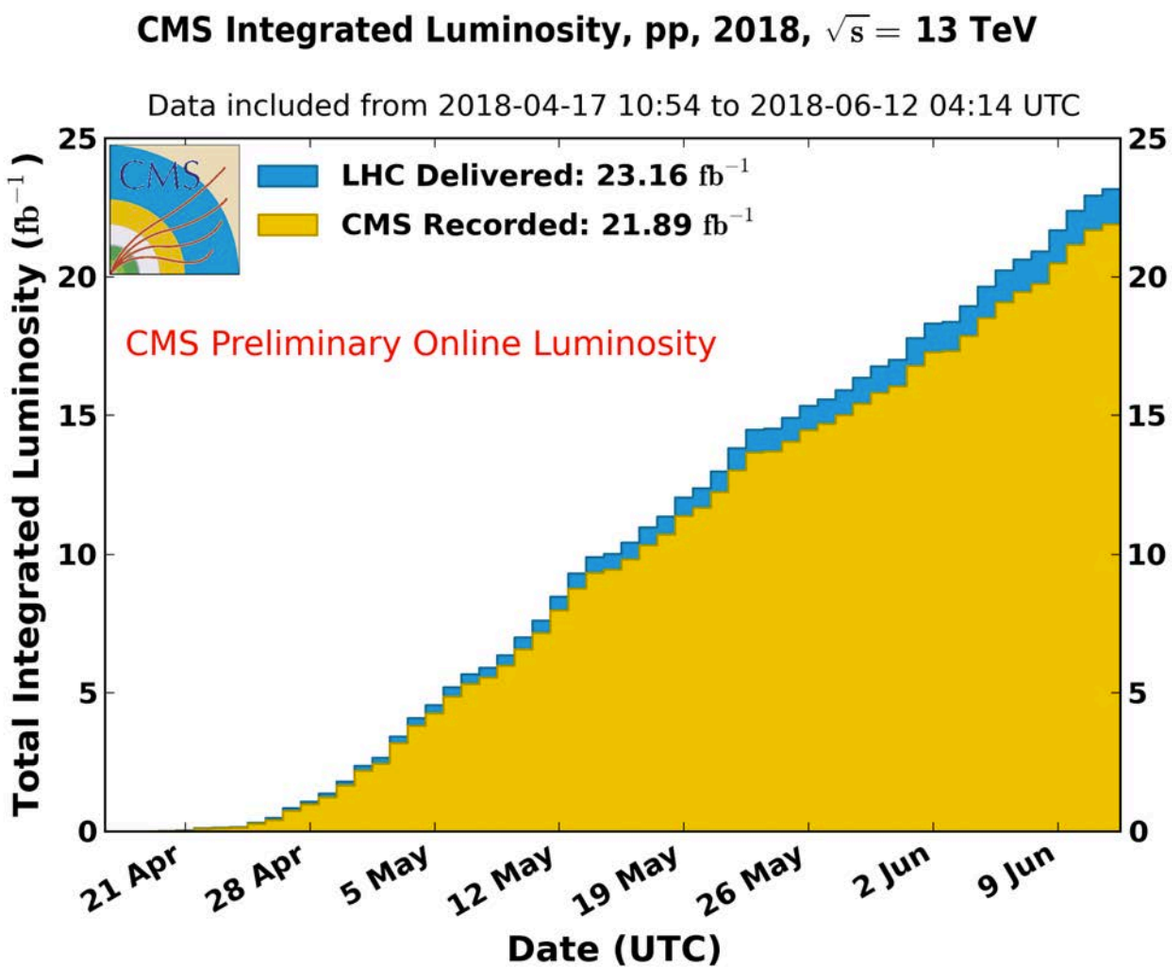
Data Certification (DC)

We run analysis only LumiSections (Sub-section of a run during which time the instantaneous luminosity is unchanging, ~23s of RUN) which we consider to be good. This task is done by team of experts in detector and physics objects.

- **Golden**: require all sub-detectors/POGs to be “GOOD”
- **Muon-only**: no requirements on calorimeters
- **DCS-only**: require only tracker to be powered

The format of file is in JSON format

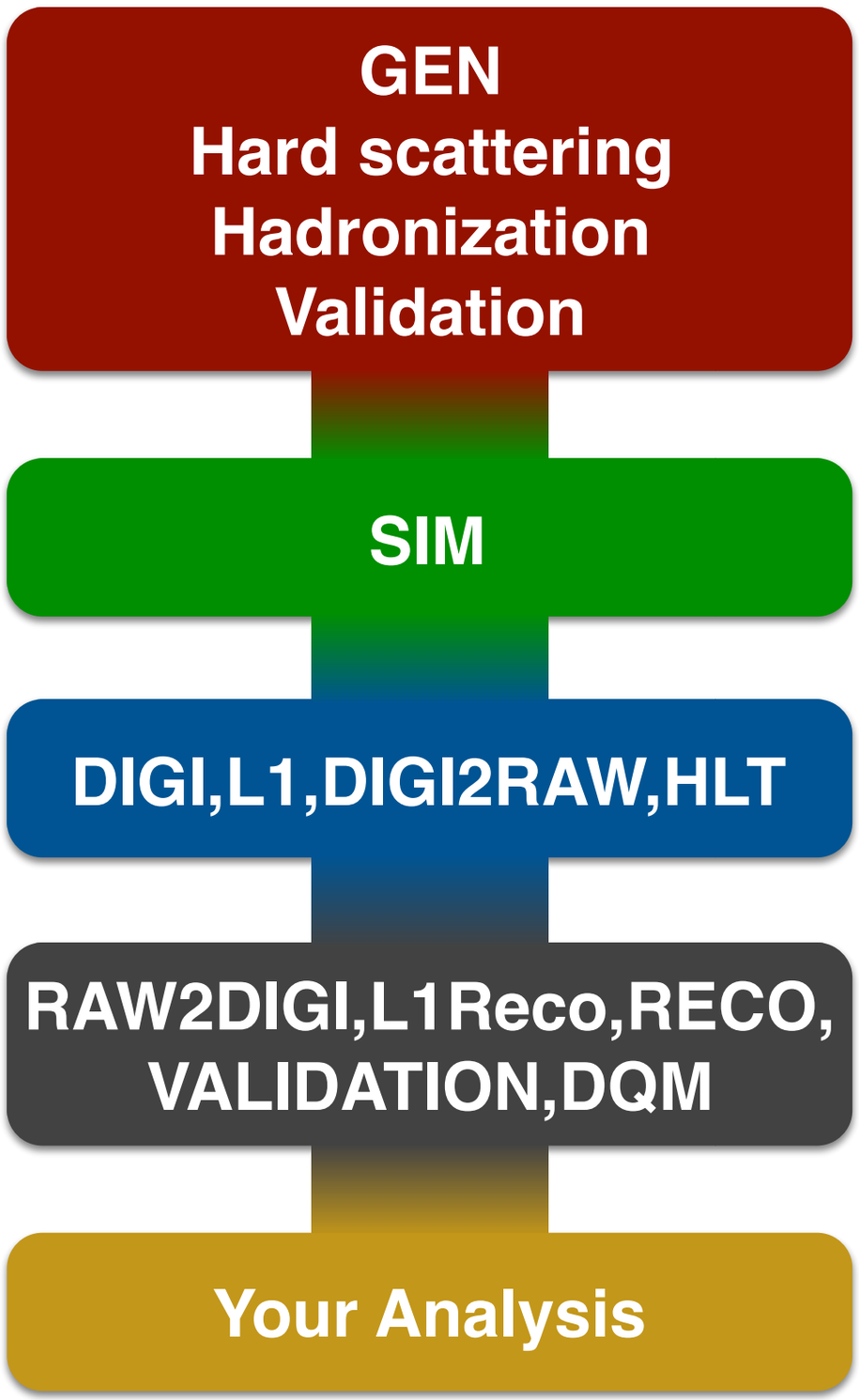
- Weekly for PromptReco
- After each major reprocessing




```
{"273158": [[1, 1279]], "273302": [[1, 45
"273408": [[1, 6]], "273409": [[1, 309]],
113], [115, 412]], "273448": [[1, 391]],
"273493": [[1, 233]], "273494": [[1, 192]
[[1, 173]], "273725": [[83, 252], [254, 2
"274159": [[1, 43]], "274160": [[1, 207]]
"274240": [[1, 40], [42, 82]], "274241":
```

Run no. **Range of LumiSections**

From collision/simulation to physics analysis



What physics are we studying?
What generator should be used?
Which parameters should be modified?

CMS Geometry, Magnetic field,...

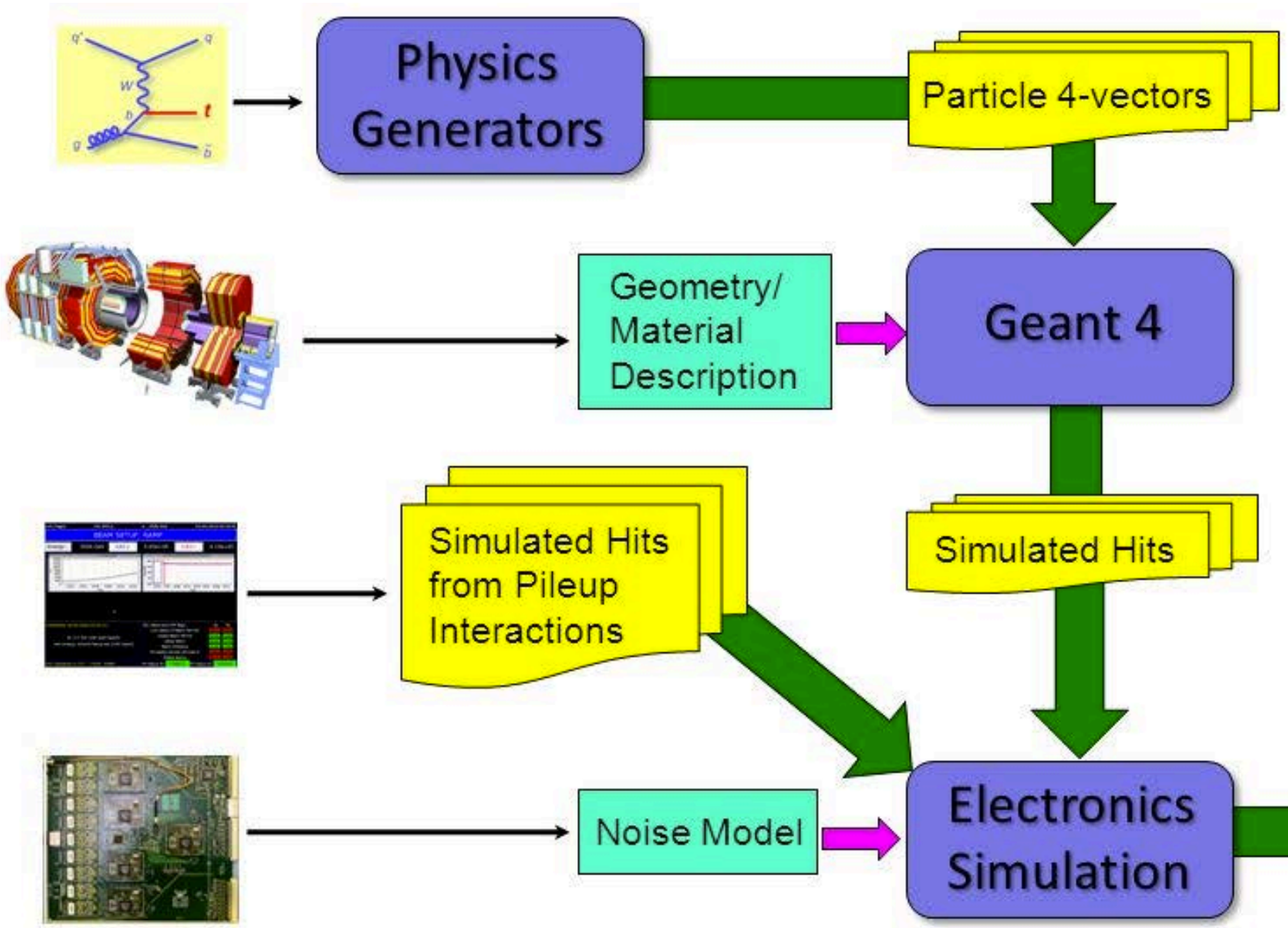
Detector electronics simulation, Pileup situation,
Alignment-Calibration, Trigger menu,...

Reconstruction algorithms,
High level object creation

Your analysis code

Simulation framework

- Event simulation algorithms are implemented as module, communicating via the Event.
- The simulation sequence aims at producing MC truth and RAWDATA as it comes from point 5.



