



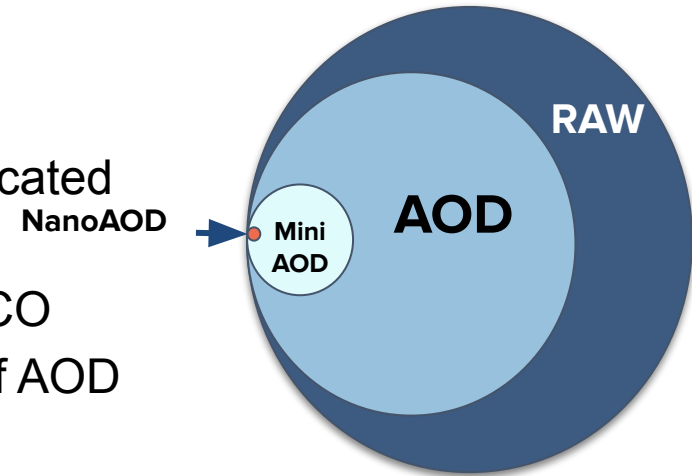
# CMS Analysis Facilities

Oksana Shadura, University of Nebraska-Lincoln  
on behalf of CMS O&C team

# Current Run-3 analysis model

*Main derived data formats for physics analysis in CMS:*

- RECO, ~2 MB/event - *RECO*nstructed data, used for dedicated studies and detector commissioning
- **AOD**, ~500 kB/event - Analysis Object Data, 40% of RECO
- **MiniAOD**, ~60 kB/event - lightweight data tier, 10-15% of AOD
- **NanoAOD**, ~1 kB/event - ntuple like format



relative sizes of CMS data formats

Each format also exists with similar event size for simulated data: **AODSIM**, **MiniAODSIM**, **NanoAODSIM**.

In addition, **specialized HLTSCOUT (~10 kB/event) and L1SCOUT (~360 kB/orbit) formats**.

HLT scouting information is now being included in the MiniAODSIM to avoid analysts' having to use AODSIM for searches using scouting data.

# Current Run-3 analysis model

- **The most common data format for analysis is NanoAOD, which contains high-level physics object information. It is estimated that over half of CMS analysis currently use NanoAOD.**
- MiniAOD is also in active use primarily as an input to custom data reduction steps, *where analysis codes have not migrated to the NanoAOD format. 30-40% of analyses still use MiniAOD (e.g. mostly through ntuple-like MiniAOD data products)*
- *AOD is rarely accessed (~10% as often as MiniAOD) but still made available (including automated tape recall) with CRAB. However, as AOD is 500kB/event, the total volume actively accessed by analysis is similar to that of MiniAOD.*



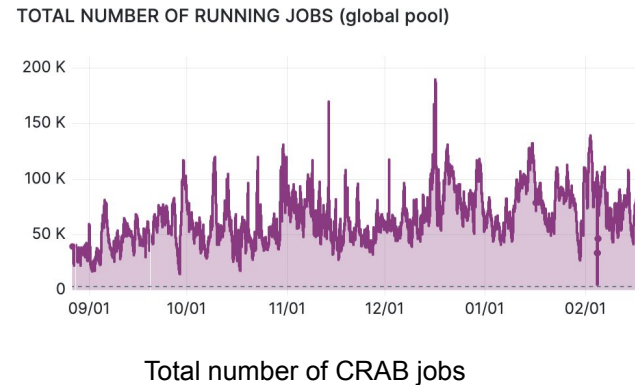
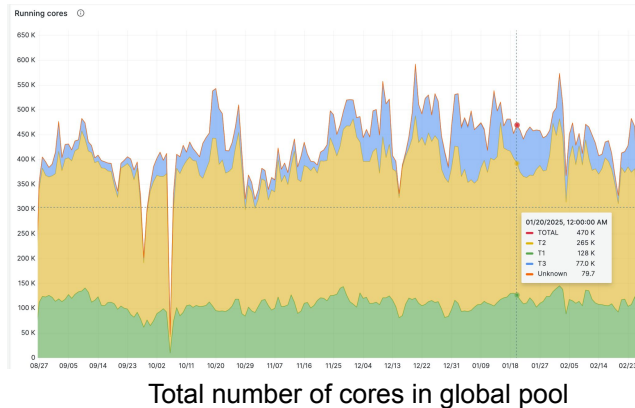
# CMS Remote Analysis Builder (CRAB)

CRAB is a utility that distributes CMSSW jobs to the CMS grid (typically, but not always, *using CPUs at the site where the input data is stored*).

