Machine Learning for Tracker DQM for CMS Experiment

Guillermo Fidalgo

(Physics Department)

Mentor: Sudhir Malik

Internado realizado en: Fermi National Accelerator Laboratory

Campo estudiado: Alta Energía

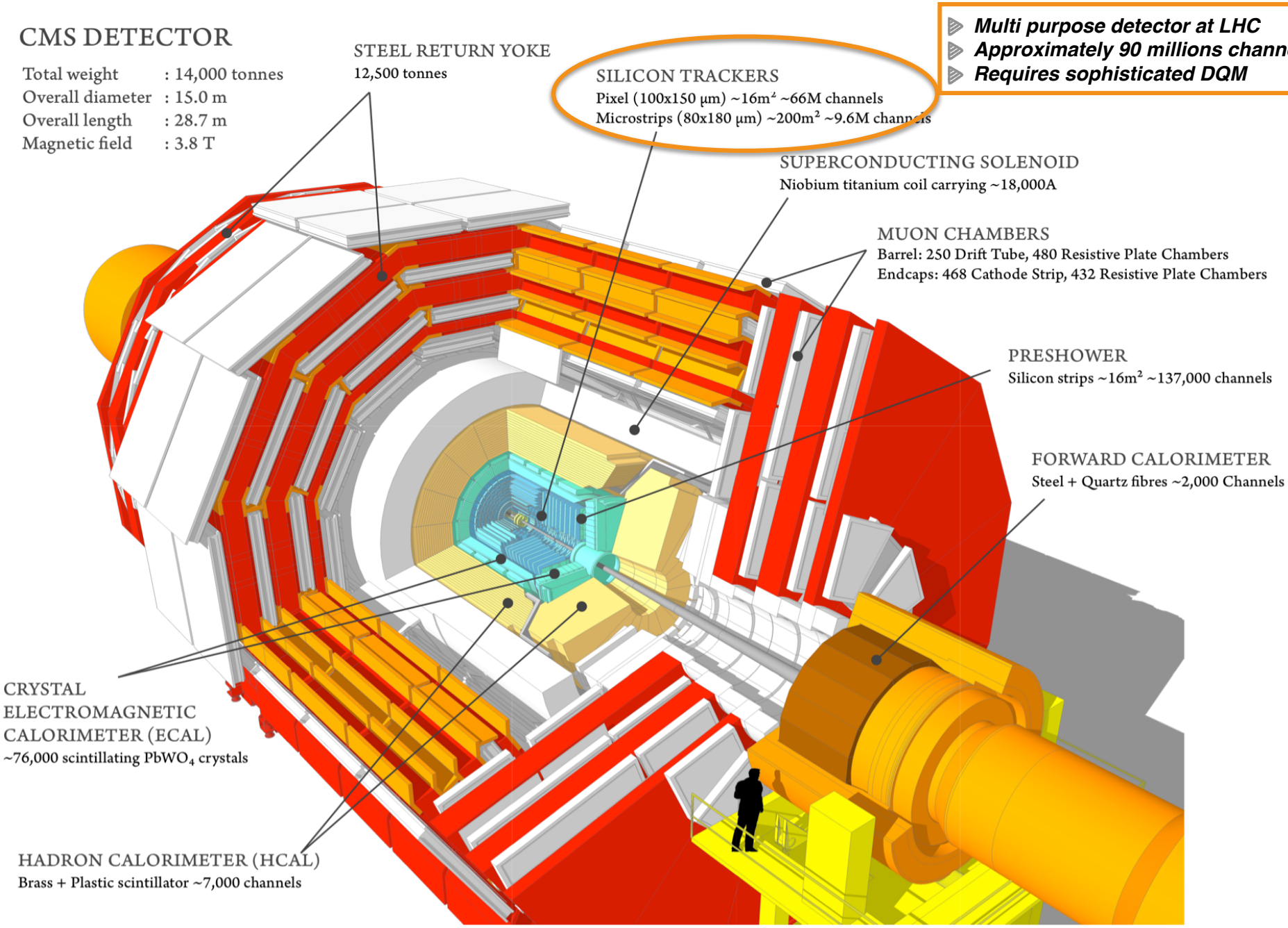
Support: NSF

Simposio de investigaciones en Física, Verano 2019

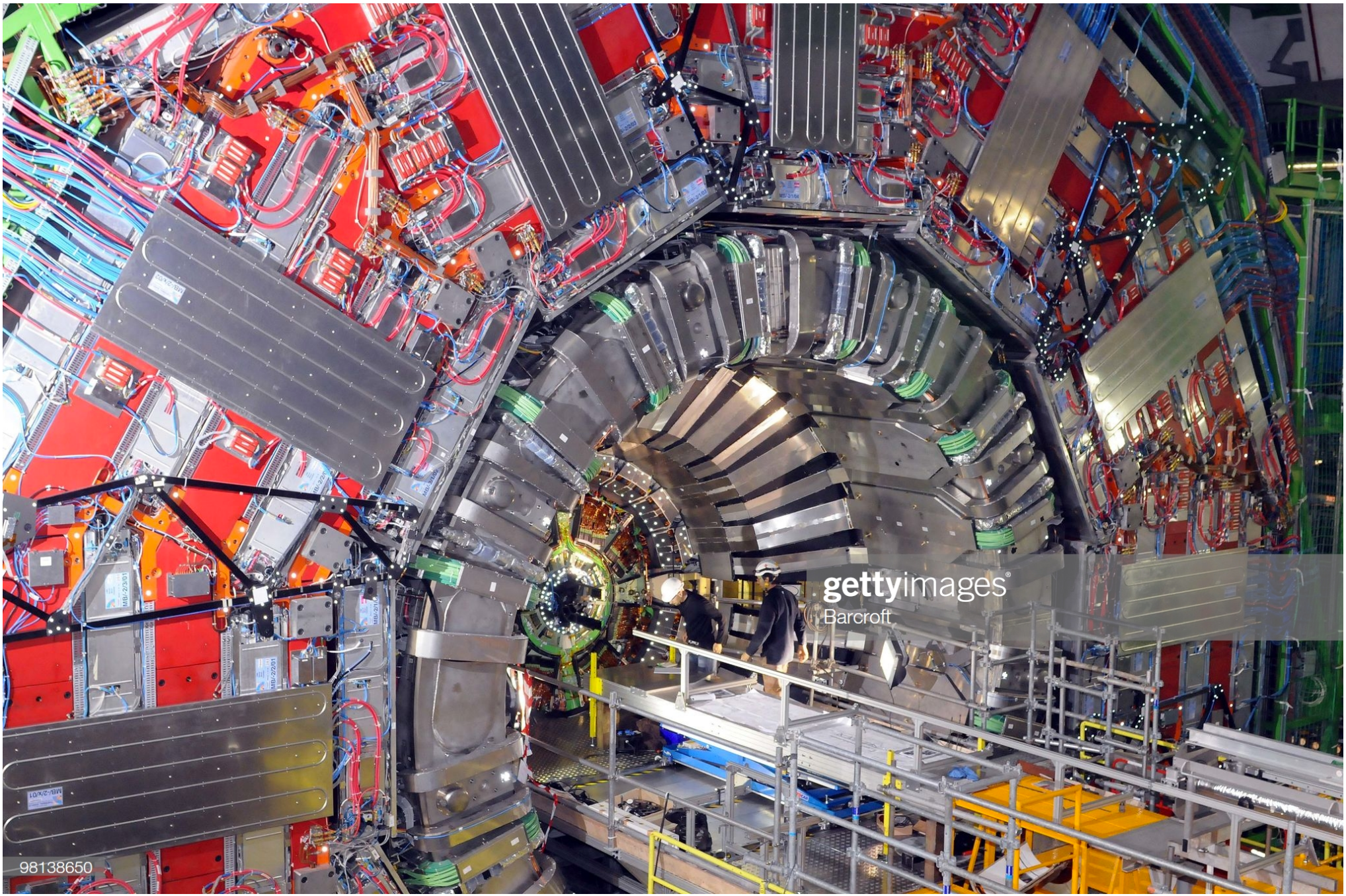
Large Hadron Collider (LHC)