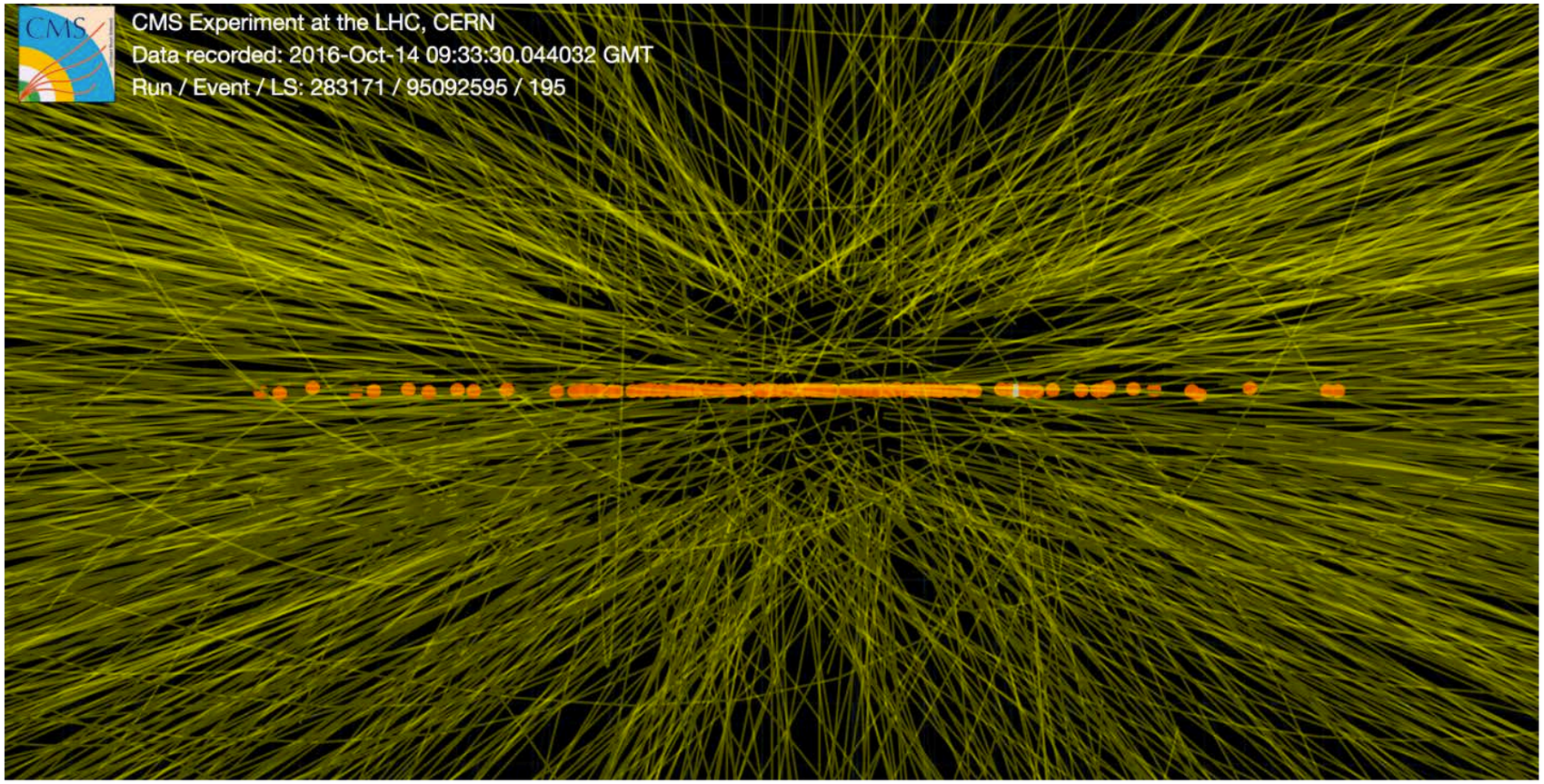
Data tiers

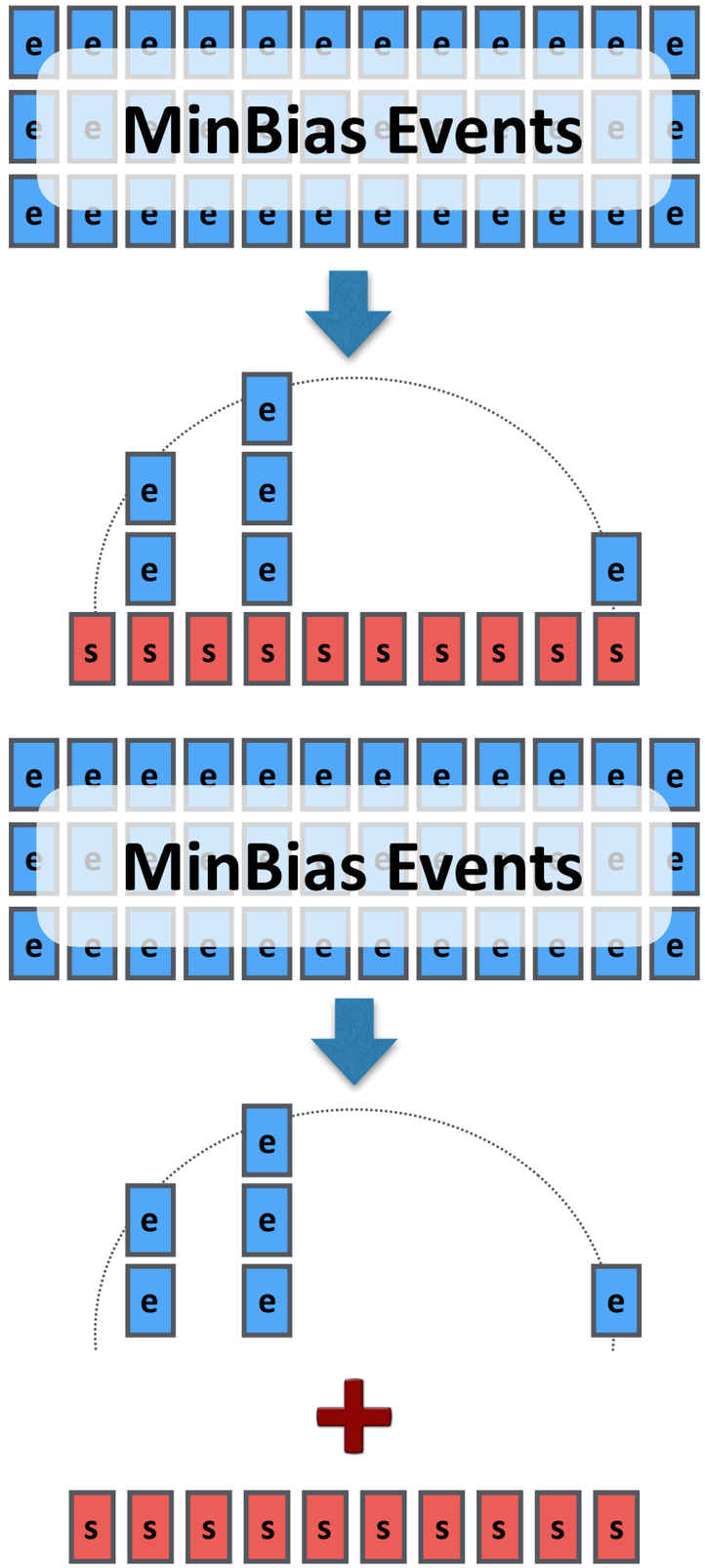
GEN, LHE

GEN-SIM

GEN-SIM-RAW

Pileup





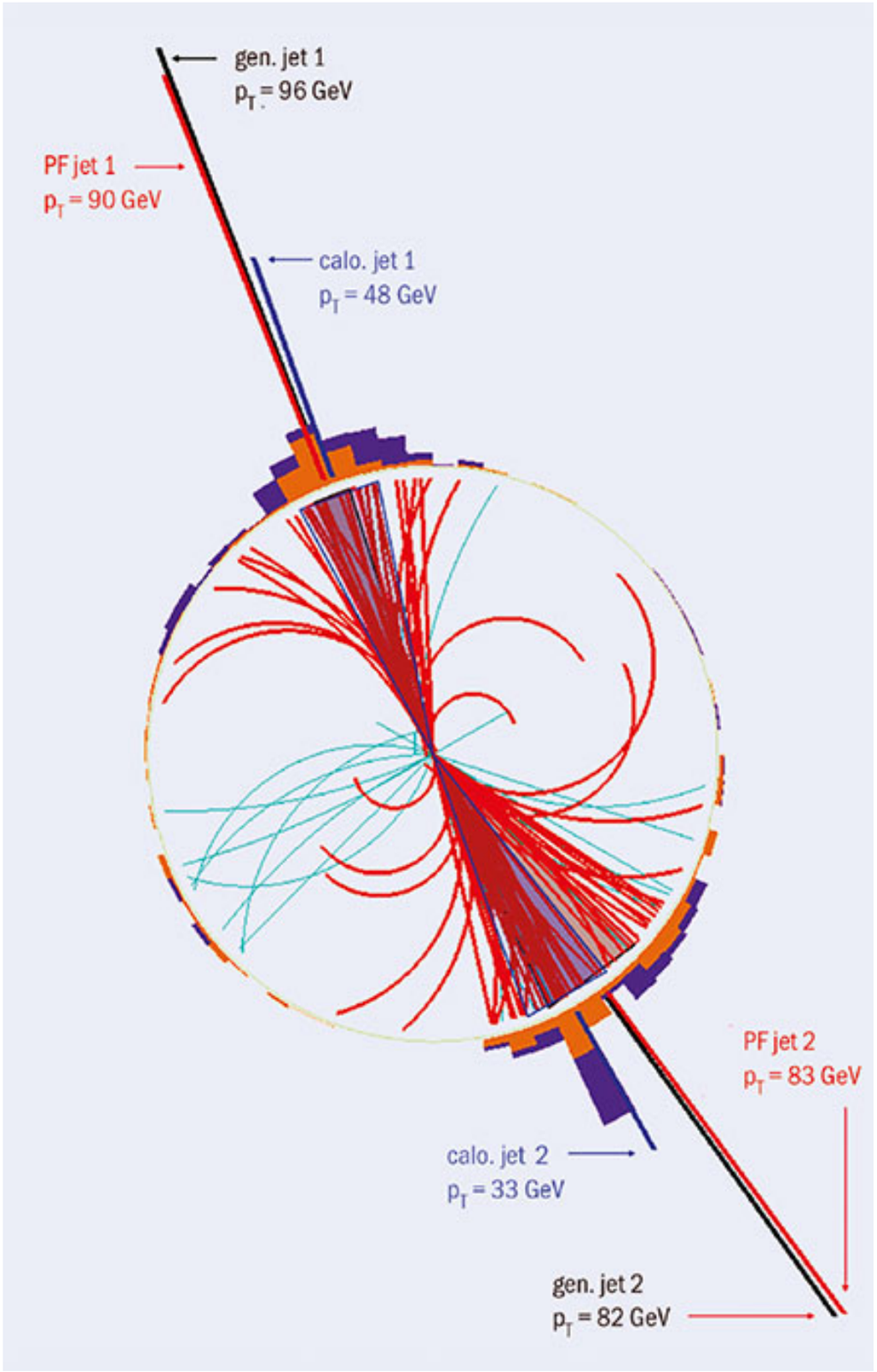
Classic mixing

- GENSIM Signal (MC Hard-scatter event) is overlaid with GENSIM MinBias with chosen pileup configuration.

Pre-mixing

- MinBias events in RAWSIM format are overlaid on empty single neutrino events using a chosen pileup configuration. Digis made in this step are converted to RAW.
- 1-1 combination of PreMixed event - signal event. RawToDigi is done on-the-fly to premixed events before overlay.

Reconstruction



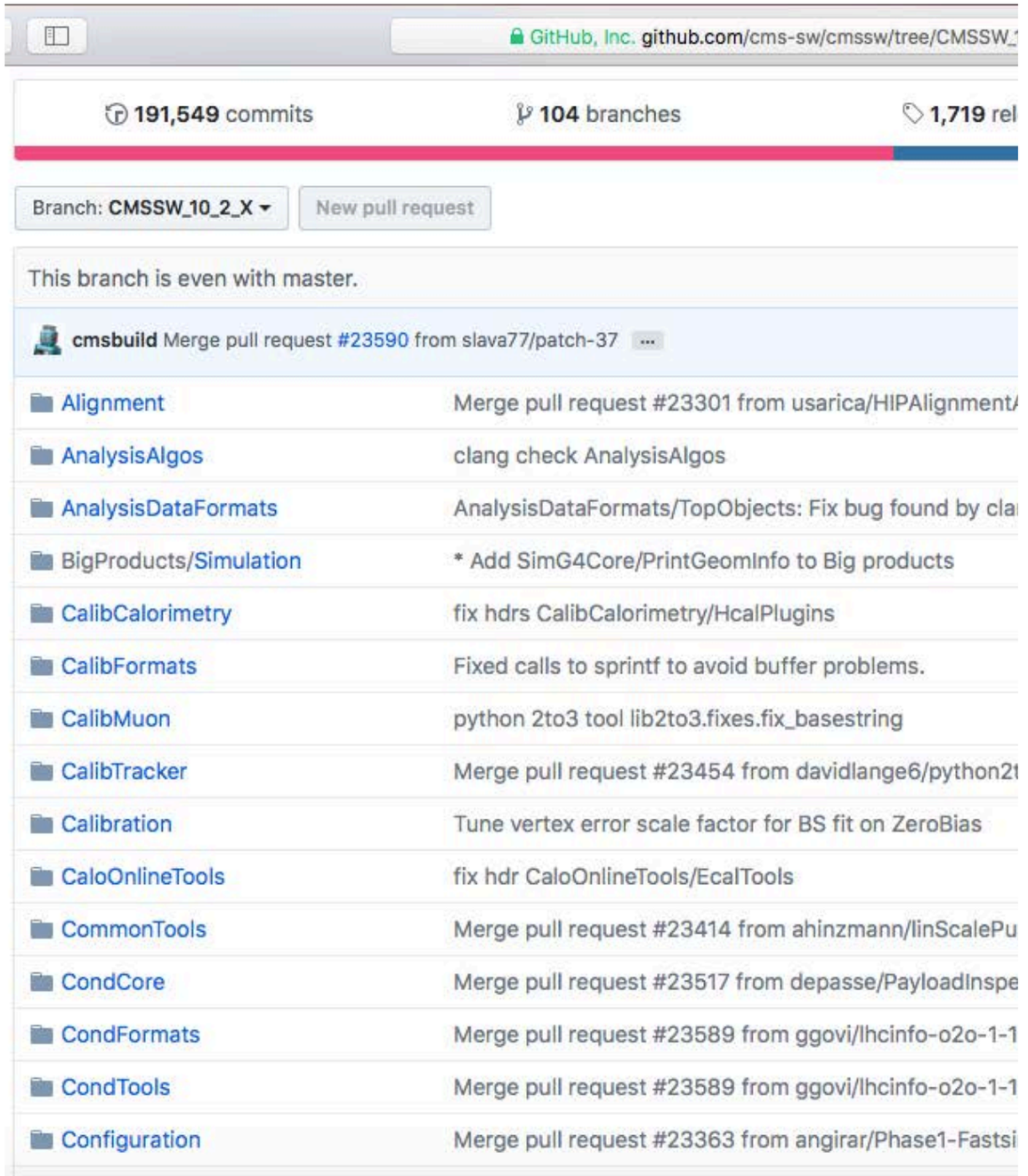
- The reconstruction sequence turns the binary output (RAW) from CMS/DIGI into physically interpretable quantities ready for data analysis
- Hits in the detector are aggregated in cluster and tracks, which in turn are matched to create particle candidates (Particle Flow): Tracks, muons, electrons, photon, jet, …
- Data tiers include
  - RECO
  - AOD
  - MINIAOD
  - NANOAOD

Data Tier	Size (kB)
RAW	1000
GEN	< 50
SIM	1000
DIGI	3000
RECO(SIM)	3000
AOD(SIM)	400 (8x reduction)
MINIAOD(SIM)	50 (8x reduction)
NANOAOD(SIM)	1 (50x reduction)