By using CRAB user is able to:

- Access CMS data and Monte-Carlo which are distributed to CMS associated centres worldwide.
- Exploit the CPU and storage resources at CMS associated centres.

The jobs will then transfer the reduced output (e.g., skimmed/slimmed ntuples or even histograms) to user /store/user/ space.



# Evolution of the Analysis Infrastructure during Run-3

US CMS Facilities: FNAL EAF, Nebraska coffea-casa, MIT SubMiT, Purdue AF

German Facilities: DESY NAF

Italian Facilities: INFN AF

Spanish Facilities: CIEMAT AF

## Traditional resources:

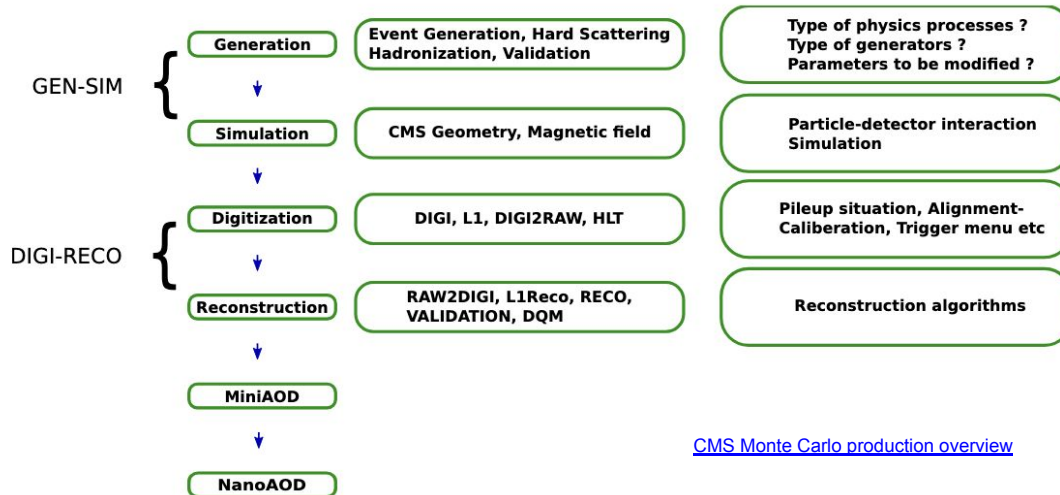
- Lxplus / Lxbatch, LPC
- Local university cluster
- Laptop

XCache



# Current Run-3 analysis model

- Main analysis workflows are:
  - **(Private) Signal MC production:** generation (gridpack, LHE, Pythia, etc.), detector simulation, digitization, reconstruction, reduction (Mini/NanoAOD); much signal MC is handled centrally but individual analysts also produce some additional samples.
  - Due to the large size of the outputs of some of these steps, GRID resources are most efficiently used if the steps are closely chained.