CMS Detector

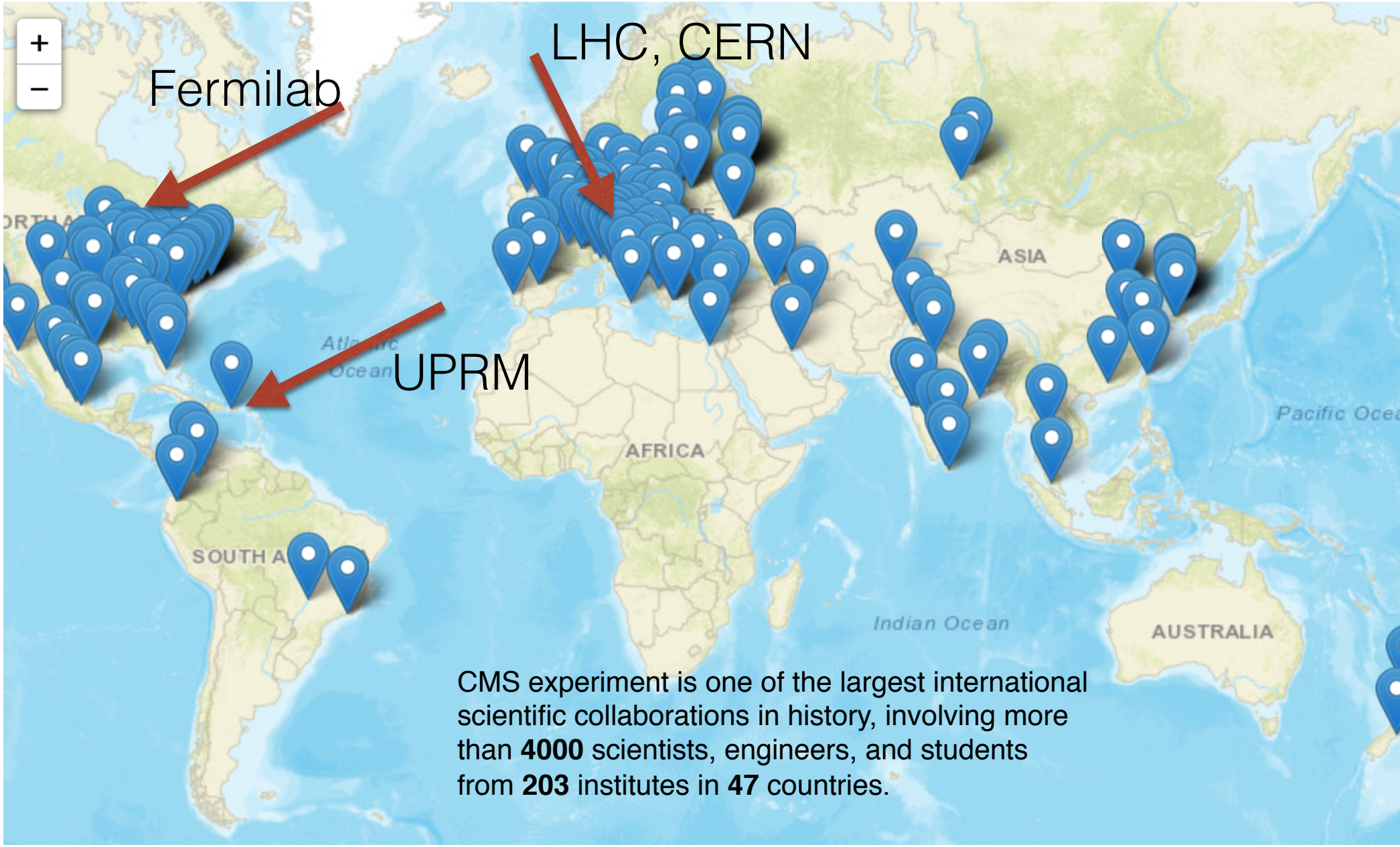
- ***Multi purpose detector at LHC***
- ***Approximately 90 millions channels***
- ***Requires sophisticated DQM***