Analysis data formats

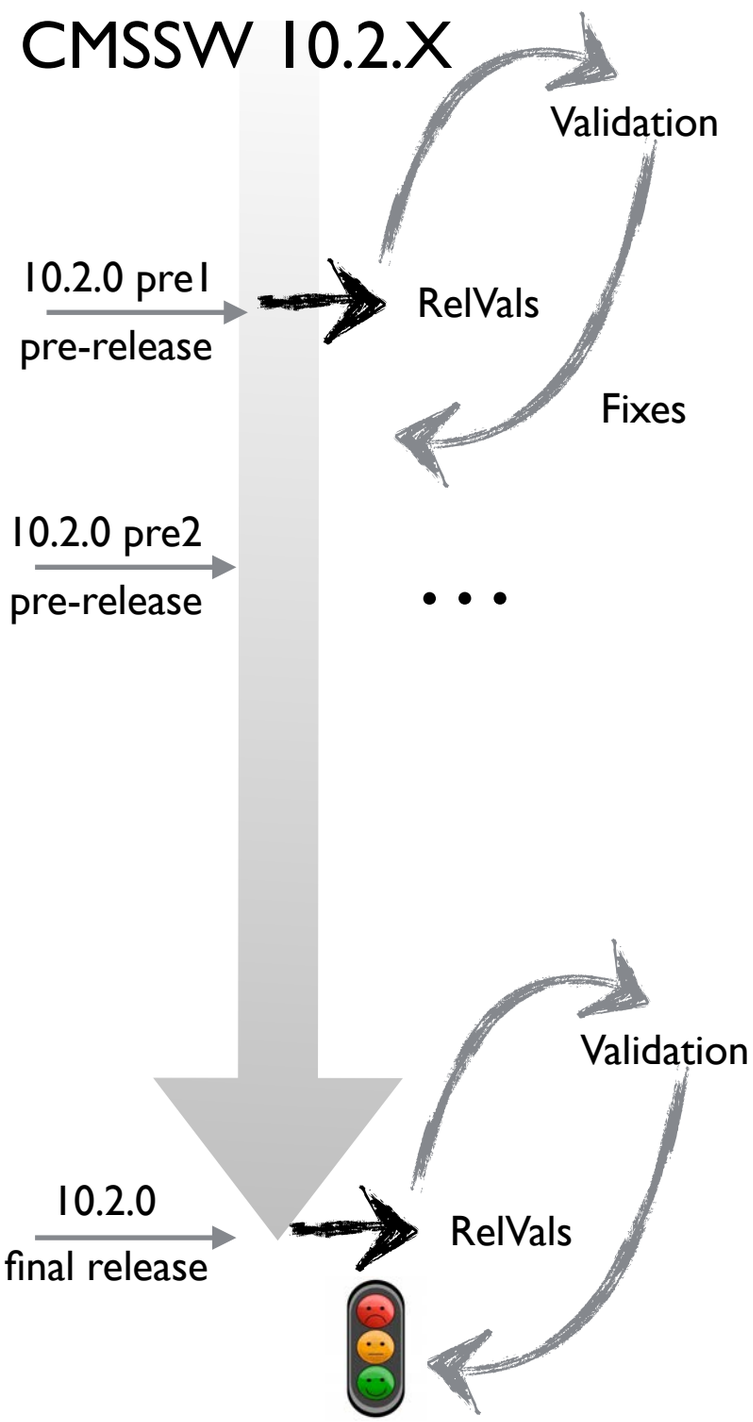
CMSSW

- CMSSW: one release to rule them all
  - ☑ GENerator, SIMulation, HLT, RECOnstruction ANALYSIS workflows...
- C++ code and configuration handled via Python
  - ☑ “git” used for code versioning and integration
- Release schedule follows a “train model”: dear developer: catch this train or wait for the next one
  - ☑ regular timetable of ~6 months (slightly tuned for major conferences or physics needs)
  - ☑ pre-releases are regularly produced while the release is under development



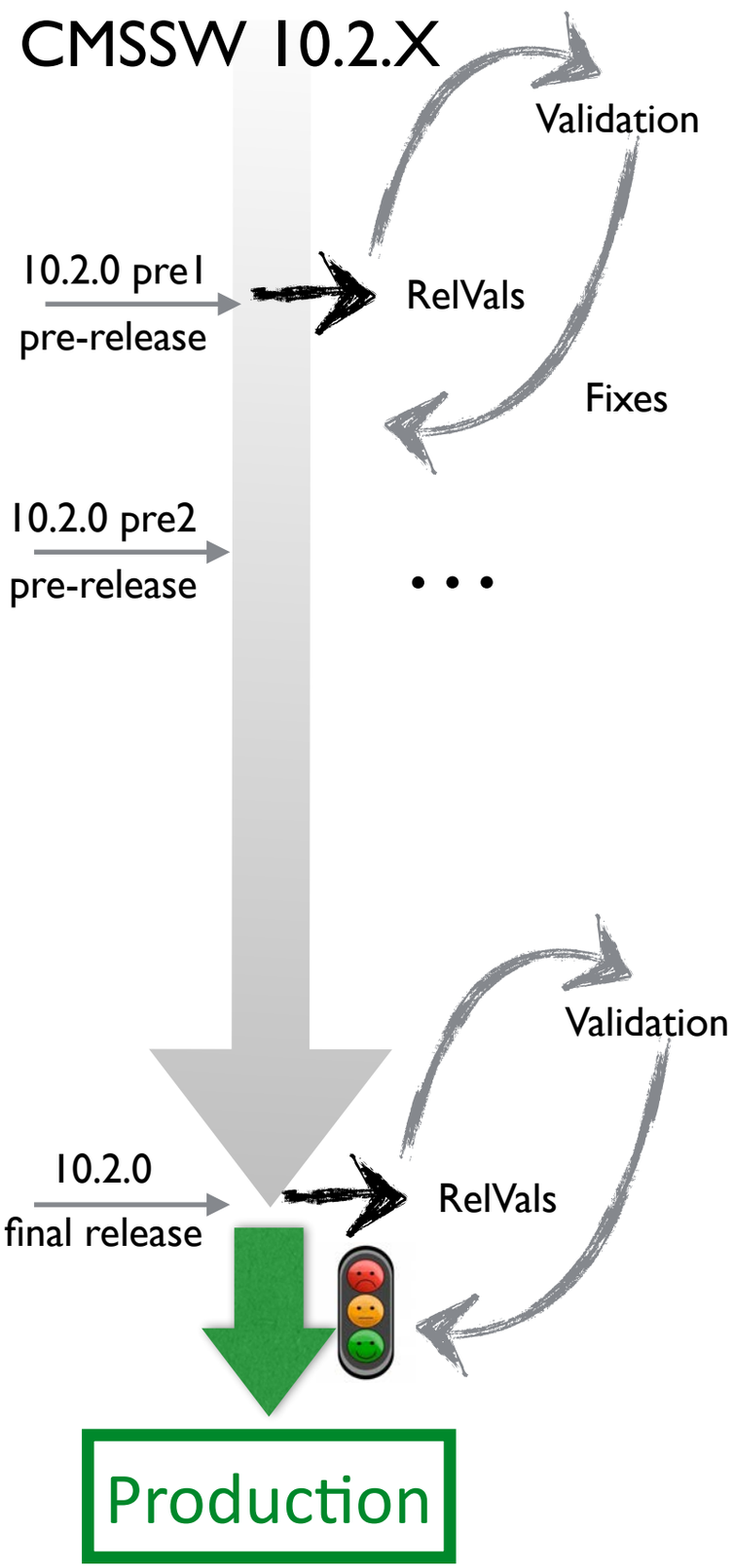
Start from http://cms-sw.github.io/

CMSSW release validation



- Release integration bound to Quality Assurance Tests
  ⟶ Data Quality Monitoring (DQM)
  - ☑ unit tests & regression tests
  - ☑ small scale production tests: Release Validation Test (RelVal) producing DQM plots
- Validation: iterative process performed all along release cycle (from pre-releases to final version)
  - ☑ DPG, POG and PAG validation experts check the plots
  - ☑ PPD/PdmV group coordinates validation campaigns
  - ☑ sign-off on quality of release and calibrations
  - differential validation compares plots of each release w.r.t reference (ex. last validated pre-release)
  - ☑ each cycle takes 4-5 days to have the samples + 1 week for feedback from the validators

CMSSW release validation & production preparation



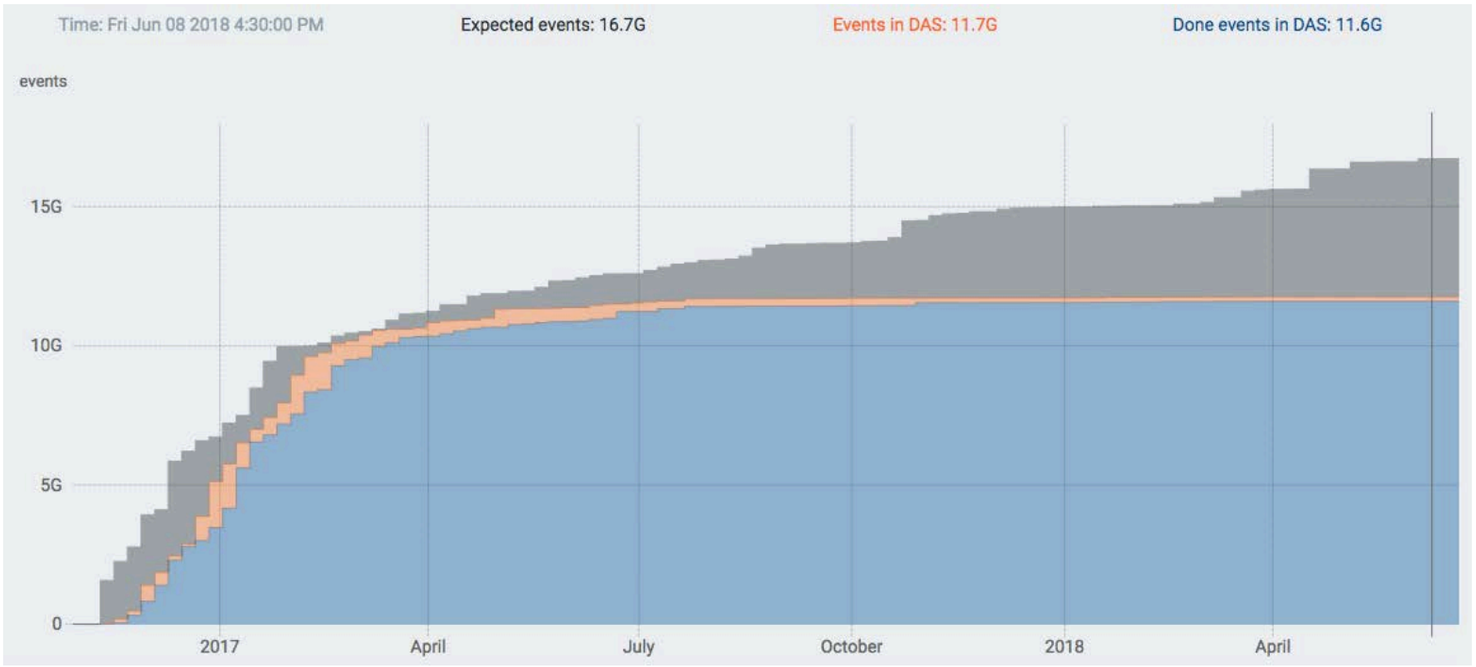
- Once a major release (X.Y.Z) is green-lighted
  ⟶ start preparation of the campaign
  (re-reco or MC production)
  - ☑ finalization of the alignment and calibration conditions (and their validation)
  - ☑ finalization of the parameters for the Pile-Up overlay (PU scenario)
  - ☑ preparation of the injection machinery for the central processing by computing

CMSSW & Production

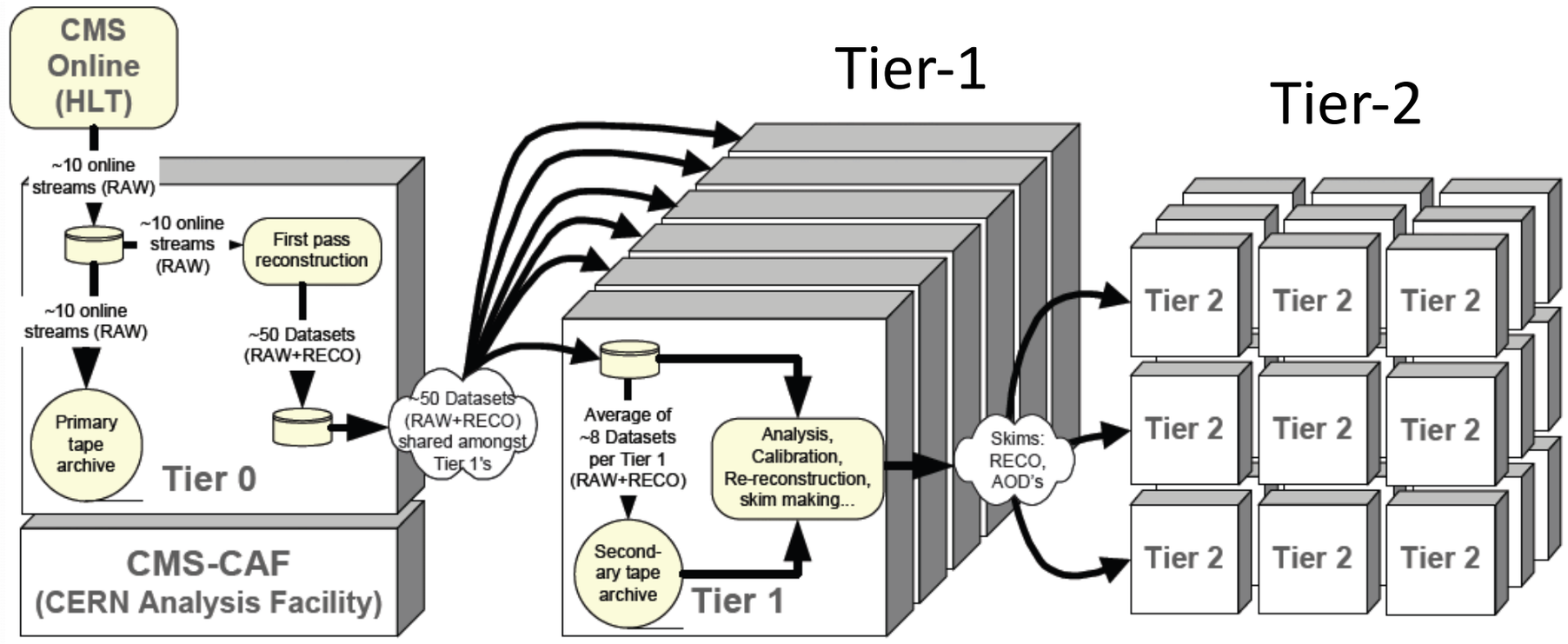
- Feature planning:
  - ☑ production releases: driven by physics/machine constraints & goals
  - ☑ e.g. RunIISummer16DR80Premix production is producing MC compatible to 2016 data