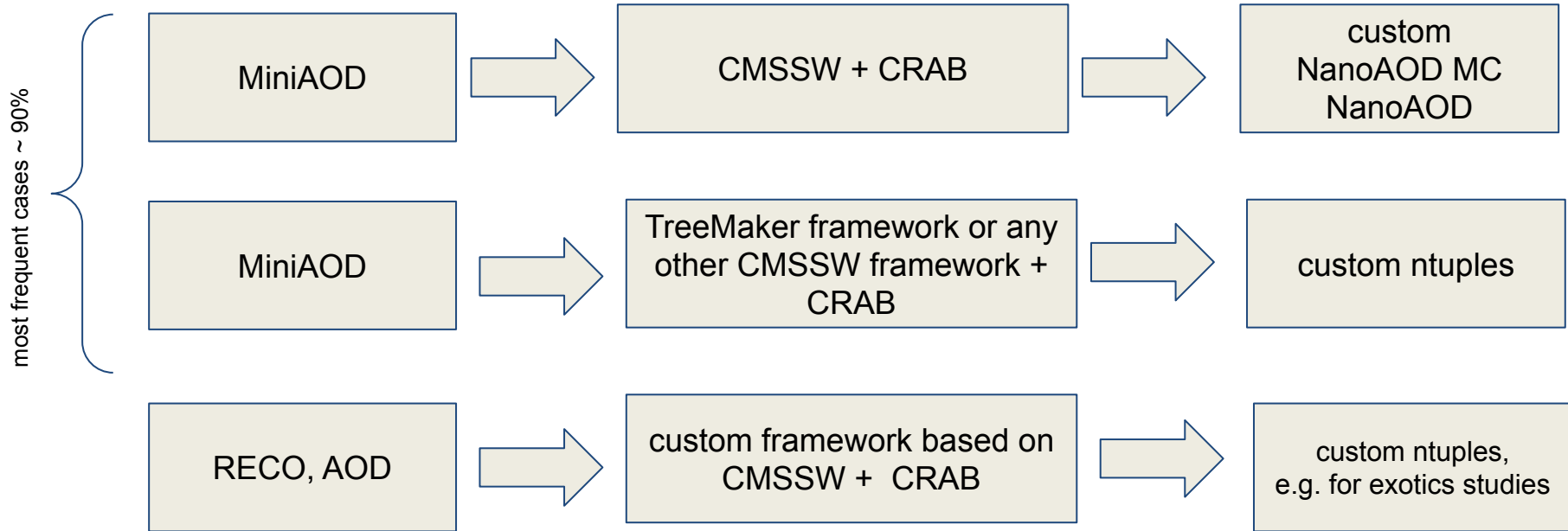
[CMS Monte Carlo production overview](#)

Various generators, software frameworks and services used:

- [Pythia](#), [Herwig](#), [Tauola](#).
- [Powheg](#), [Sherpa](#), [MadGraph5\\_aMCatNLO](#), [Alpgen](#).
- CMSSW framework (CMSDriver)
- CRAB

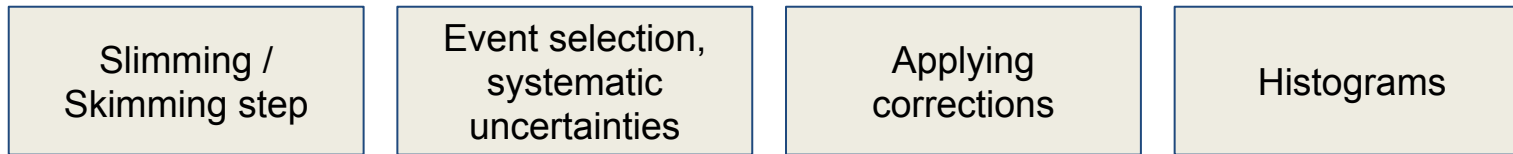
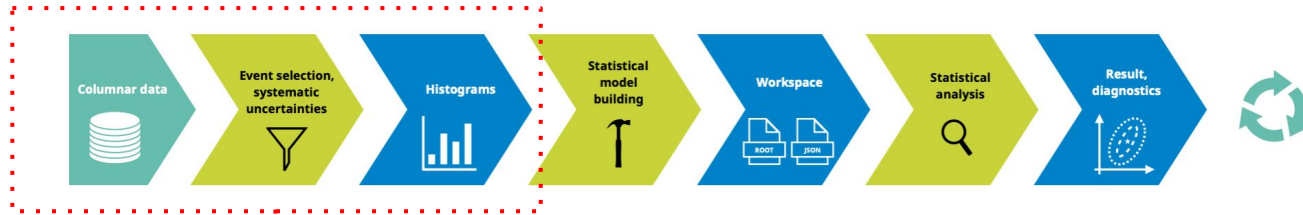
# Current Run-3 analysis model

- Main analysis workflows are:
  - **NTuple production:** read central MC and data, produce private format or (custom) nano



