



CMS Run-3 Model

Nick Smith, Fermilab

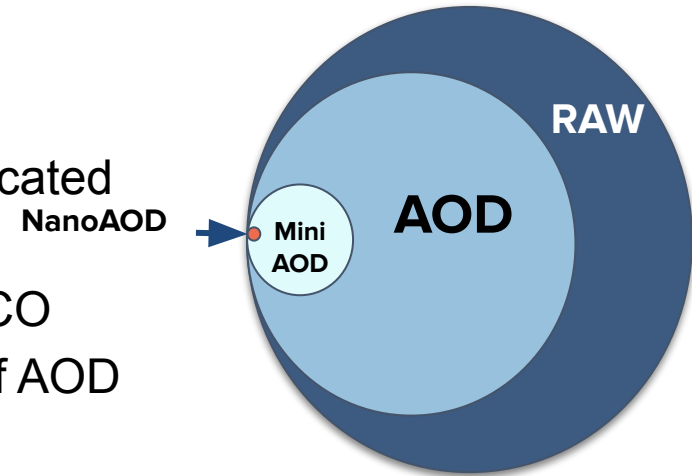
Oksana Shadura, UNL

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.



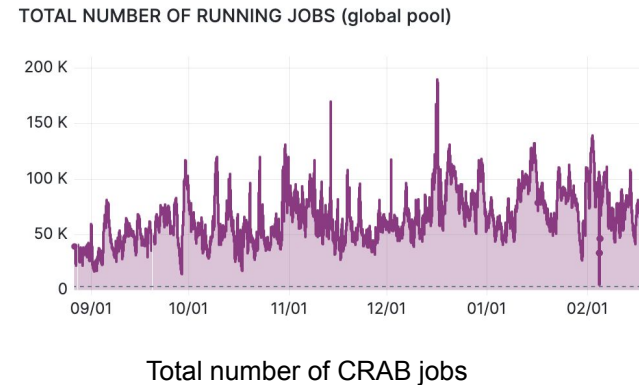
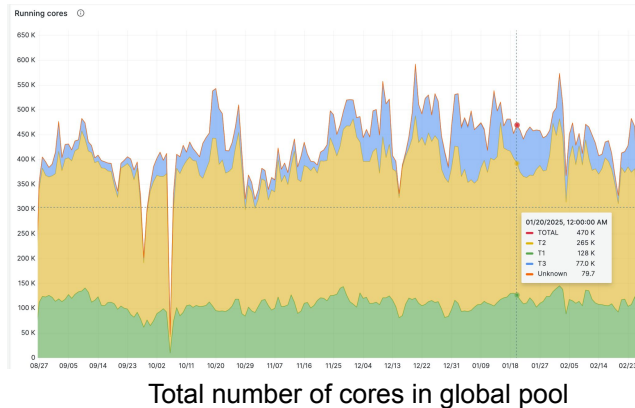
CMS Remote Analysis Builder (CRAB)

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:

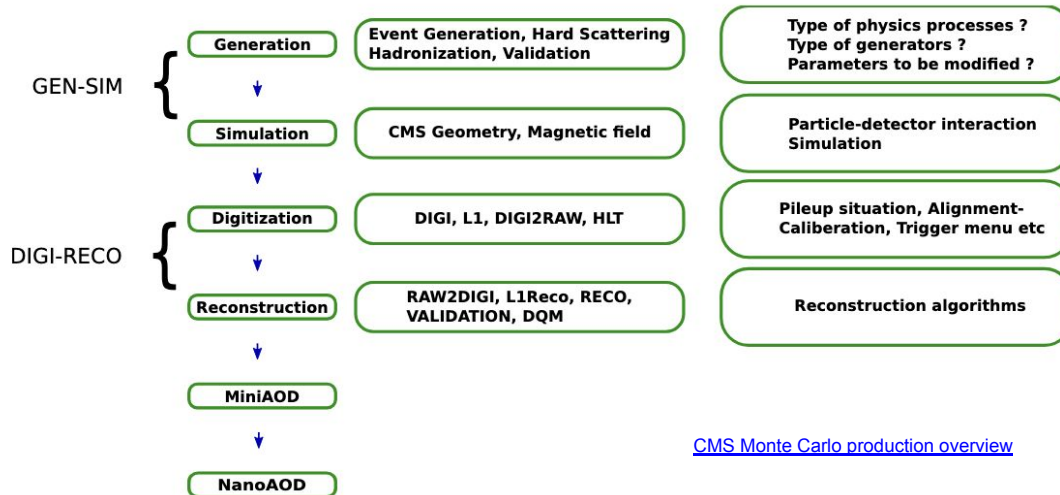
- Access CMS data and Monte-Carlo which are distributed to CMS associated centres worldwide.
- Exploit the CPU and storage resources at CMS associated centres.