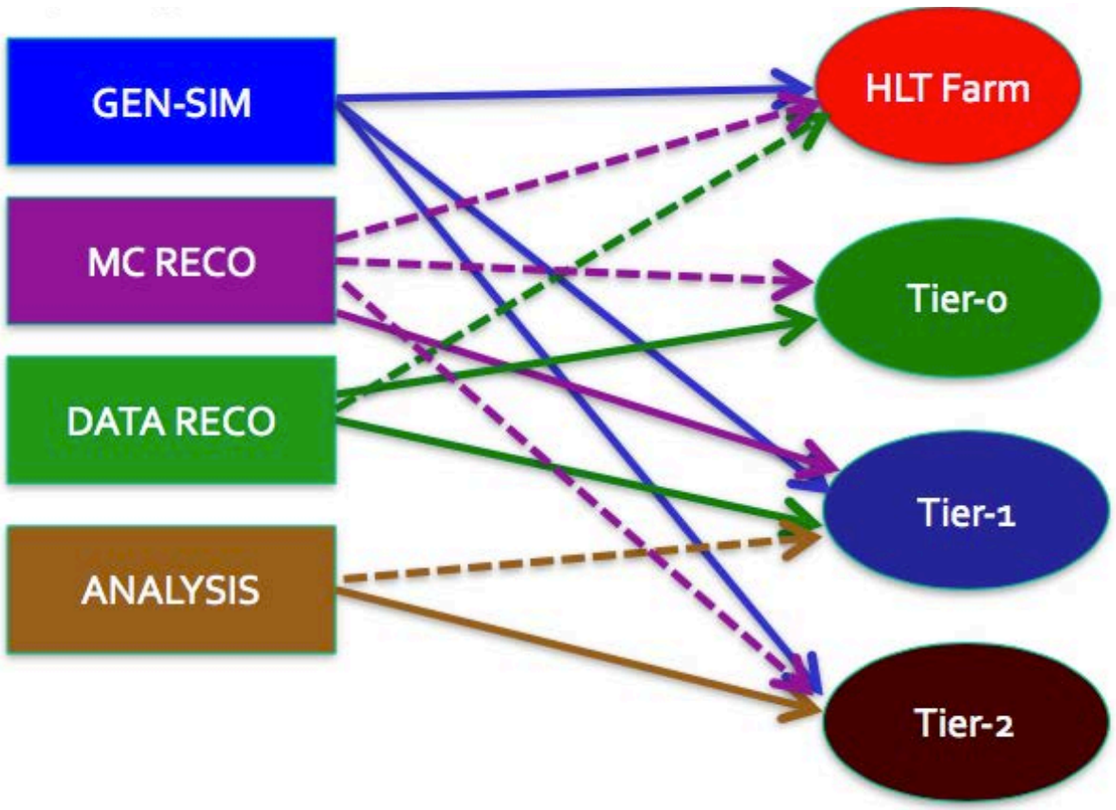
23Sep2016 re-reco data: 4.9G events

RunIISummer16DR80Premix: 11.7G events



Distributed computing





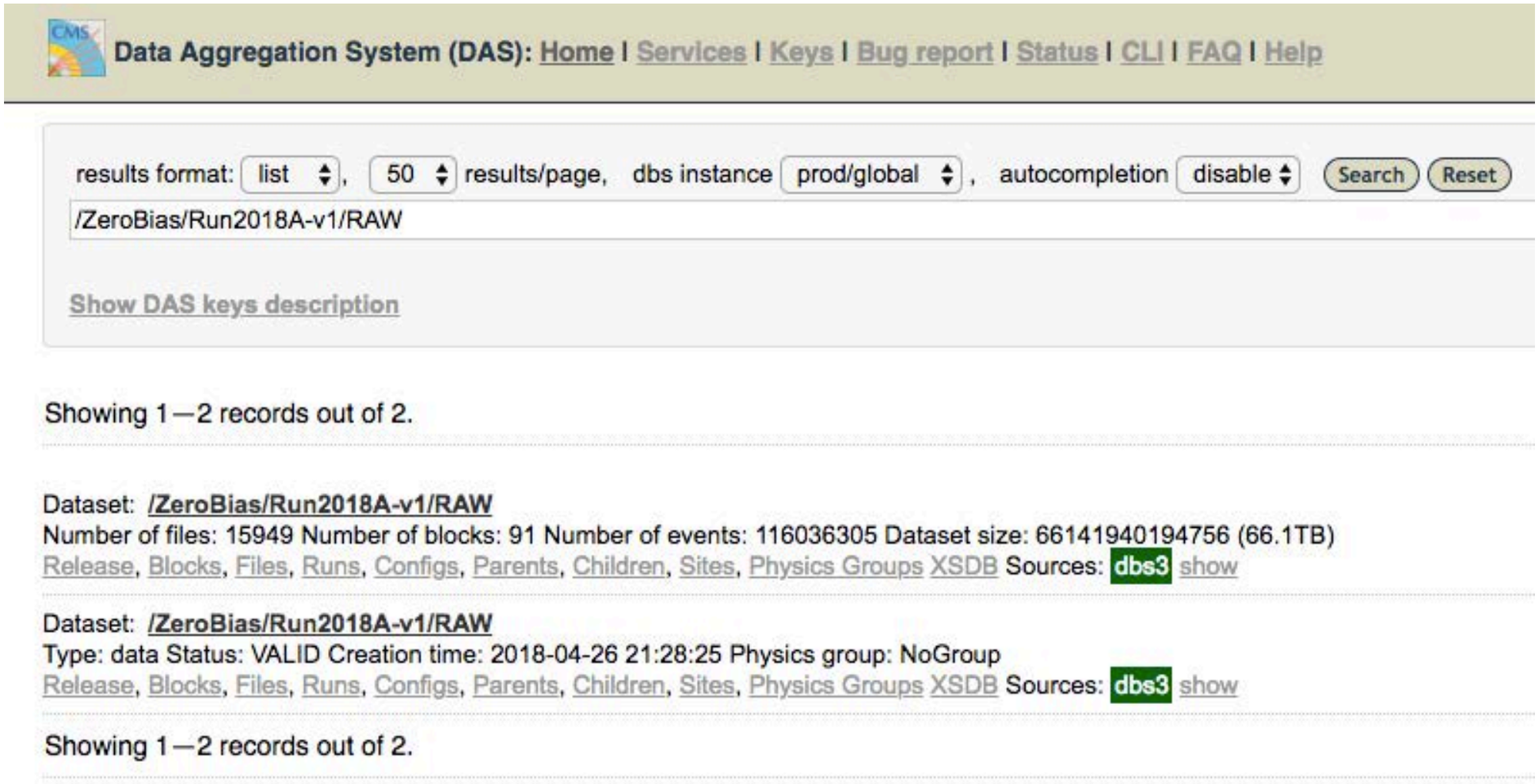
- **Increasing the flexibility for facilities and workflows**
  - ☑ More places that jobs can run

Run-1 - - - - Run-2

Search, access, transfer datasets

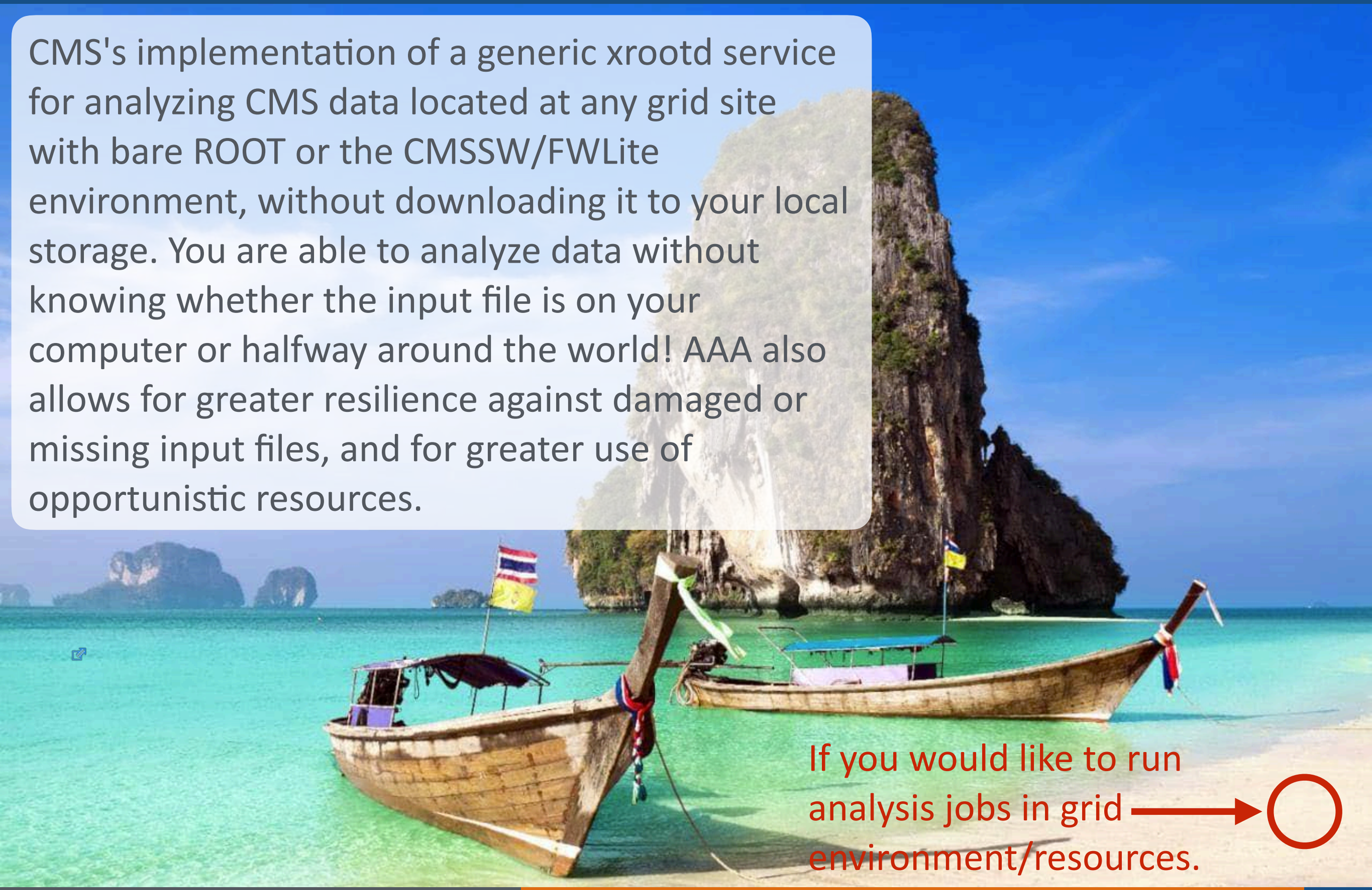
Data Aggregation Service (DAS) is the place where you can look for samples.

- https://cmsweb.cern.ch/das/
  - ☑ Certificate is needed (CERN, or mapping to CERN account)
- DAS lists datasets and their properties (requestID, sites, run # and LS #....) aggregating information from various services.
- Dataset name structure is mentioned in the backup.
- PhEDEx service will help to transfer datasets between sites (backup).



Analysis: Any Data, Anytime, Anywhere (AAA)

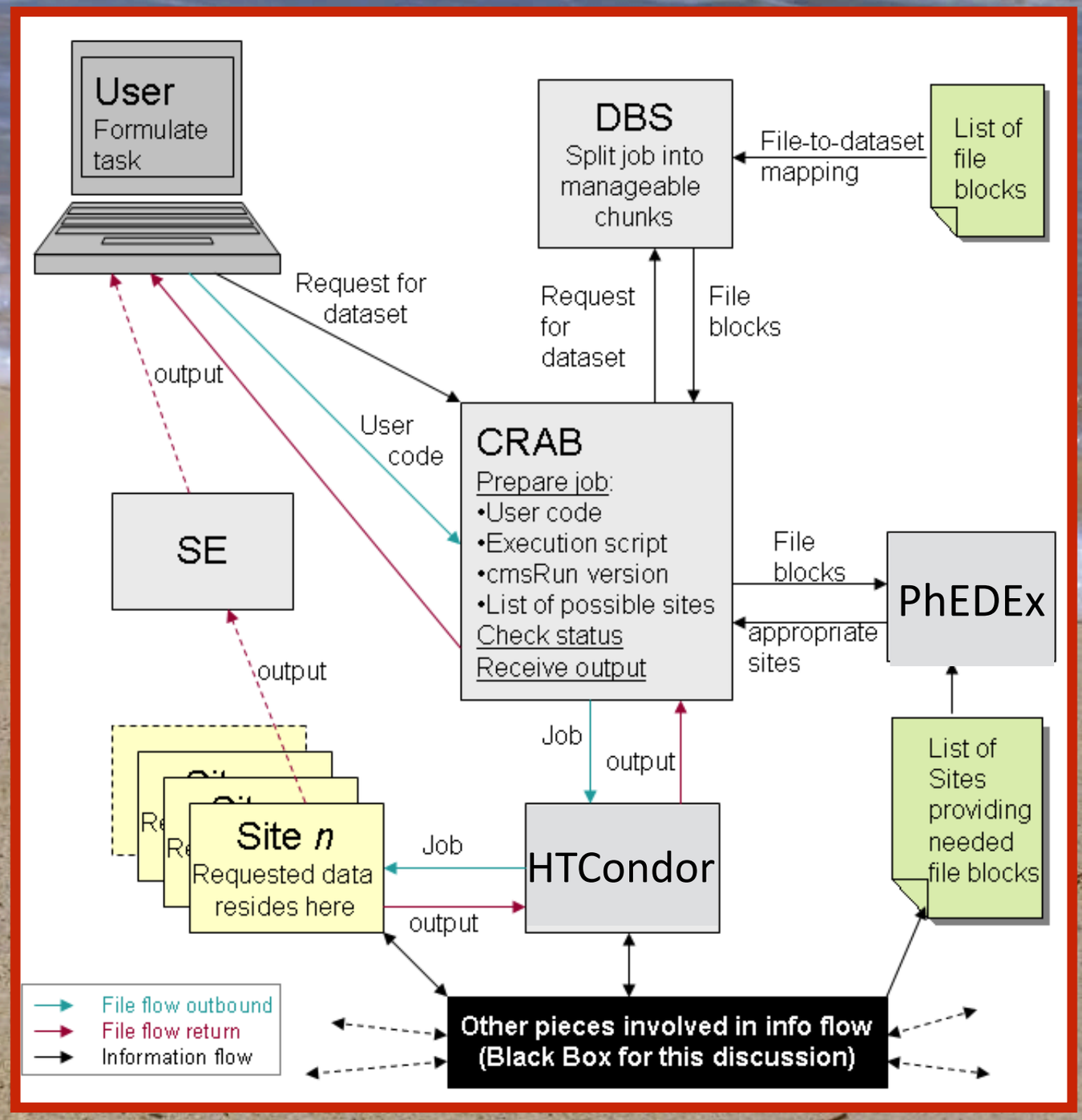
CMS's implementation of a generic xrootd service for analyzing CMS data located at any grid site with bare ROOT or the CMSSW/FWLite environment, without downloading it to your local storage. You are able to analyze data without knowing whether the input file is on your computer or halfway around the world! AAA also allows for greater resilience against damaged or missing input files, and for greater use of opportunistic resources.