CMS Detector



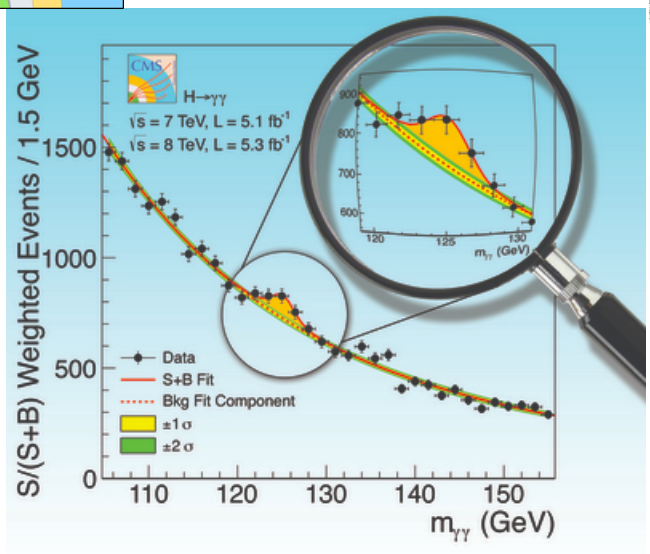
CMS people



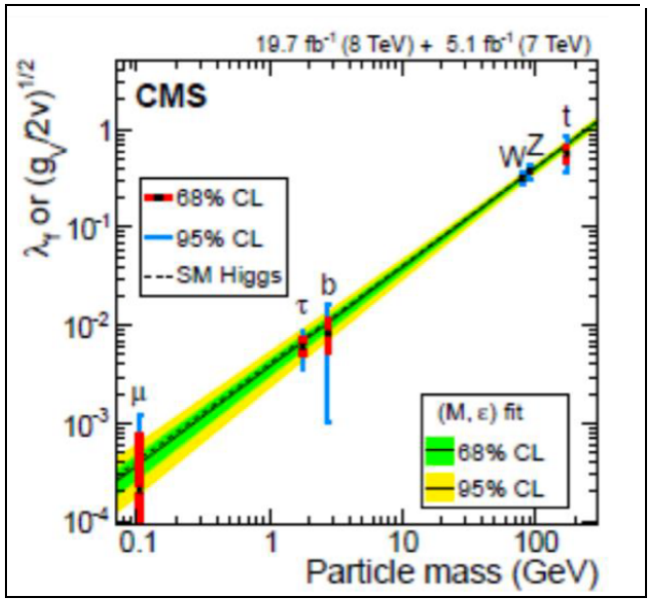
CMS experiment is one of the largest international scientific collaborations in history, involving more than **4000** scientists, engineers, and students from **203** institutes in **47** countries.

Fermilab

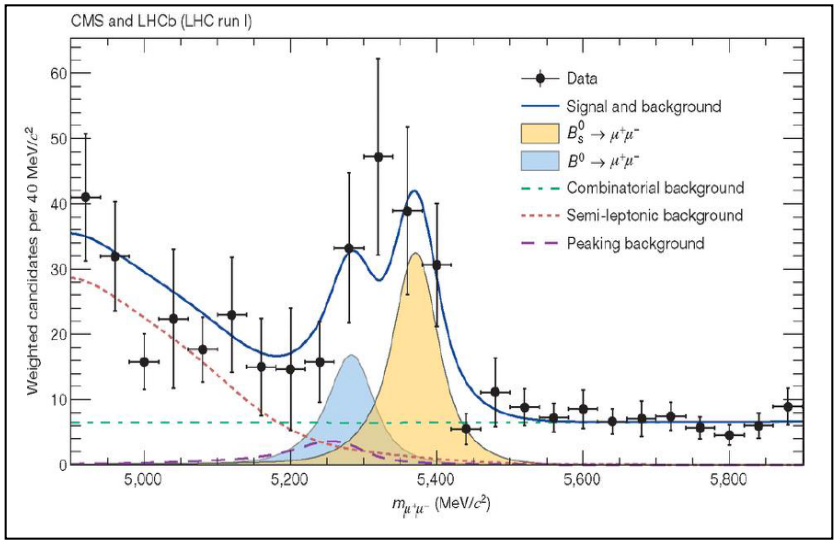
Accomplishments



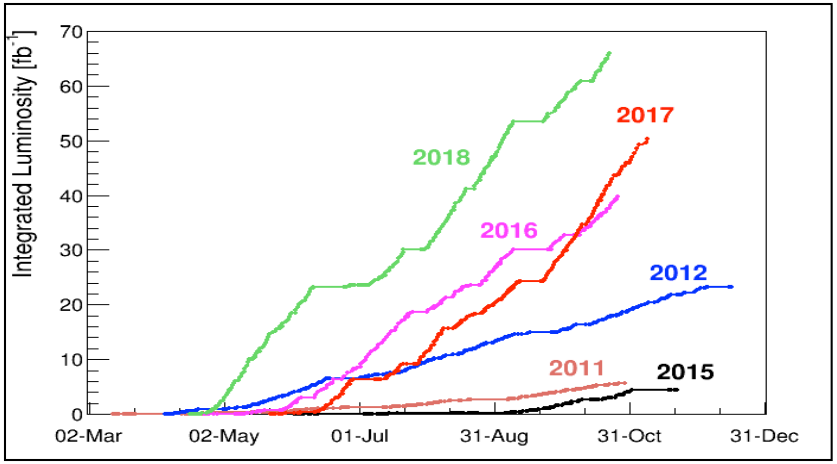
- Higgs discovery (completing the Standard Model) Precise masses of the Higgs boson



- Higgs boson coupling to the heaviest, third generation of quarks and leptons

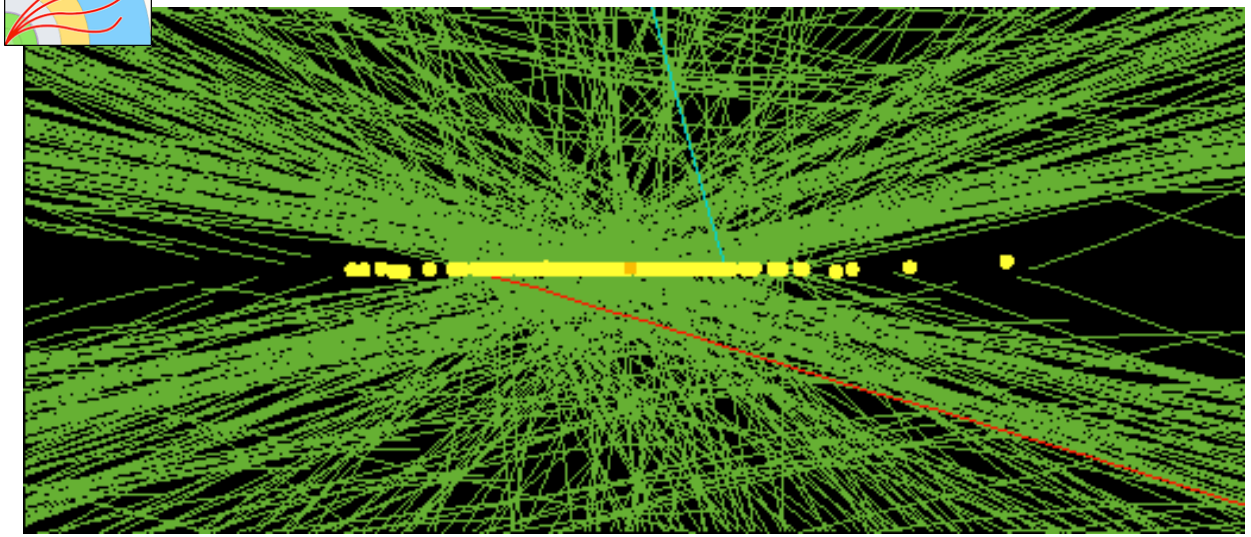


- $B_s \rightarrow \mu\mu$ branching ratio - constraining the Standard Model



- Integrated Luminosity CMS ~ 190 fb^{-1}

Challenges



CMS event with 78 collisions in one pp bunch crossing – less than half as dense as a typical beam crossing in HL-LHC