# Current Run-3 analysis model

- Main analysis workflows are:
  - **Primary analysis:** slimming/skimming, corrections, histograms

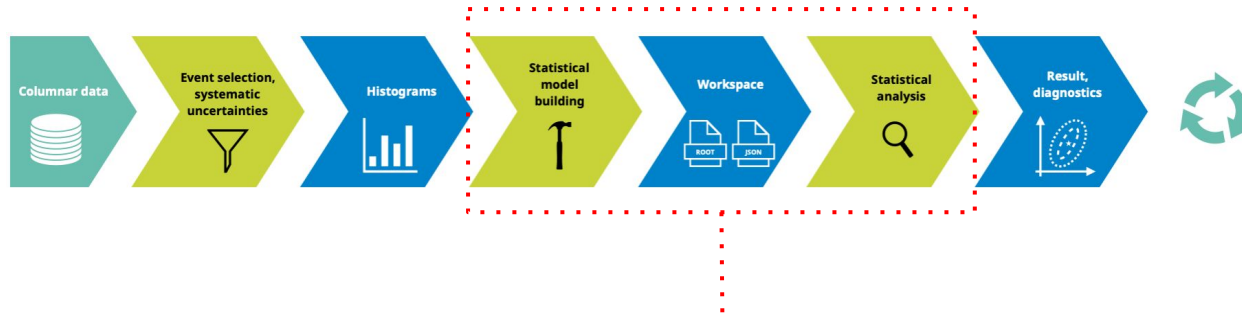


Skimming / slimming: the goal is to reduce the initial datasets by filtering suitable events and the selection of the interesting observables

**Software stack:**  
 Various analysis frameworks based on python tools and ROOT / RDataframe for NanoAOD format and custom ntuples, rarely CMSSW based frameworks for MiniAOD format

# Current Run-3 analysis model

- Main analysis workflows are:
  - **Interpretation: fitting (Combine)** - usually on aggregated data



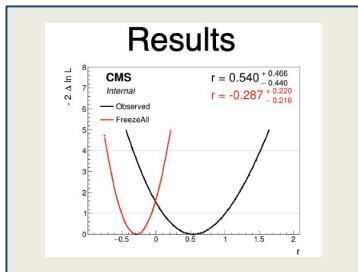
Create CMS statistical model via Combine datacards:

- Signal and background distributions
- Systematic uncertainties

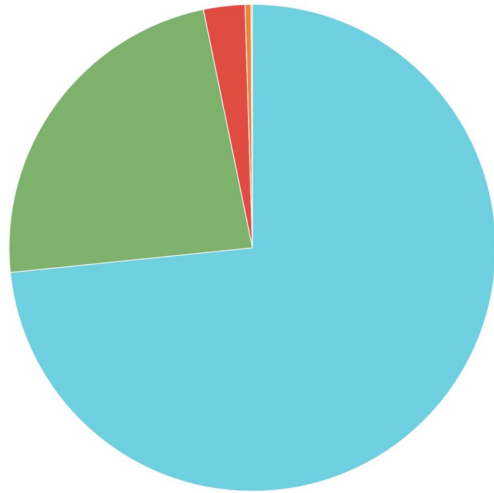
Convert them to RooFit workspace

**Combine:**

- Limit setting
- Significance / p-value calculation
- Confidence intervals ...