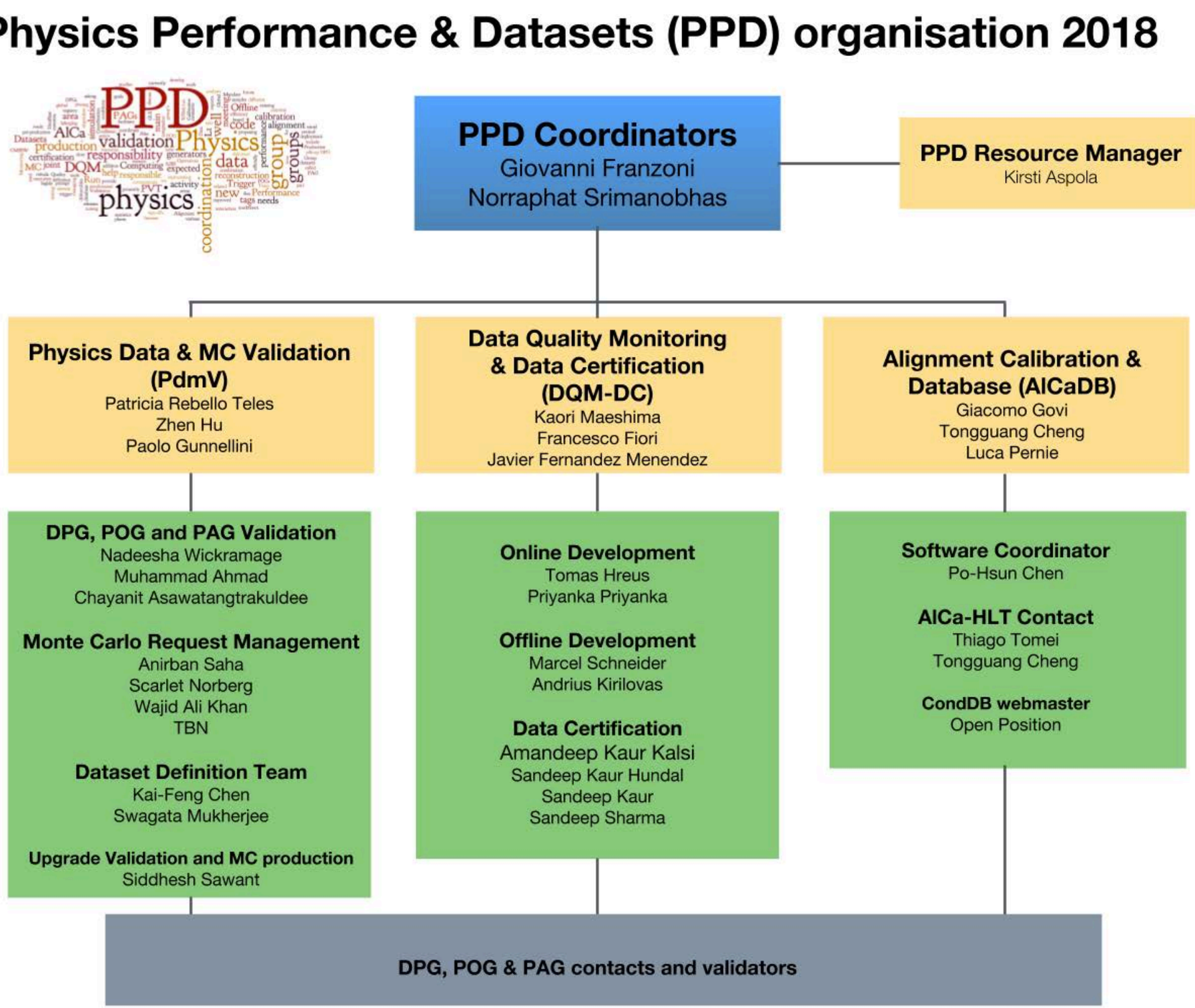
CRAB

CRAB is a Python program intended to simplify the process of creation and submission of CMS analysis jobs into a grid environment.

If datasets you want to use are not on disk everywhere, they will be transfer from tape to disk for you automatically.

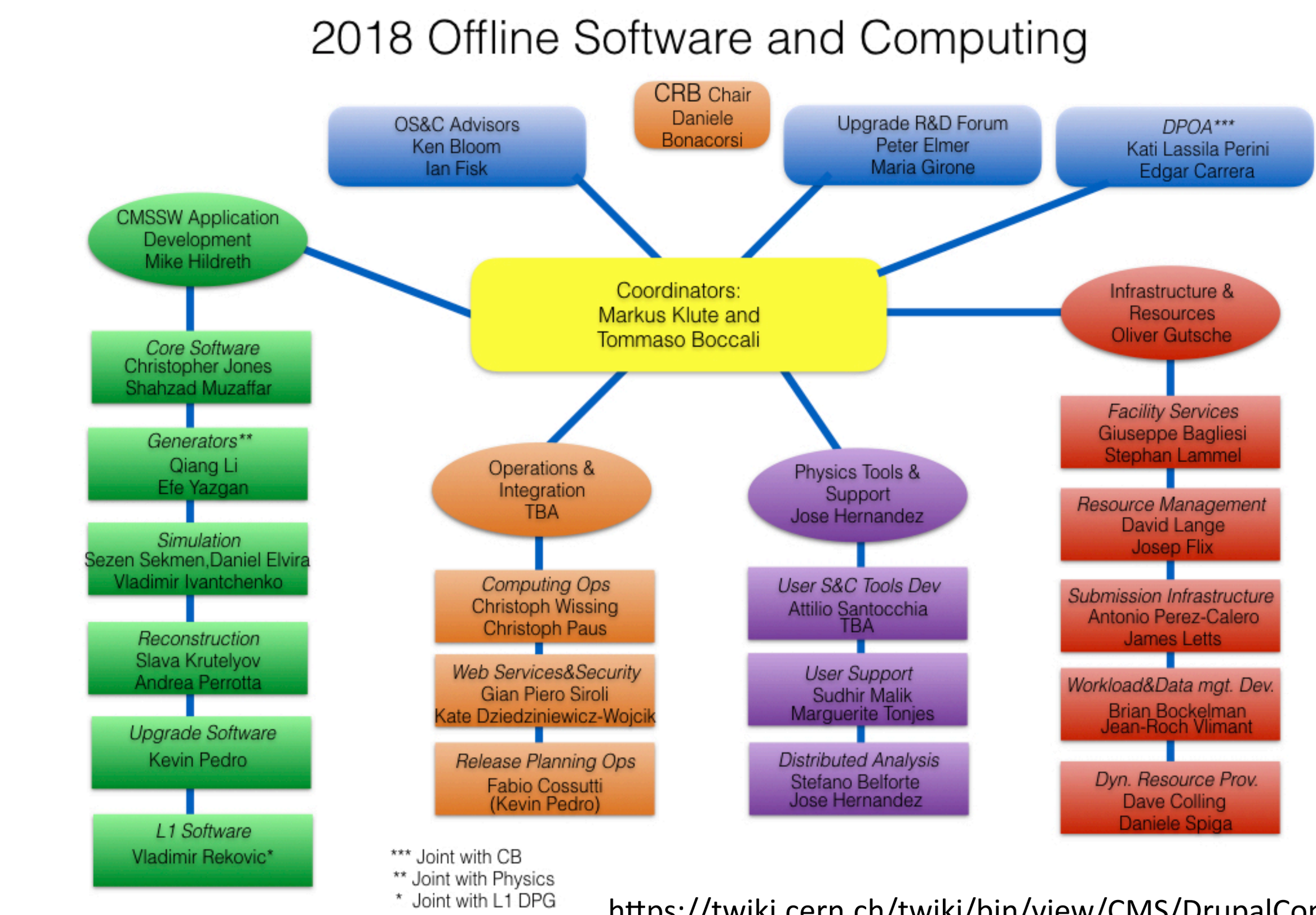


Physics Performance & Datasets (PPD) Organization



https://twiki.cern.ch/twiki/bin/view/CMS/PhysicsPerfomanceDatasetHome

Offline & Computing (O&C) Organization



https://twiki.cern.ch/twiki/bin/view/CMS/DrupalComputing

Outline: Exercises

https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideCMSDataAnalysisSchoolPisa2019PPDExercise

In this set of exercises, we have

- Exercise 0: Starting your RunII analysis
- Exercise 1: Exploring dataset using DAS : searching 2018 EGamma dataset as an example
- Exercise 2: Explore information for a Monte Carlo miniAODSIM sample from DAS, McM and pMp
- Exercise 3: Miscellaneous details about datasets
- Exercise 4: Generating MC NanoAOD events from scratch
- Exercise 5: Explore GlobalTag through cmsDBbrowser
- Exercise 6: Compute the integrated luminosity collected by CMS in RunII

Exercise 0: Starting your RunII analysis

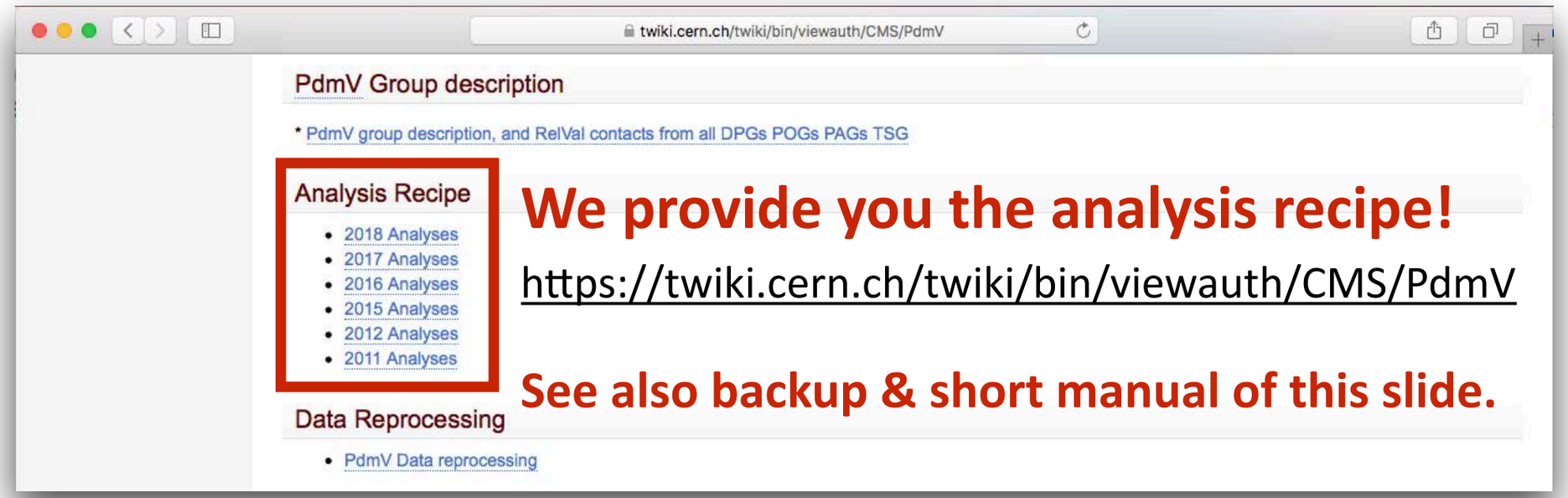
When you start your analysis

Several questions will come to you:

- Which datasets you are using?
- Do you need Monte-Carlo samples for your analysis?
  - ☑ Are they produced centrally, all of them?
  - ☑ Are they currently producing? How can you monitor them.
  - ☑ How to produce it privately?

    (For learning, not suggestion for private production)

- How many fb^{-1} of data you are using?
  - ☑ Are you sure that you use all available data? How?



A table for RunII analysis

PPD RunII Analysis Recipe

Updated: 2018-12-06A

https://twiki.cern.ch/twiki/bin/view/CMS/PdmV

	2016	2017	2018
CMSSW	CMSSW_8_0_X (AOD) CMSSW_9_4_X (Mini,Nano)	CMSSW_9_4_X	CMSSW_10_2_X
Data	**AOD:** 07Aug17* **MiniAOD:** 17Jul2018* **NanoAOD:** 22Aug2018*	**AOD:** 17Nov2017* **MiniAOD/NanoAOD:** 31Mar2018*	**AOD/MiniAOD:** 17Sep2018 (ABC), Prompt (D) **NanoAOD:** 14Sep2018 (ABCD)
Int. lumi. (Golden)	35.92 /fb	41.53 /fb	59.97 /fb
GT for data Analysis	80X_dataRun2_2016SeptRepro_v7 94X_dataRun2_v10	94X_dataRun2_v11	102X_dataRun2_Sep2018Rereco_v1 102X_dataRun2_Prompt_v11
MC Digi-Reco	RunIISummer16DRPremix	RunIIFall17DRPremix	RunIIAutumn18DRPremix
MC MiniAOD	RunIISummer16MiniAODv3	RunIIFall17MiniAODv2	RunIIAutumn18MiniAOD
MC NanoAOD	RunIISummer16NanoAODv3	RunIIFall17NanoAOD	RunIIAutumn18NanoAOD
GT for MC Analysis	80X_mcRun2_asymptotic_2016_TrancheIV_v8 94X_mcRun2_asymptotic_v3	94X_mc2017_realistic_v17	102X_upgrade2018_realistic_v12

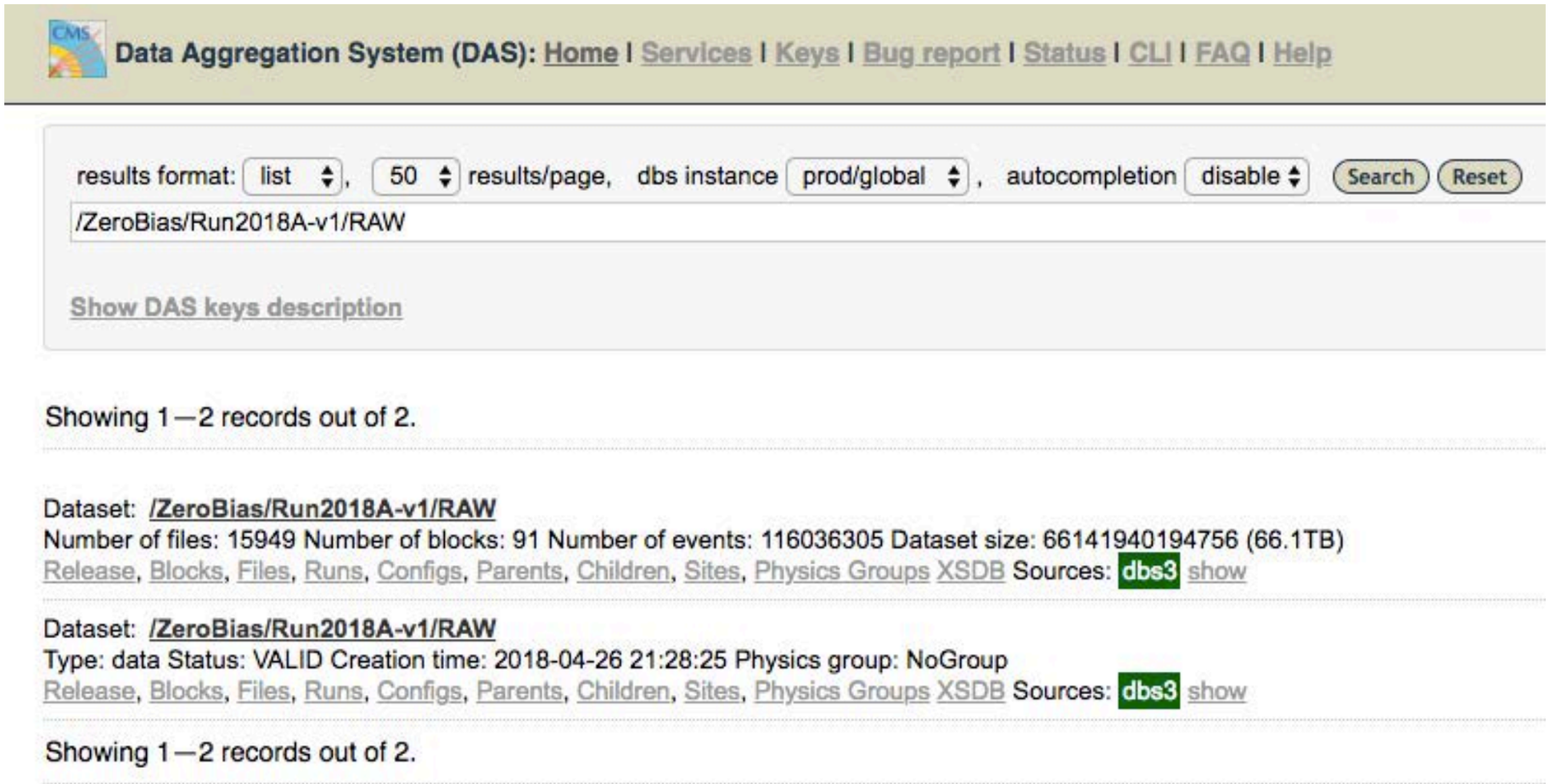
https://twiki.cern.ch/twiki/bin/viewauth/CMS/PdmVAnalysisSummaryTable

Exercise 1: Exploring dataset using DAS : searching 2018 EGamma dataset as an example

Data Aggregation Service (DAS)

Data Aggregation Service (DAS) is the place where you can look for samples.