- HL-LHC will deliver ~ 140-200 collisions/beam crossing (40MHz)
  - Of which a small number are interesting,
  - Additional collisions in the same beam crossing called pile up (PU)
- This means CMS must be able to:
  - Separate and identify particles in extremely dense collision debris
  - Trigger on events at 40 MHz rate -> keeping only 7.5 kHz of data
- And do all of this in a high radiation environment
  - Must develop novel radiation hard particle detectors and electronics
- Certify good and big data efficiently
  - Must develop efficient way of DQM

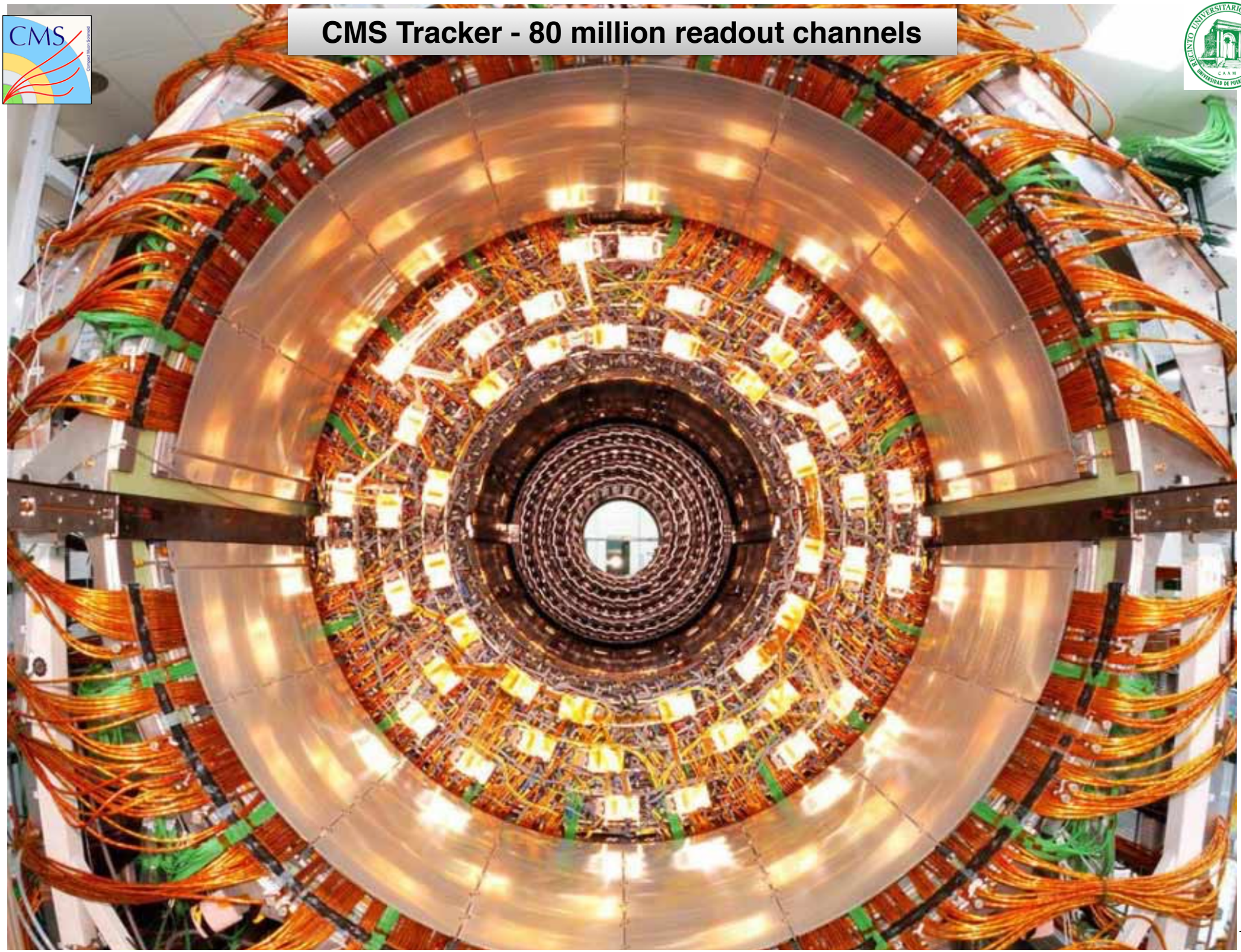
The amount of data collected for each event is around 1 MB (1 Megabyte).

109 collisions/s x 1 Mbyte/collision = 1015 bytes/s = 1 PB/s (1 Petabyte/second)

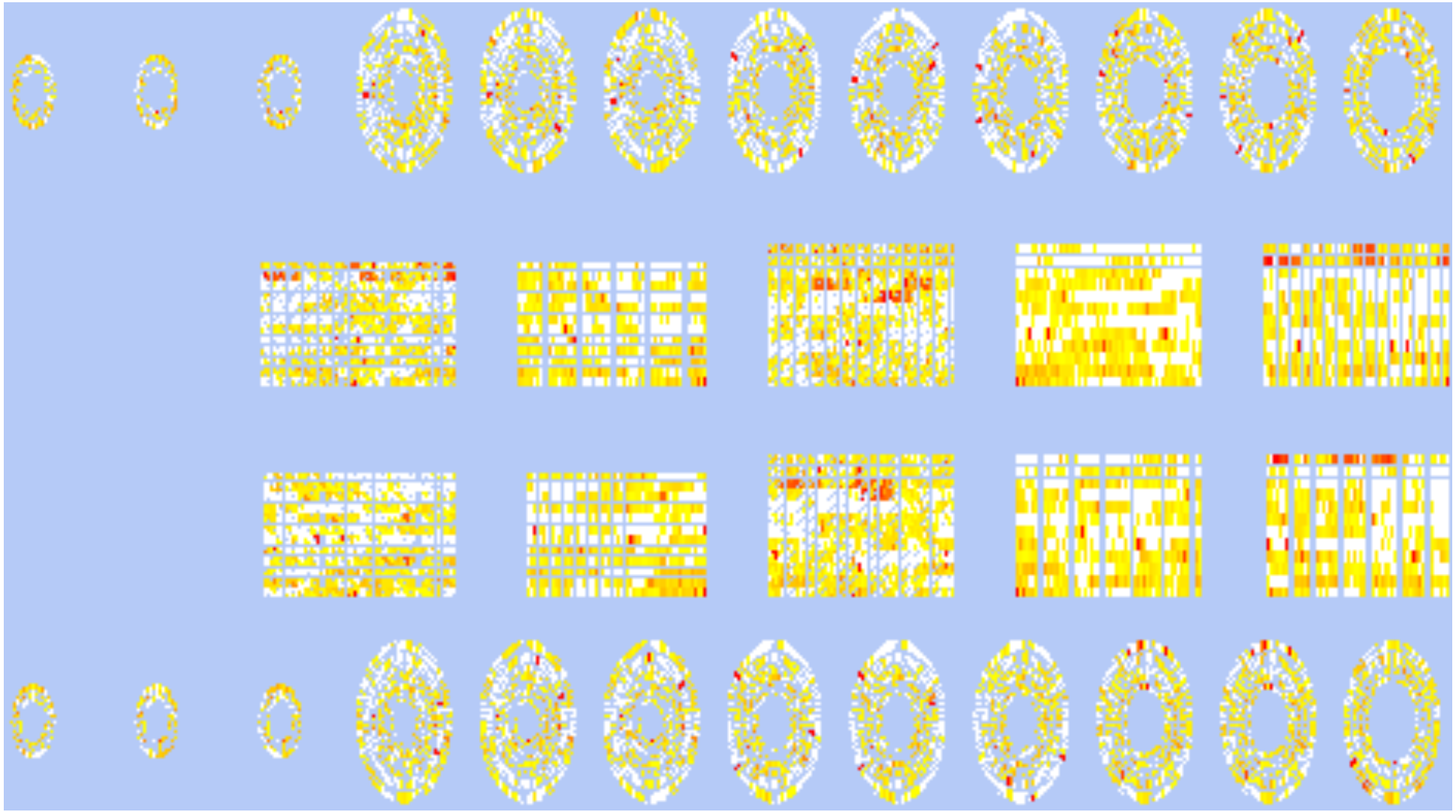
Since 1 DVD ~ 5 GB : 200000 DVDs per second would be filled or about 6000 IPods (ones with 160 GB of storage) per second!