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



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.



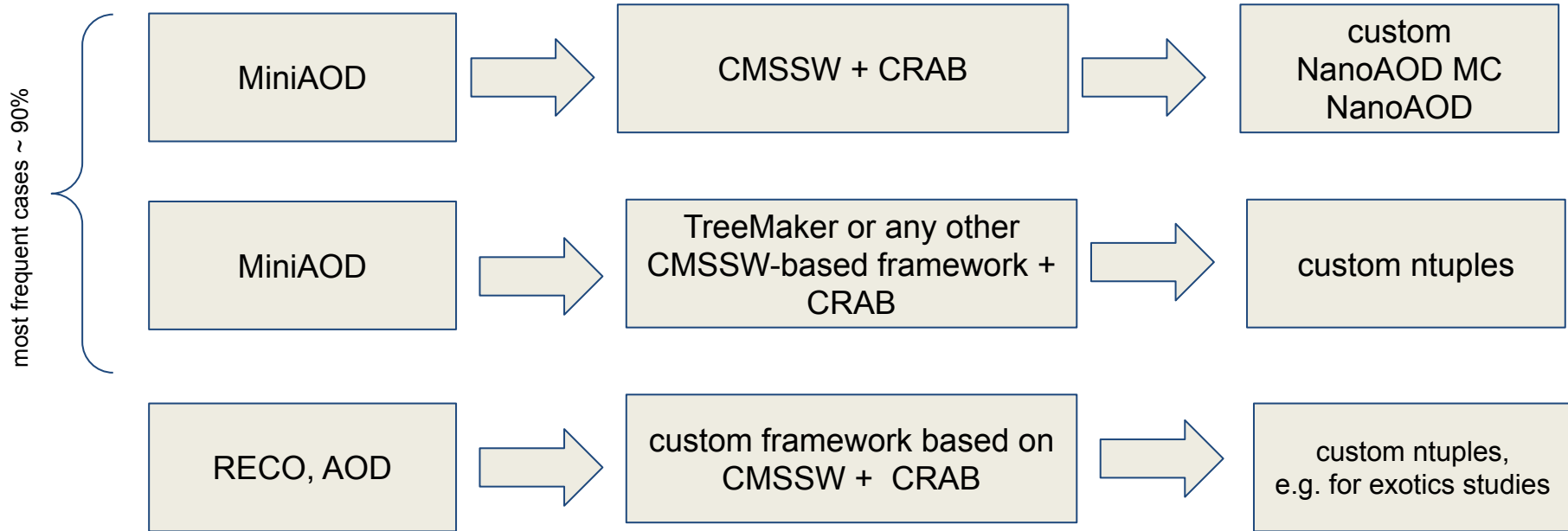
[CMS Monte Carlo production overview](#)

Various generators, software frameworks and services used:

- [Pythia](#), [Herwig](#), [Tauola](#).
- [Powheg](#), [Sherpa](#), [MadGraph5_aMCatNLO](#), [Alpgen](#).
- CMSSW framework (CMSDriver)
- CRAB