- https://cmsweb.cern.ch/das/
  - ☑ Certificate is needed (CERN, or mapping to CERN account)
  - ☑ Use *dasgoclient*, see in Exercise 3
- DAS lists datasets and their properties (requestID, sites, run # and LS #....) aggregating information from various services.
- Dataset name structure is mentioned in the backup.



Data Tier

Datasets anatomy for both Data and MC

/**<PrimaryDataset>**/**<Campaign>-<Processing string>-<Processing version>**/**<data tier>**

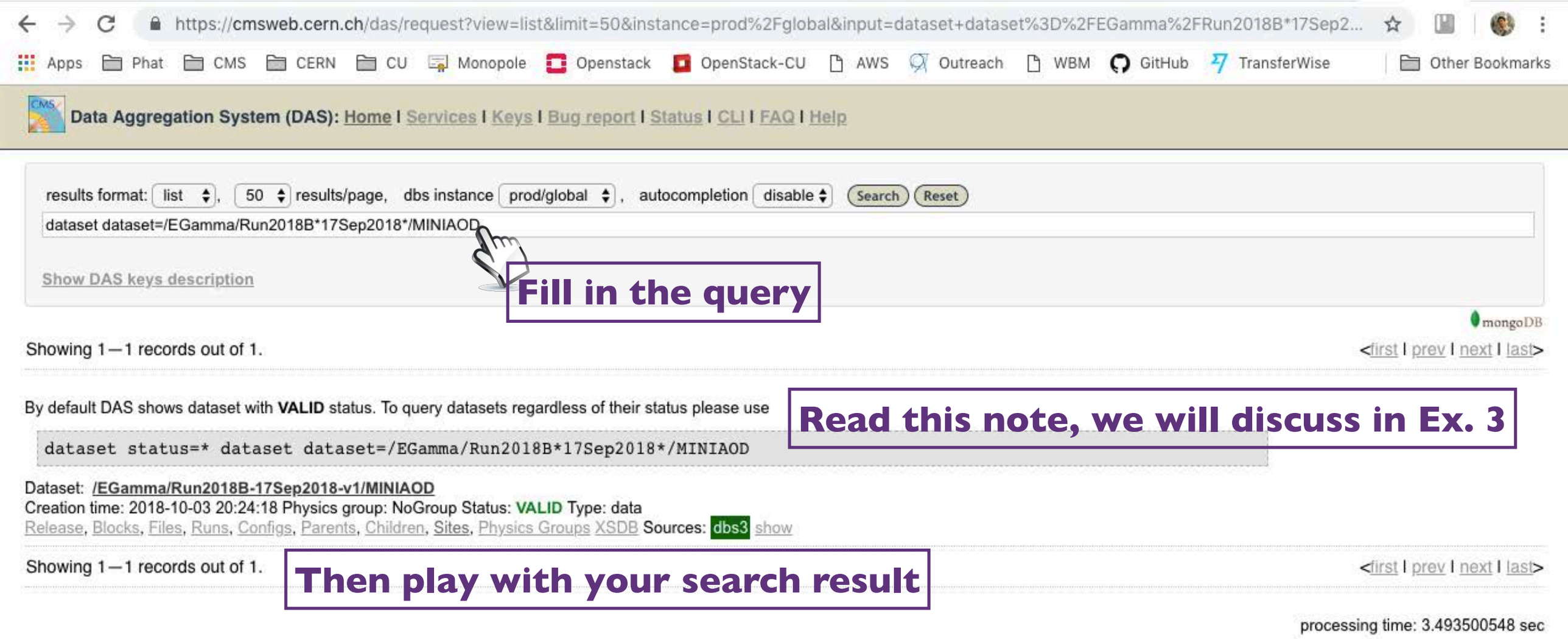
For example

- Data: /**EGamma**/**Run2018B-17Sep2018-v1**/**MINIAOD**
  - ☑ Campaign = CMS Run Period = **Run2018B**
  - ☑ Process string = **17Sep2018**
  - ☑ Processing version = **v1** (Exercise 3)
- MC: /**TTJets_TuneCP5_13TeV-amcatnloFXFX-pythia8**/**RunIIFall17MiniAODv2-PU2017_12Apr2018_new_pmx_94X_mc2017_realistic_v14-v1**/**MINIAODSIM**
  - ☑ Campaign = McM production campaign = **RunIIFall17MiniAODv2**
  - ☑ Process string
    - **PU2017_12Apr2018_new_pmx**; PU scenario + other string
    - **94X_mc2017_realistic_v14**; Global tag
    - Sometimes, you may see "**extX**" as sample is extension
  - ☑ Processing version = **v1** (Exercise 3)

- We use ***dash ("-")*** to separate between <Campaign>, <Processing string>, and <Processing version>. And ***underscore ("_")*** to link within <Processing string>.

DAS

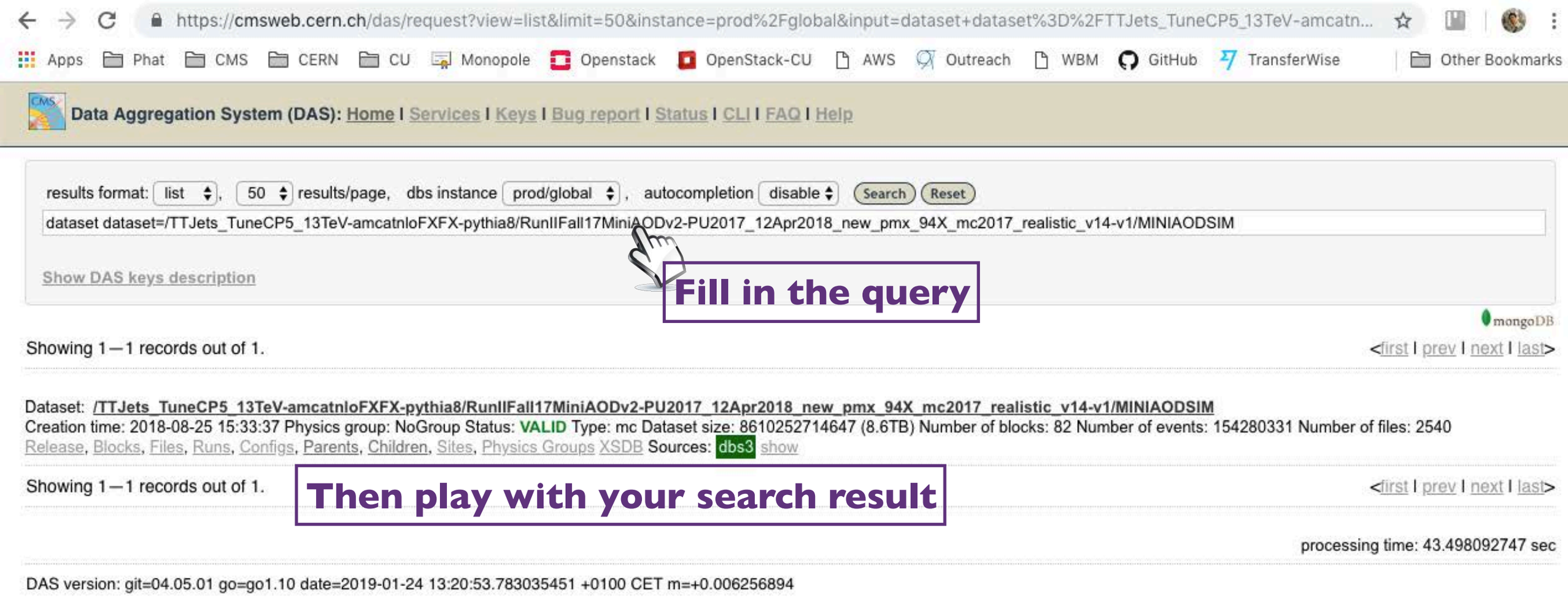
https://cmsweb.cern.ch/das/



Exercise 2: Explore information for a Monte Carlo MiniAODSIM sample from DAS, McM and pMp