CMS control room for DQM

CMS Tracker - 80 million readout channels



Readout divided into different sectors

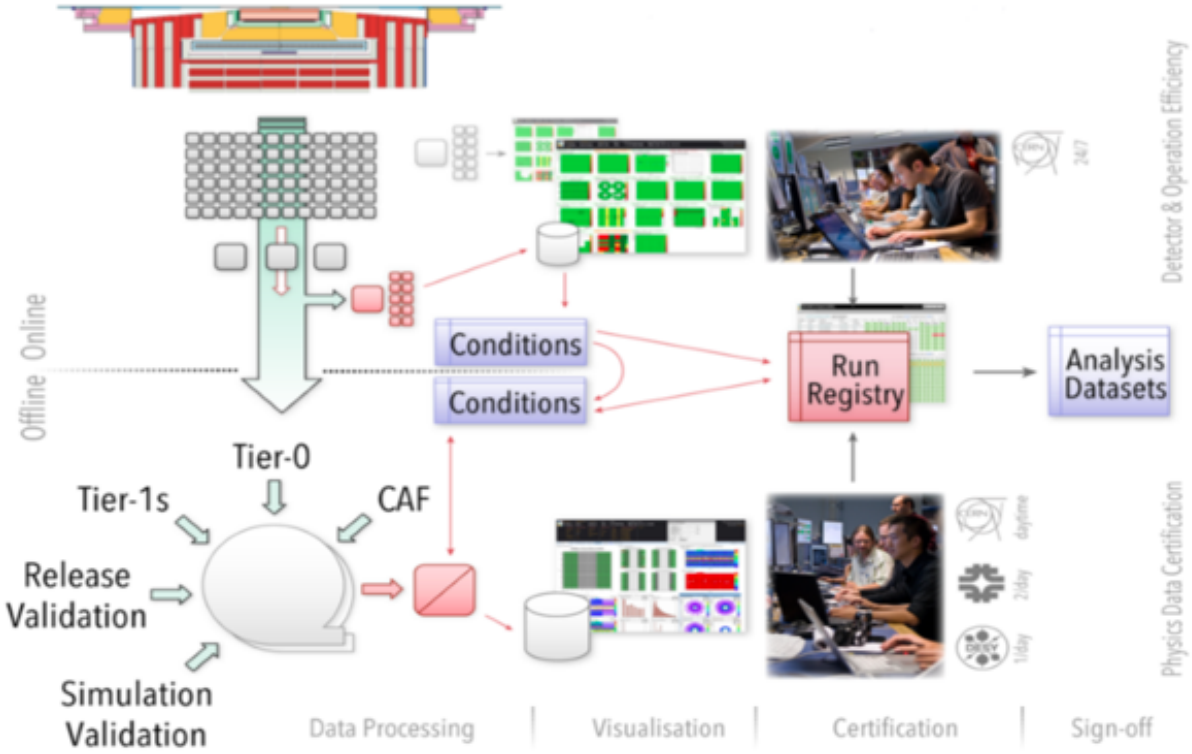


Data Quality Monitoring flow

Collection of tools and processes to provide

- Monitoring - Detector and operation performance and malfunctions
- Certification - Assess and record quality of data and software releases
- Debugging - Provide detailed information in case of problems

Humans are a central part of DQM



OFFLINE

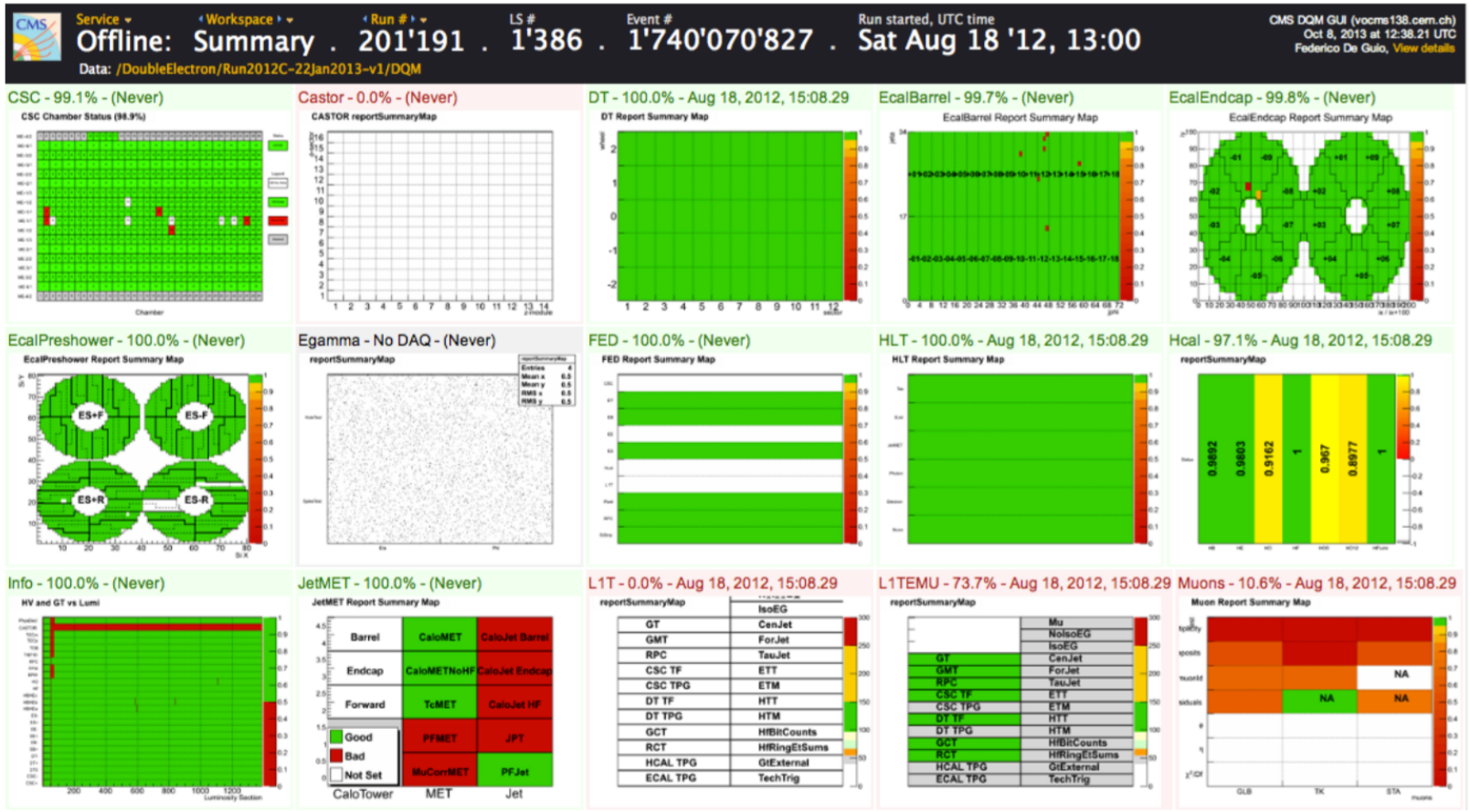
- **After data taking**
- **Responsible for bookkeeping and certifying the final data with fine time granularity**