# Current Run-3 analysis model



**Analysis currently uses on average 20%+ of the CMS global pool compute used by CRAB**  
(last 90 days statistics)

## Challenges:

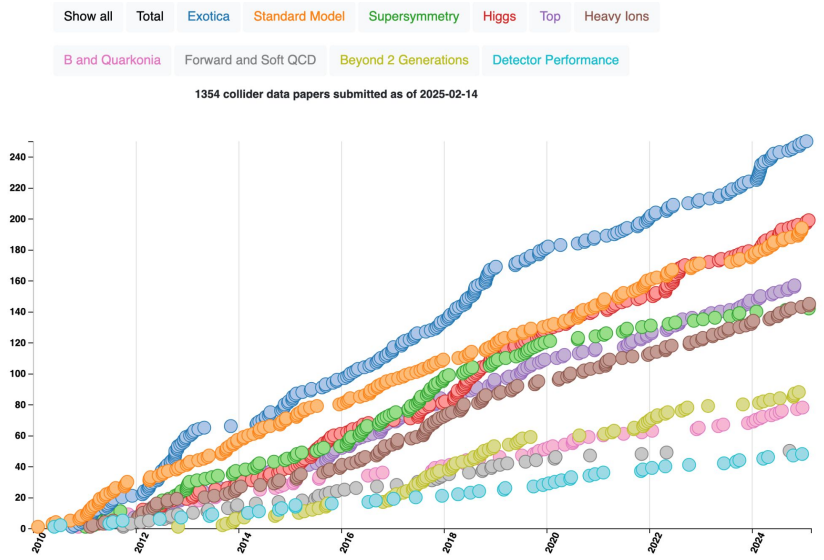
Analysis compute outside the CMS global pool is challenging to quantify

Streaming data transfers for analysis are challenging to quantify, as xrootd monitoring is not robust

Interactive resources are typically modest in comparison to the above and hard to quantify

# Current Run-3 analysis model

- Analysis is getting done, there are no major restrictions on what people are able to implement.



# Future analysis model in Run-4 and Run-5

We expect that **private signal Monte Carlo production in Run-4 will not scale**,  
in case it will be affected by CRAB sustainability

Linear scaling of **primary analysis step turnaround (3-5x) will be challenging** for human productivity  
and we have a very good idea how to beat linearity with caching

**Scale of ML training will be growing** larger and may require a different infrastructure

**More efficient low tier data access** (e.g. *provide a column-on-demand delivery service*),  
which will allow significantly to reduce duplication of data on disk and etc.

# Future analysis model in Run-4 and Run-5

Data Tier	Location	Lifetime	Replicas	Purpose
RAW	Tape	Permanent	1.5	(Re)reconstruction
Express FEVT	Disk	3 months	1	Monitoring
Express AlCaRECO	Disk	2 years	1	Alignment, Calibration
AOD(SIM)	Partial disk	Latest on tape	1	(Re)MiniAOD
MiniAOD(SIM)	Disk	Latest 3 on disk	2	Analysis, Nano creation
NanoAOD(SIM)	Disk	Flexible, on disk	Many	Analysis
HLTScout	Disk	Latest on disk	1	Analysis
L1Scout	Disk	Latest on disk	1	Analysis
Skims	Disk	6 months	1	Analysis
Pileup Libraries	Disk	As needed	1	Monte Carlo Production

# Future analysis model in Run-4 and Run-5

- **Investing in caches at Analysis Facilities**
  - Fractions are subject of R&D but not known yet
- **CMS is prioritizing the use of smaller data tiers by analysts**
- **We would like to keep batch and also interactive (both CPU and GPU) and the goal is to provide automated tools for managing batch infrastructure as a part interactive workflows**
- We will also need to provide network infrastructure commensurate to interactive timescales on expected datasets

# Future analysis model in Run-4 and Run-5

- Annual volume expected for the different data formats, both data and MC.
  - In Run 4, approximately 6e10 data and 14e10 simulation events per year are expected.

Current estimates for Run-4:

- **RAW event sizes** are **4.3 MB** and **5.9 MB** on average for 140 and 200 pileup events.
- Similarly, sizes for **AOD events** are **1.4 MB** and **2 MB**, and **180 kB** and **250 kB** for **MiniAOD**.
- The size of **NanoAOD events** is **4 kB**, independent of pileup.