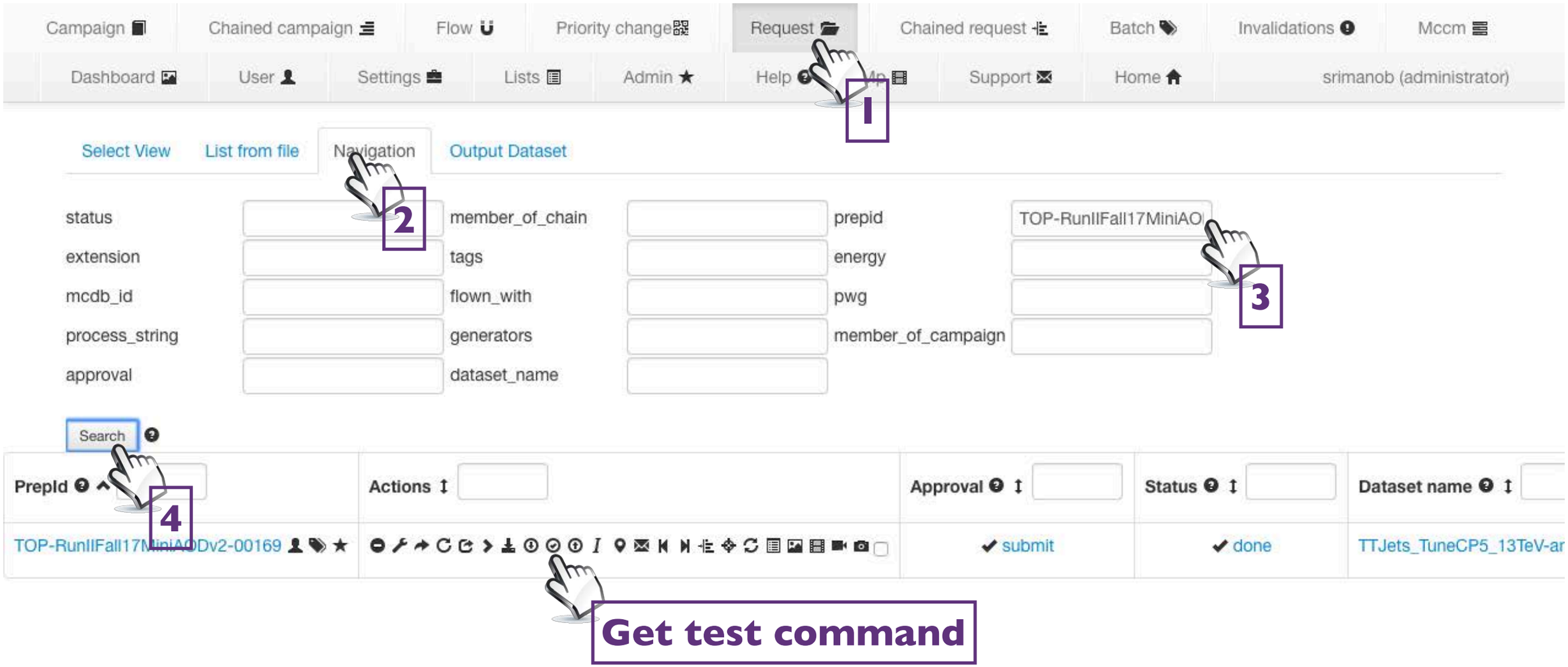
DAS

https://cmsweb.cern.ch/das/



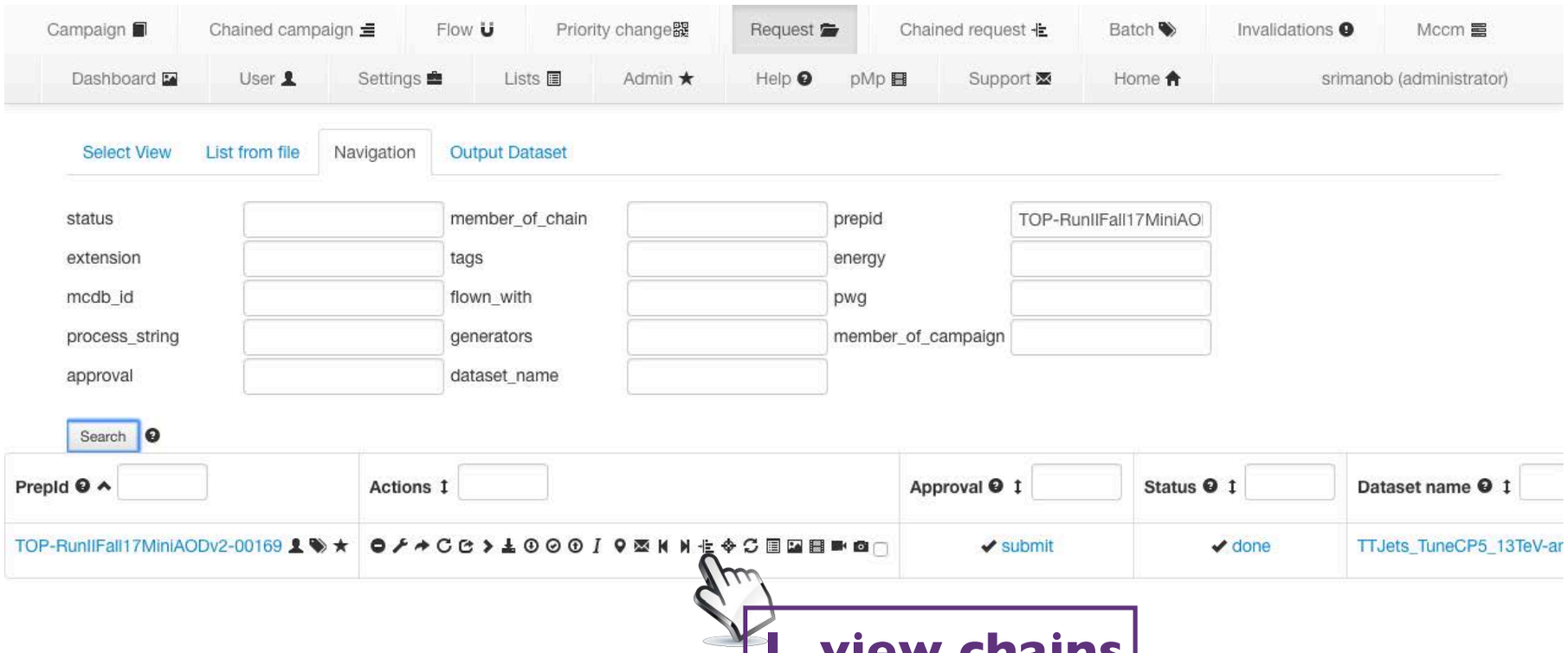
McM

https://cms-pdmv.cern.ch/mcm/

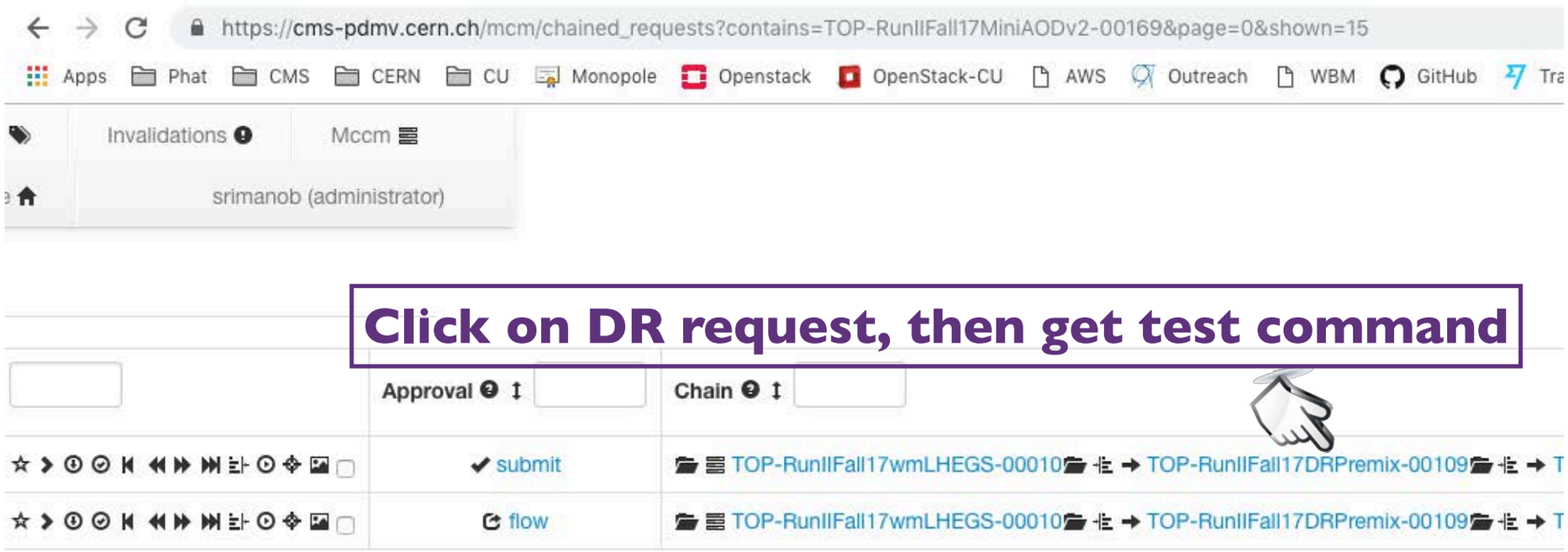


McM

Q 2.6: Get the McM test command sequence with the full **digi-reco** configuration.



1, view chains

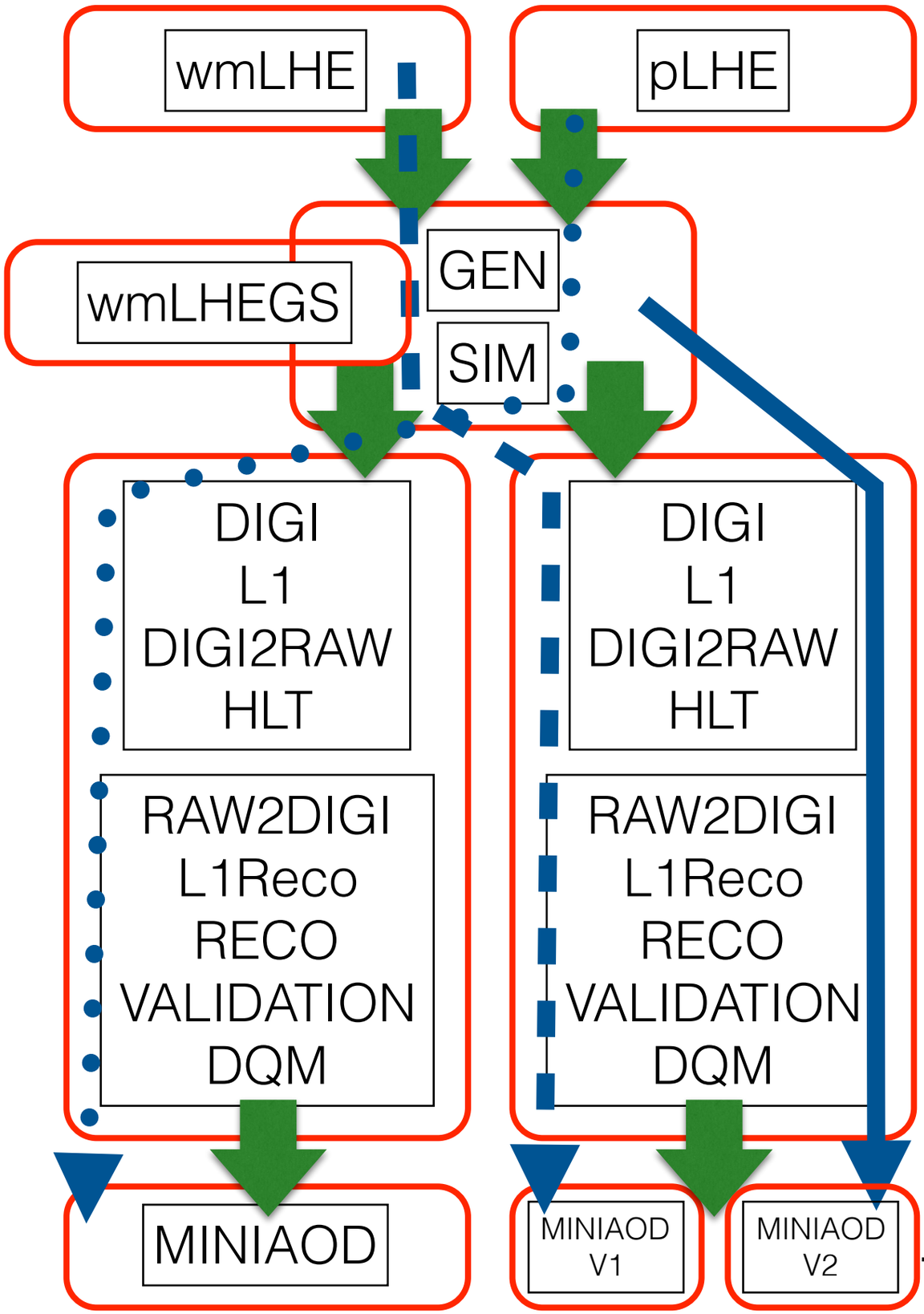


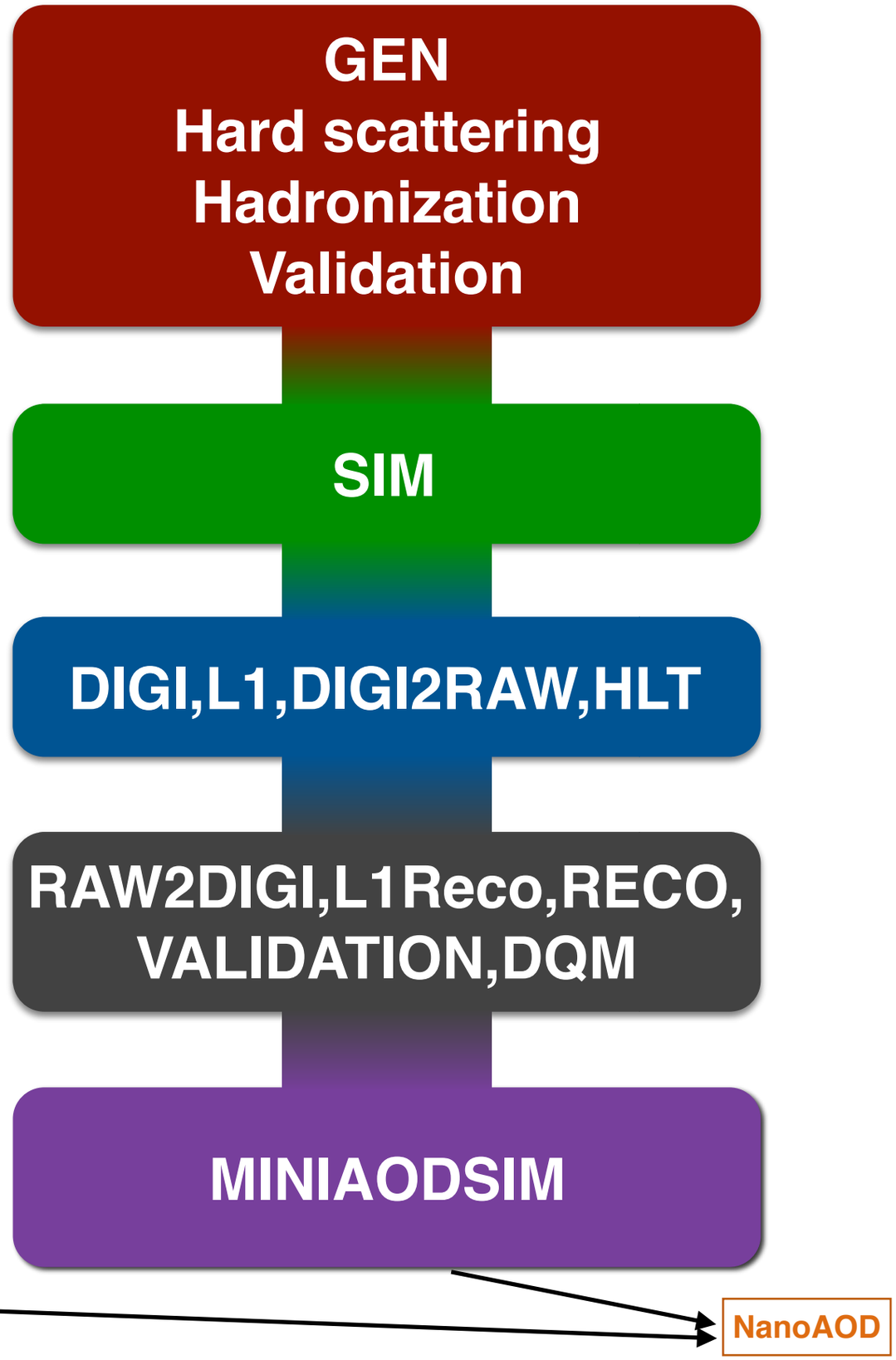
Click on DR request, then get test command

Exercise 3: Miscellaneous details about datasets

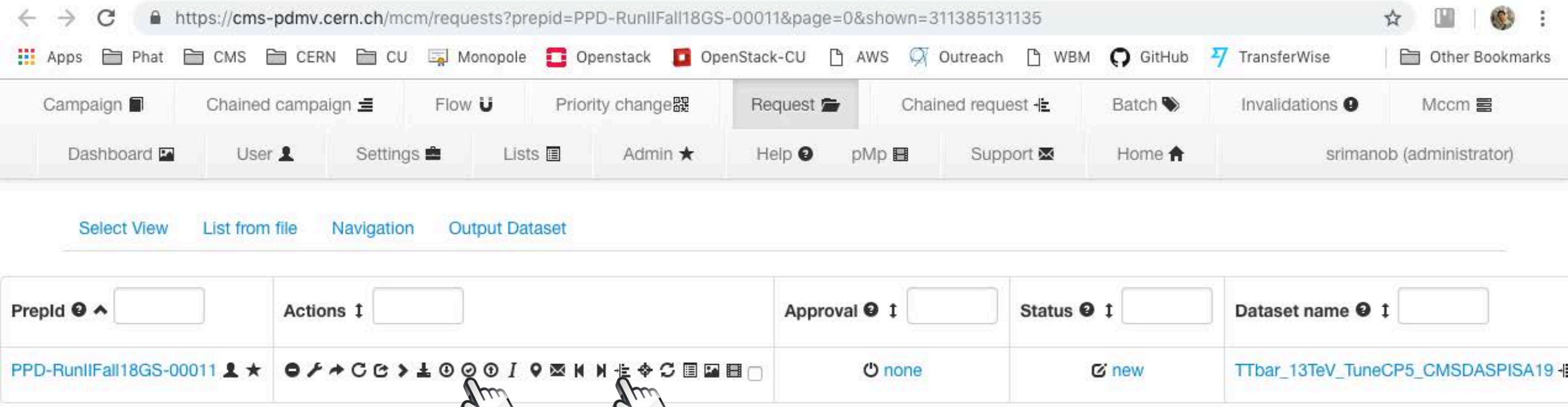
Exercise 4: Generating MC NanoAOD events from scratch

How do we organise MC production in CMS?