ONLINE

- **Provides feedback of live data taking.**
- **Alarms if something goes wrong**

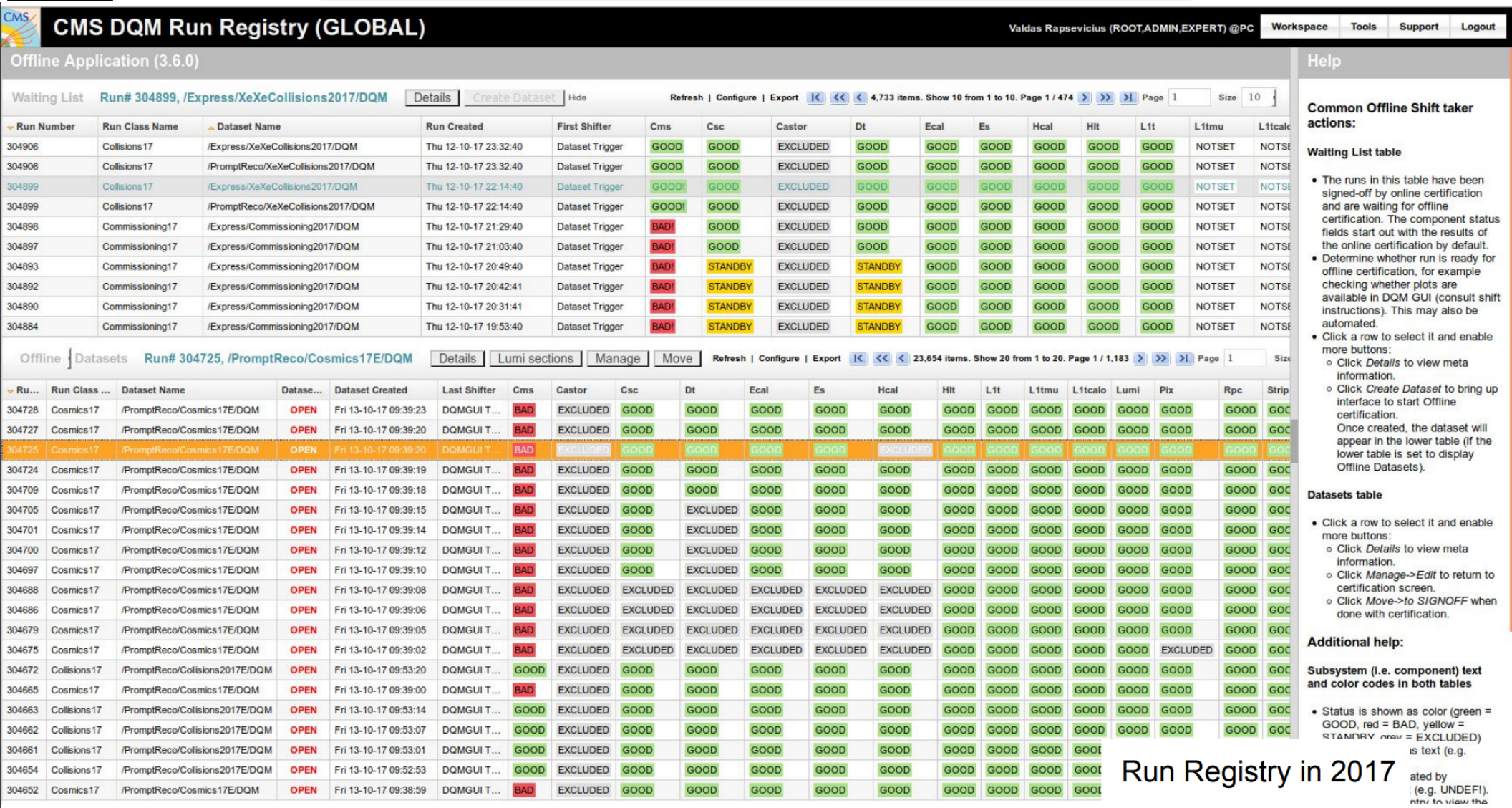
DQM GUI

20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013) IOP Publishing
Journal of Physics: Conference Series **513** (2014) 032024 doi:10.1088/1742-6596/513/3/032024



Web service to collect and archive monitoring elements (ME), API scripts
Web based interface to browse realtime and historical data
DQM GUI provides access to e.g. - Online: 22,000 runs, 650 GB
- Offline: 400,000 datasets, 4100 GB
- 100,000 MEs per Run

Run Registry - central tool for tracking the data quality and data certification



CMS DQM Run Registry (GLOBAL)

Valdas Rapsevicius (ROOT,ADMIN,EXPERT) @PC | Workspace | Tools | Support | Logout

Offline Application (3.6.0)

Waiting List — Run# 304899, /Express/XeXeCollisions2017/DQM — Details | Create Dataset | Hide — Refresh | Configure | Export — 4,733 items. Show 10 from 1 to 10. Page 1 / 474