**AOD - 280 PB**



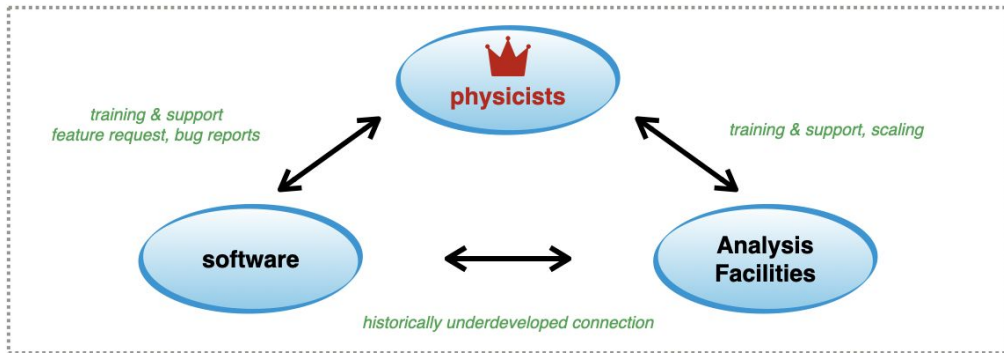
**MiniAOD - 36 PB**



**NanoAOD - 1 PB**

# Evolution of the Analysis Infrastructure

This is still a challenge, but CMS is exploring the user requirements via various approaches:



CMS user surveys:

[AF users UX feedback](#) designed for HSF WLCG workshop, [CMS TOP physics group survey](#), [CMS CAT Analysis Survey](#), [U.S. CMS S&C Ops Program - User Community Survey](#)

## [HSF Analysis Forum Whitepaper](#)

### Building blocks used for designing AFs

- Columnar analysis and support new pythonic ecosystem
- Efficient data delivery and data management technologies
- Machine learning services and tools
- Efficient data caching solutions
- Support for object storage
- Easy integration with scalable computing resources
- Modern authentication (IAM/OIDC), tokens, macaroons
- Modern deployment and integration techniques

# Evolution of the Analysis Infrastructure: new technologies to explore

## ***Changes in analysis paradigms & use of ML***

- Simulation-based inference
- Automatic differentiation
- Inference at HL-LHC scale

*“people use more and more ML in creative ways and that all costs significant amount of GPU resources”*

## ***A place of ML in infrastructure***

- The interface between analysis facilities and ML training needs further and broader consideration. Should this be built in as a first class operation, or only a specialized case offered at specific facilities? e.g. Hyperparameter scanning in particular is highly compute-intensive and benefits from central coordination at dedicated facilities.
- As ML becomes more commonplace, there needs to be **facility(s) with the right hardware to do large-scale training, which is something not readily available from the grid.**

## ***Data access***

**More actively investigate object stores** for fine-grained data access

## ***Network***

Interactive analysis turnaround will challenge the network infrastructure

# Evolution of the Analysis Infrastructure: next steps

[Blueprint workshop 15-16 May 2025](#) together with **ATLAS, CMS and IRIS-HEP**

to identify a representative set of physics analyses, described in terms of workflow and computational needs

Expected outcomes:

- **a survey** aimed to gather input for physics analysis examples
- an event **showcasing the examples of analyses**
- **a document summarizing these analyses**, alongside an extrapolation of how we expect them to evolve at the HL-LHC

# Evolution of the Analysis Infrastructure: current limitations

- The columnar analysis facility concept is based on idea to keep recently and/or frequently accessed columns in fast-access memory rather than having to cold start the analysis from disk every time, in order to allow rapid iteration and promote column sharing among analyzers/groups.
- Some interactive or machine learning workflows, in order to match the HL-LHC magnitude, will be not compatible with current Grid specs
  - e.g. [some systematic tests on Tier-2 current specs](#)

# Evolution of the Analysis Infrastructure: R&D

**CMS believes that specialized interactive resources for analysis are essential for the scientific productivity of HL-LHC researchers**

Exploring models where a central hub provides a seed of resources and scales out over heterogeneous resources based on user needs (INFN, CIEMAT AF, Coffea-casa, Purdue AF, EAF, SubMIT)

Scaling computing capabilities on AFs to HL-LHC rates: introducing *200 Gbps and 400 Gbps* challenges (Coffea-casa)

Hardware benchmarking activities (INFN, Purdue AF, SubMIT)

Working on the improving of “Analyst”  $\Leftrightarrow$  “Facility”  $\Leftrightarrow$  “Framework” *feedback loop*, based on Analysis Grand Challenge benchmark (**All CMS AF Facilities**)

Investigation and adoption of WLCG bearer tokens in AF (Coffea-casa early adopter)