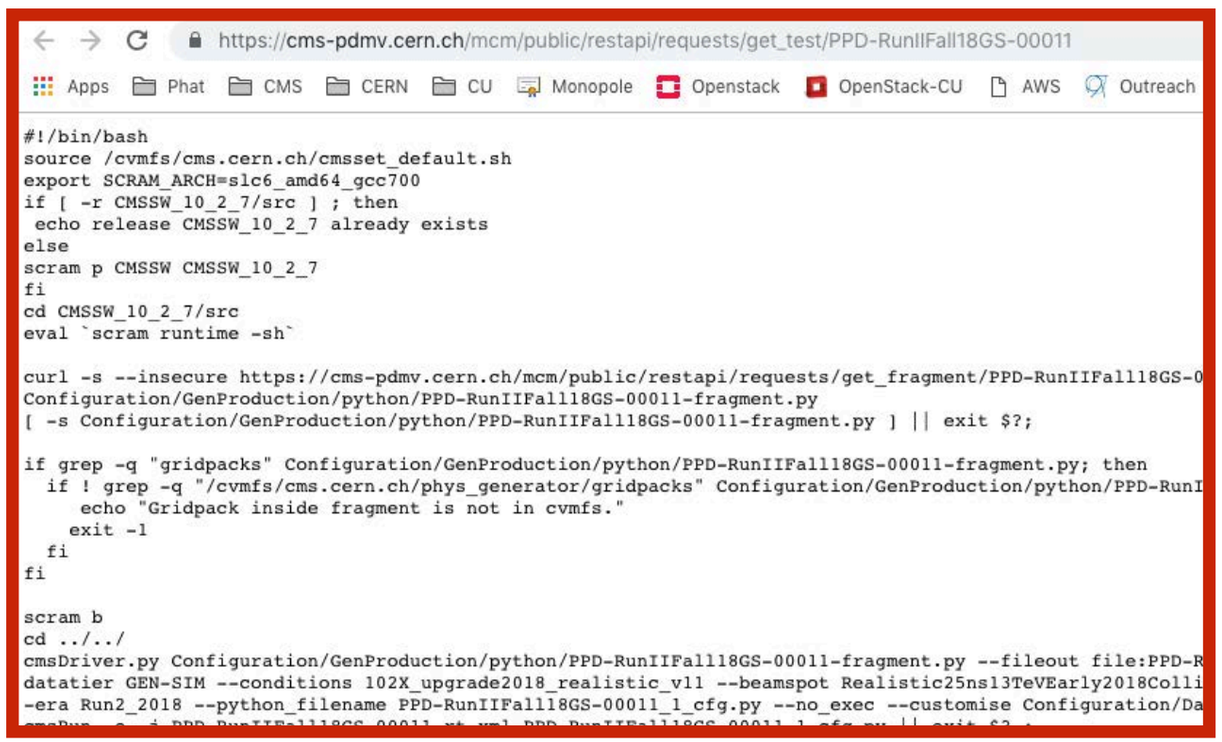


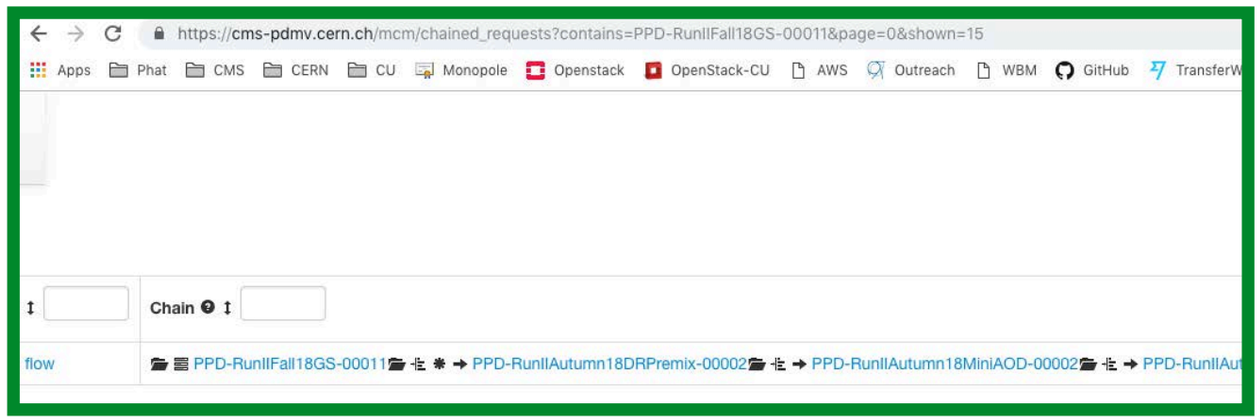
McM



Get test command

View chains





RunIIFall18GS-00011

Exercise 5: Explore GlobalTag through cmsDBbrowser

Exercise 6: Compute the integrated luminosity collected by CMS in RunII

BACKUP & SHORT MANUAL

Important Twiki links for Analysers

- Information about Golden JSONs and recent primary datasets with conditions : https://twiki.cern.ch/twiki/bin/viewauth/CMS/PdmV2017Analysis
  - Previous year twikis also exists with a change of year in url
- Global tags for both data and MC for all the scenarios : https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideFrontierConditions
- Documentation for BrilCalc : script to know luminosity which are considering for your analysis : https://cms-service-lumi.web.cern.ch/cms-service-lumi/brilwsdoc.html
- Information about the pile-up calculation for data : https://twiki.cern.ch/twiki/bin/view/CMS/PileupJSONFileforData#Pileup_JSON_Files_For_Run_II
- A plethora of information about data recorded by CMS is available here : https://cmswbm.cern.ch/ (just enter the run number and press “enter”)
- To know whats inside miniAOD and what are the latest recipes : https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookMiniAOD
- And there are many more from each POG for the latest recipes.

Important Twiki links for Analysers

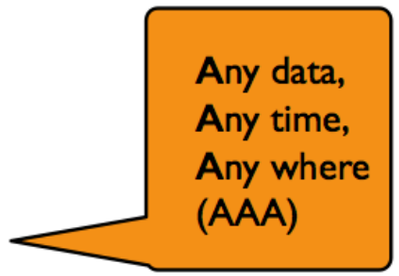
- Egamma Recipes for Run-II (id working points, scale factors and other recommendations) : https://twiki.cern.ch/twiki/bin/view/CMS/EgammaIDRecipesRun2
- Muon POG main twiki : https://twiki.cern.ch/twiki/bin/viewauth/CMS/MuonPOG
- B-Tag recommendations : https://twiki.cern.ch/twiki/bin/view/CMS/BtagRecommendation80XReReco
- JetMET main twiki: https://twiki.cern.ch/twiki/bin/view/CMS/JetMET

References

Contacts	Documentation
cms-ppd-coordinator@cern.ch cms-offcomp-coordinator@cern.ch	PPD Main Twiki Offline Main Twiki Computing Main Twiki
hn-cms-dataset-definition@cern.ch	DDT Twiki
hn-cms-computing-tools@cern.ch	DAS
hn-cms-prep-ops@cern.ch	PdmV Twiki
	Computing Model Workbook
hn-cms-offlineAnnounce@cern.ch hn-cms-relAnnounce@cern.ch	Offline Workbook SW Guide
hn-cms-physTools@cern.ch	MiniAOD Workbook
hn-cms-phedex@cern.ch	XROOTD doc Phedex - Phedex Workbook
hn-cms-relval@cern.ch hn-cms-physics-validation@cern.ch	PdmV Twiki
hn-cms-evfdqmannounce@cern.ch	DQM Twiki
hn-cms-data-certification@cern.ch	DQM-DC Twiki
hn-cms-data-certification@cern.ch	JSON File Twiki
hn-cms-luminosity@cern.ch cms-dpg-conveners-bril@cern.ch cms-pog-conveners-lum@cern.ch	brilcalc Doc bril dpg lumi pog
hn-cms-alca@cern.ch	AlCaDB Twiki GlobalTag Twiki

Accessing Datasets

- Datasets for analysis ⟶ they need to be on disk @ Tier2/3 (GRID = "where" doesn't matter)
  - **AnalysisOps** usually subscribes the most common samples to "central" space ⟶ no action needed
  - what if you need something more "exotic"? (are you sure you do need it??)
- You can ask the transfer @ T2s using the Phedex service
  - **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 (usually 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 Management** tools
    ⟶ optimal data replication and quota control
- before submitting I want to run on a single file locally
  ⟶ how do I do it
  - most (almost all) datasets accessible remotely via XROOTD protocol
    ⟶ e.g. can run @ CERN reading files @ FNAL



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

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:

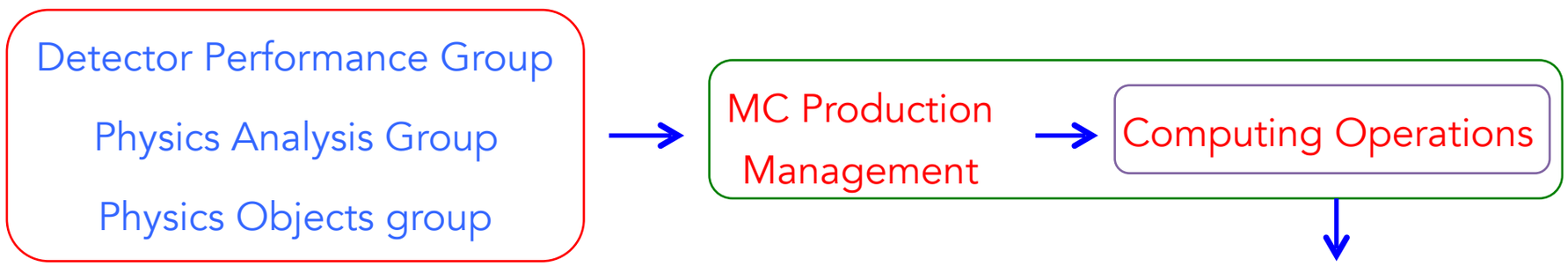
- data (prompt reco): /SingleElectron/Run2015D-PromptReco-v3/AOD
- data (re-reco): /SingleElectron/Run2015D-16Dec2015-v1/MINIAOD
- MC (RunIIFall15DR76):
  /WprimeToMuNu_M-1600_TuneCUETP8M1_13TeV-pythia8/
  RunIISpring15MiniAODv2-74X_mcRun2_asymptotic_v2-v1/MINIAODSIM

Data tier indicates the collections available at each event

Contact&Doc

MC production management

- About 20 groups in CMS: physics analyses (PAGs), detector performance (DPGs), physics object studies (POGs)
- 100s of physics analyses; 1000s of MC samples needed; Billions of events required for ongoing Run-II of LHC
  - CMS has produced over 20 Billion simulated events in 1 year production in 2016-17
- Strong and efficient production infrastructure required: bookkeeping and interface to computing resources



Multiple actions are performed by following 4 roles

Generator Contact: collects the needs of simulated samples from detector or physics groups, produces generator configuration and presents them to the production team for execution

Generator Convener: scrutinizes and approves event generator configurations

Request Manager: configures campaigns and flows, performs requests' chaining, sets their priority and submits requests to the production infrastructure

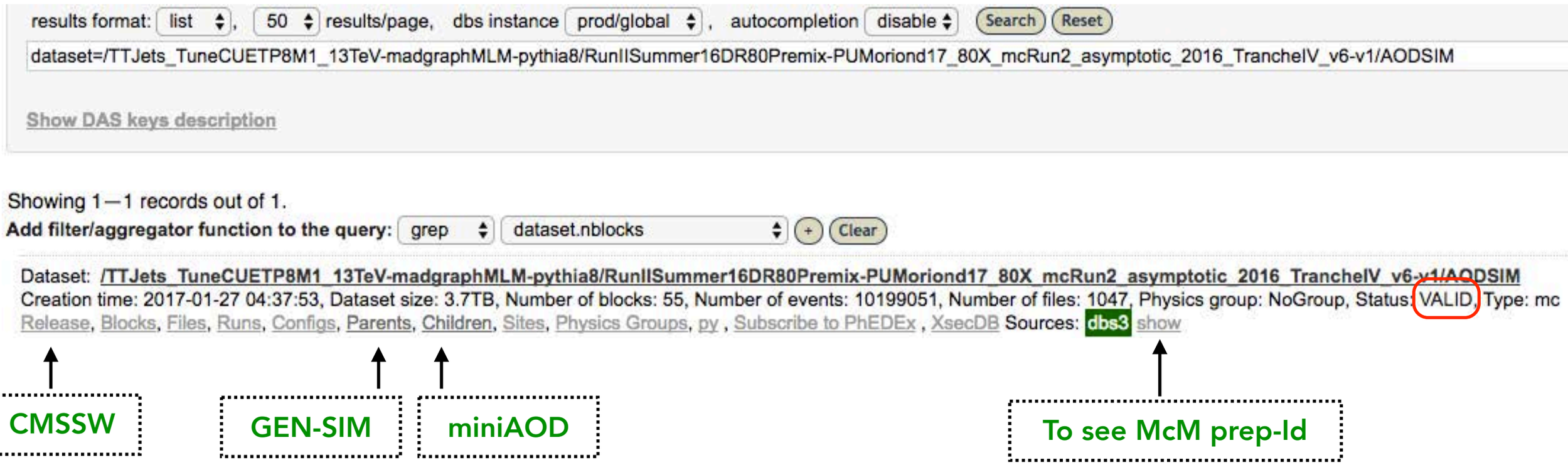
} McM

Production Manager: handle workflows during production and send datasets to DAS ← Computing tools

CMS Analysts: performs analyses, and communicates needs with production team

Data Aggregation System (DAS)

- All completed samples will be available in DAS
- Analysts can find various type of data samples and can transfer files to laptop/Tier-2/3
- General queries about DAS, command line tool, etc. are available at DAS-FAQs
- Find datasets in DAS: https://cmsweb.cern.ch/das/
- Wildcards (*) can be used but try to be as specific as possible as DAS may show many combinations



DAS

- Status **VALID**: production is finished and dataset is available
- Status **PRODUCTION**: Its statistics is still growing due to running production jobs and dataset is not yet announced. But analysts can still run over the existing statistics by using crab parameter "allowNonValidInputDataset" parameter in [CRAB3 configuration parameters]

Dataset: /TT_JpsiFilter_TuneCUETP8M1_mtop166_5_14TeV-powheg-tauola-pythia8/PhaseIITDRSpring17MiniAOD-noPU_PhaseIITDRSpring17_91X_upgrade2023_realistic_v3-v1/MINIAODSIM
Creation time: 2017-06-12 14:24:31, Physics group: NoGroup, Status: PRODUCTION, Type: mc
Release, Blocks, Files, Runs, Configs, Parents, Children, Sites, Physics Groups, py , Subscribe to PhEDEx , XsecDB Sources: dbs3 show

Collection of ≥1 files

Global tag and CMSSW release

You can transfer full dataset or only a Block to your Tier-2/Tier3: Subscribe to PhEDEx. But please remember that:

- DISK SPACE IS VERY CRITICAL
- Very high threshold on private transfer requests

Showing 1—3 records out of 3.
Add filter/aggregator function to the query: grep site.dataset_fraction + Clear

Site: T1_ES_PIC_Disk
Block completion: 100%, Block presence: 100%, Dataset presence: 100%, File-replica presence: 100%, Site type: Disk, StorageElement: srmcms.pic.es
Datasets, SiteDB Sources: phedex combined show

Site: T1_US_FNAL_Buffer
Block completion: 100%, Block presence: 100%, Dataset presence: 100%, File-replica presence: 100%, Site type: **TAPE no user access**, StorageElement: cmsdcatape01.fnal.gov
Datasets, SiteDB Sources: phedex combined show

Site: T1_US_FNAL_MSS
Block completion: 100%, Block presence: 100%, Dataset presence: 100%, File-replica presence: 100%, Site type: **TAPE no user access**, StorageElement: cmsdcatape01.fnal.gov
Datasets, SiteDB Sources: phedex combined show

PhEDEx Transfer of Datasets

Subscribe to PhEDEx

Replace this data item with a block address if full dataset is not added

Subscribe data

Data Items [?] /TTJets_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/RunIISummer16DR80Premix-PUMoriond17_80X_mcRun2_asymptotic_2016_TrancheIV_v6-v1/AODSIM

DBS https://cmsweb.cern.ch/dbs/prod/global/DBSReader change

Destination

☐ T3_GR_IASA_HG	☐ T3_HR_IRB	☐ T3_HU_Debrecen	☐ T3_IN_VBU
☐ T3_IR_IPM	☑ T3_IT_Bologna	☐ T3_IT_Firenze	☐ T3_IT_MIB
☐ T3_IT_Napoli	☐ T3_IT_Perugia	☐ T3_IT_Trieste	☐ T3_KR_KISTI

Site Custodial [?] ○ yes ◉ no

Subscription Type [?] ◉ growing ○ static

Transfer Type [?] ◉ replica ○ move

Priority [?] ○ high ◉ normal ○ low

User Group [?] higgs

Data injected after [?] YYYY-MM-DD [hh:mm:ss]

Email gurpreet.singh@cern.ch

Comments enter any additional comments here

Reset | Preview | Accept

Luminosity: HowTo

- Need to know the luminosity of ALL the LSs you run on and ONLY those
  - 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 →

CRAB output {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]]
}
```


- Lumi POG provides tools to compute luminosity starting from JSON file:
  - brilCalc.py: reports measurements of LHC delivered, CMS recorded luminosity for LSs in the JSON
- NOTE: using directly the certification JSON is not correct for several reasons:
  - LS that failed prompt reco or re-reco will appear in the cert. JSON
  - LS that failed in your jobs will appear in the cert. JSON

Contact&Doc