Run Number	Run Class Name	Dataset Name	Run Created	First Shifter	Cms	Csc	Castor	Dt	Ecal	Es	Hcal	Hlt	L1t	L1tmu	L1tcalc
304906	Collisions17	/Express/XeXeCollisions2017/DQM	Thu 12-10-17 23:32:40	Dataset Trigger	GOOD	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	NOTSET	NOTSE
304906	Collisions17	/PromptReco/XeXeCollisions2017/DQM	Thu 12-10-17 23:32:40	Dataset Trigger	GOOD	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	NOTSET	NOTSE
304899	Collisions17	/Express/XeXeCollisions2017/DQM	Thu 12-10-17 22:14:40	Dataset Trigger	GOOD!	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	NOTSET	NOTSE
304899	Collisions17	/PromptReco/XeXeCollisions2017/DQM	Thu 12-10-17 22:14:40	Dataset Trigger	GOOD!	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	NOTSET	NOTSE
304898	Commissioning17	/Express/Commissioning2017/DQM	Thu 12-10-17 21:29:40	Dataset Trigger	BAD!	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	NOTSET	NOTSE
304897	Commissioning17	/Express/Commissioning2017/DQM	Thu 12-10-17 21:03:40	Dataset Trigger	BAD!	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	NOTSET	NOTSE
304893	Commissioning17	/Express/Commissioning2017/DQM	Thu 12-10-17 20:49:40	Dataset Trigger	BAD!	STANDBY	EXCLUDED	STANDBY	GOOD	GOOD	GOOD	GOOD	GOOD	NOTSET	NOTSE
304892	Commissioning17	/Express/Commissioning2017/DQM	Thu 12-10-17 20:42:41	Dataset Trigger	BAD!	STANDBY	EXCLUDED	STANDBY	GOOD	GOOD	GOOD	GOOD	GOOD	NOTSET	NOTSE
304890	Commissioning17	/Express/Commissioning2017/DQM	Thu 12-10-17 20:31:41	Dataset Trigger	BAD!	STANDBY	EXCLUDED	STANDBY	GOOD	GOOD	GOOD	GOOD	GOOD	NOTSET	NOTSE
304884	Commissioning17	/Express/Commissioning2017/DQM	Thu 12-10-17 19:53:40	Dataset Trigger	BAD!	STANDBY	EXCLUDED	STANDBY	GOOD	GOOD	GOOD	GOOD	GOOD	NOTSET	NOTSE

Offline Datasets — Run# 304725, /PromptReco/Cosmics17E/DQM — Details | Lumi sections | Manage | Move — Refresh | Configure | Export — 23,654 items. Show 20 from 1 to 20. Page 1 / 1,183

Ru...	Run Class ...	Dataset Name	Datase...	Dataset Created	Last Shifter	Cms	Castor	Csc	Dt	Ecal	Es	Hcal	Hlt	L1t	L1tmu	L1tcalo	Lumi	Pix	Rpc	Strip
304728	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:23	DQMGUI T...	BAD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304727	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:20	DQMGUI T...	BAD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304725	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:20	DQMGUI T...	BAD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304724	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:19	DQMGUI T...	BAD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304709	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:18	DQMGUI T...	BAD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304705	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:15	DQMGUI T...	BAD	EXCLUDED	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304701	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:14	DQMGUI T...	BAD	EXCLUDED	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304700	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:12	DQMGUI T...	BAD	EXCLUDED	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304697	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:10	DQMGUI T...	BAD	EXCLUDED	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304688	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:08	DQMGUI T...	BAD	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304686	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:06	DQMGUI T...	BAD	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304679	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:05	DQMGUI T...	BAD	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304675	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:02	DQMGUI T...	BAD	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	EXCLUDED	GOOD	GOC
304672	Collisions17	/PromptReco/Collisions2017E/DQM	OPEN	Fri 13-10-17 09:53:20	DQMGUI T...	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304665	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:39:00	DQMGUI T...	BAD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304663	Collisions17	/PromptReco/Collisions2017E/DQM	OPEN	Fri 13-10-17 09:53:14	DQMGUI T...	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304662	Collisions17	/PromptReco/Collisions2017E/DQM	OPEN	Fri 13-10-17 09:53:07	DQMGUI T...	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304661	Collisions17	/PromptReco/Collisions2017E/DQM	OPEN	Fri 13-10-17 09:53:01	DQMGUI T...	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304654	Collisions17	/PromptReco/Collisions2017E/DQM	OPEN	Fri 13-10-17 09:52:53	DQMGUI T...	GOOD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC
304652	Cosmics17	/PromptReco/Cosmics17E/DQM	OPEN	Fri 13-10-17 09:38:59	DQMGUI T...	BAD	EXCLUDED	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOC

Help

Common Offline Shift taker actions:

Waiting List table

- The runs in this table have been signed-off by online certification and are waiting for offline certification. The component status fields start out with the results of the online certification by default.
- Determine whether run is ready for offline certification, for example checking whether plots are available in DQM GUI (consult shift instructions). This may also be automated.
- Click a row to select it and enable more buttons:
  - Click *Details* to view meta information.
  - Click *Create Dataset* to bring up interface to start Offline certification.
    Once created, the dataset will appear in the lower table (if the lower table is set to display Offline Datasets).

Datasets table

- Click a row to select it and enable more buttons:
  - Click *Details* to view meta information.
  - Click *Manage->Edit* to return to certification screen.
  - Click *Move->to SIGNOFF* when done with certification.

Additional help:

Subsystem (i.e. component) text and color codes in both tables

- Status is shown as color (green = GOOD, red = BAD, yellow = STANDBY, grey = EXCLUDED)

is text (e.g.

ated by (e.g. UNDEF!).