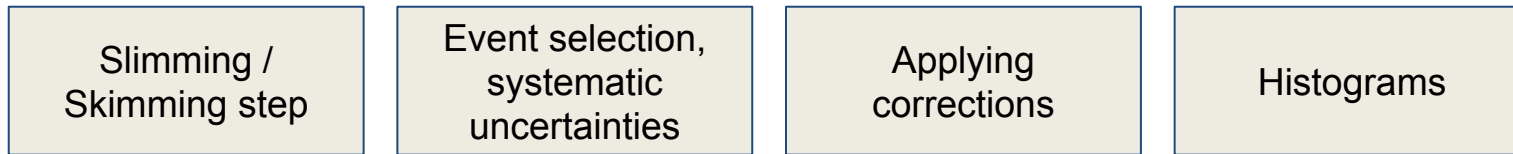
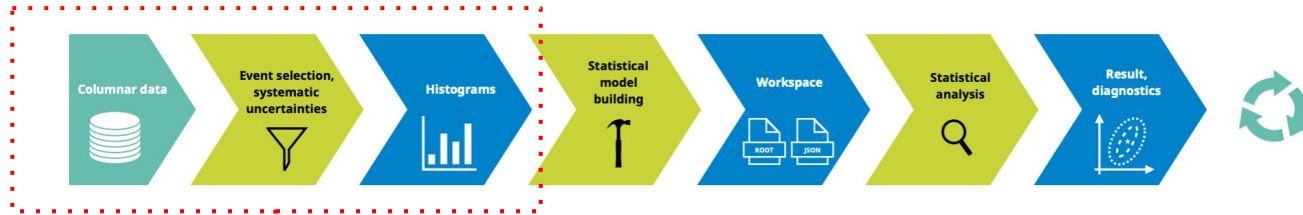
Current Run-3 analysis model

- 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

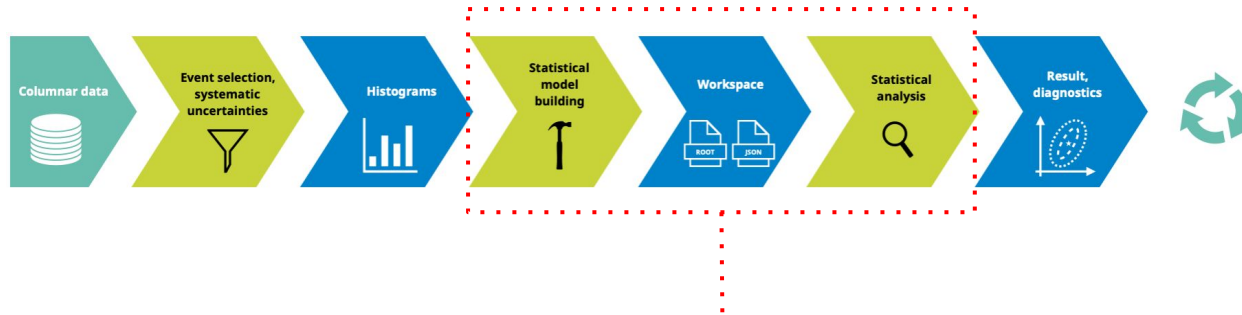


Skimming / slimming: the goal is to reduce the initial datasets by filtering suitable events and the selection of the interesting observables

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



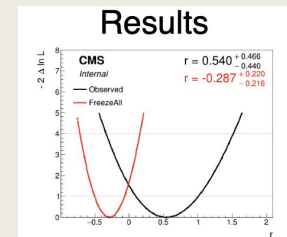
Create CMS statistical model via Combine datacards:

- Signal and background distributions
- Systematic uncertainties

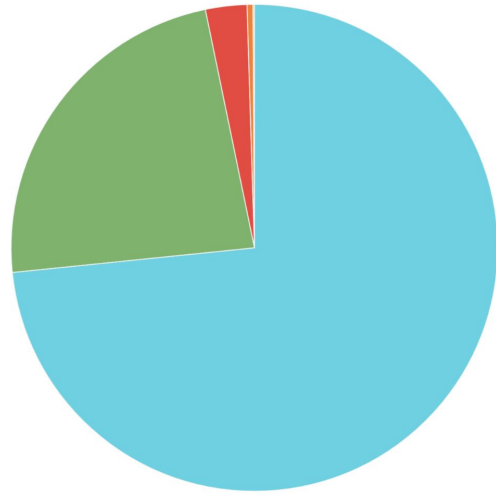
Convert them to RooFit workspace

Combine:

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



Current Run-3 analysis model



	total	percentage
production	683 Mil	73.4%
analysis	218 Mil	23.4%
tier0	26 Mil	2.7%
test	4 Mil	0.4%
domonit	881 K	0.1%

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

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