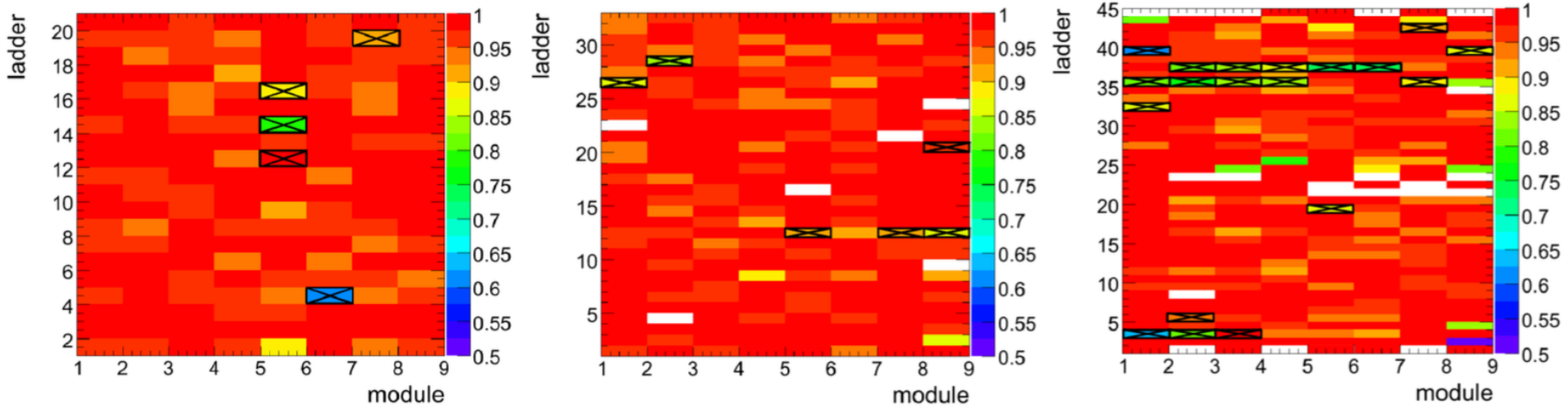
ntry to view the

Run Registry in 2017

- Automatically collects Run data
- Web interface for experts to manually set quality flags on data (GOOD/BAD)
- Provides APIs for scripts to produce final list of data ready for analysis (GoldenJSON)
- Aim to enable ML for these services

Limits of a Human-based DQM

- Problem spotting latency
- High manpower demand
- 24/7 shifts + training
- Occasional involuntary human errors
- Limit to the amount of quantities that a human can process in a finite time interval
- Transient problem can be overlooked during visual comparison
- Decision process depends on level of experience and understanding
- Changing running conditions
- Reference samples change
- Static thresholds do not scale
- Maintenance of shifter instructions
- ML can be used to develop mode intelligent tests
- checking relative position of dead ROCs



Towards ML-based DQM

- Apply recent progress in Machine Learning techniques to automate DQM
- To focus on the Online DQM
- To compare the performance of different ML algorithms.
- To compare fully supervised vs semi-supervised approach
- Impact the current workflow, make it more efficient and guarantee that the data is useful for physics analysis
- Reduce manpower to discriminate good and bad data, spot problems, save time examining hundreds of histograms by building intelligence to analyze data, raise alarms, quick feedback

- Make sure detector behaves well to perform sensible data analysis
- Implementing the best architecture for neural networks

  Underfitting - Too simple and not able to learn

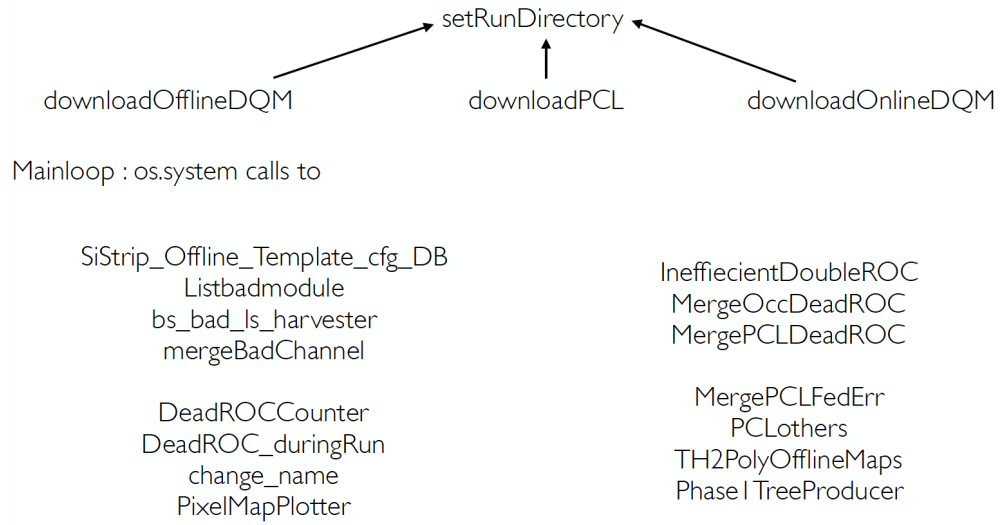
  Overfitting - Too complex and learns very specific and/or unnecessary features
- There is no rule of thumb for an ideal model

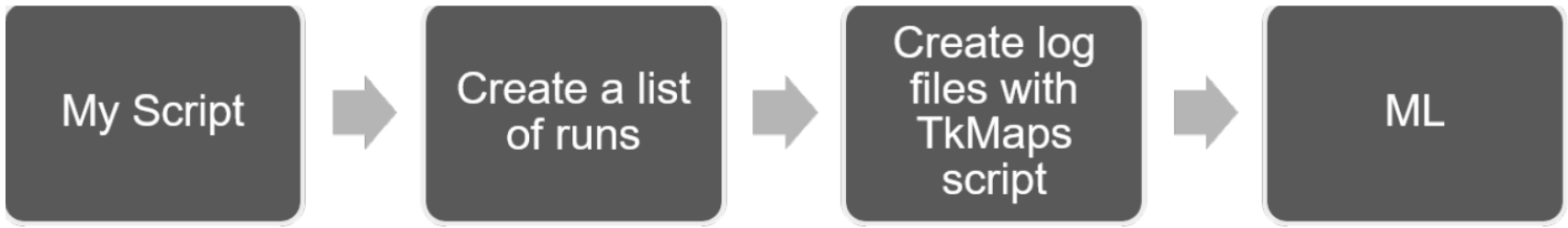
  Many, many, many. . . . . possible combinations.

Methodology

TkMaps Script

- This is work in progress
- The idea is to generate log files from a TkMaps Script and use the information as input for a ML model
- Automate this process of generating the log file using a python script to make list of available runs
- That script can be imported to the TkMaps script for easier readability and maintainability





My Python Script

ROOT/OfflineData/Run2018/ZeroBias

Up

0003257xx/ - 2018-11-04 18:07:20 UTC
0003256xx/ - 2018-11-04 12:37:28 UTC
0003255xx/ - 2018-11-02 21:45:48 UTC
0003254xx/ - 2018-11-01 17:25:14 UTC
0003253xx/ - 2018-10-31 06:12:03 UTC
0003252xx/ - 2018-10-29 06:39:30 UTC
0003251xx/ - 2018-10-29 18:36:09 UTC
0003250xx/ - 2018-10-26 09:47:05 UTC
0003249xx/ - 2018-10-26 07:15:25 UTC
0003248xx/ - 2018-10-21 23:13:14 UTC
0003247xx/ - 2018-10-20 10:18:42 UTC
0003246xx/ - 2018-10-17 14:05:59 UTC

- Made a python module that accesses and fetches the list of runs from a specific area called afs space (afs is file storage system and shown on left)
- This script will be part of TkMaps script (previous slide)
- Integrate the python module with TkMaps script
- Adapting the code to look for files in eos space (more efficient and scalable file storage system)
- Implement *argparse* for command line help for my python module
- Latest version of the script is in a new Github repo
- Runs on python 3

THANK YOU !!