nanoAODplus: news

Nano meeting

DESY, 6.5.2020

Achim Geiser, DESY Hamburg

- last items from last year (trigobj, jets) now fully integrated into NanoAnalyzer and tested on all platforms/in all configurations
  -> NanoAnalyzer.cc.jet2 no cleanup at all yet -> for experts only

- half-way cleaned-up “starting example setup” on 2011 QCD MC set is available as of today: (fully cleaned up version + more documentation will follow)

see /afs/desy.de/user/g/geiser/public/CMS/NanoAODplus/Nanoreadme.txt
(not really public, of course; comment #Compatibility if want full “plus” glory)

NanoAnalyzer.cc (now much shorter than before) is the only code you should need to understand/edit in order to make your EPR contribution (the rest is non-EPR).

- in friendly collaboration with QCD group, DY+jets will serve as a benchmark for run 1 nanoAODplus analysis in run 2 nanoAOD style

-> see contribution Armando

- is wednesday morning OK as new regular meeting slot?

Backup

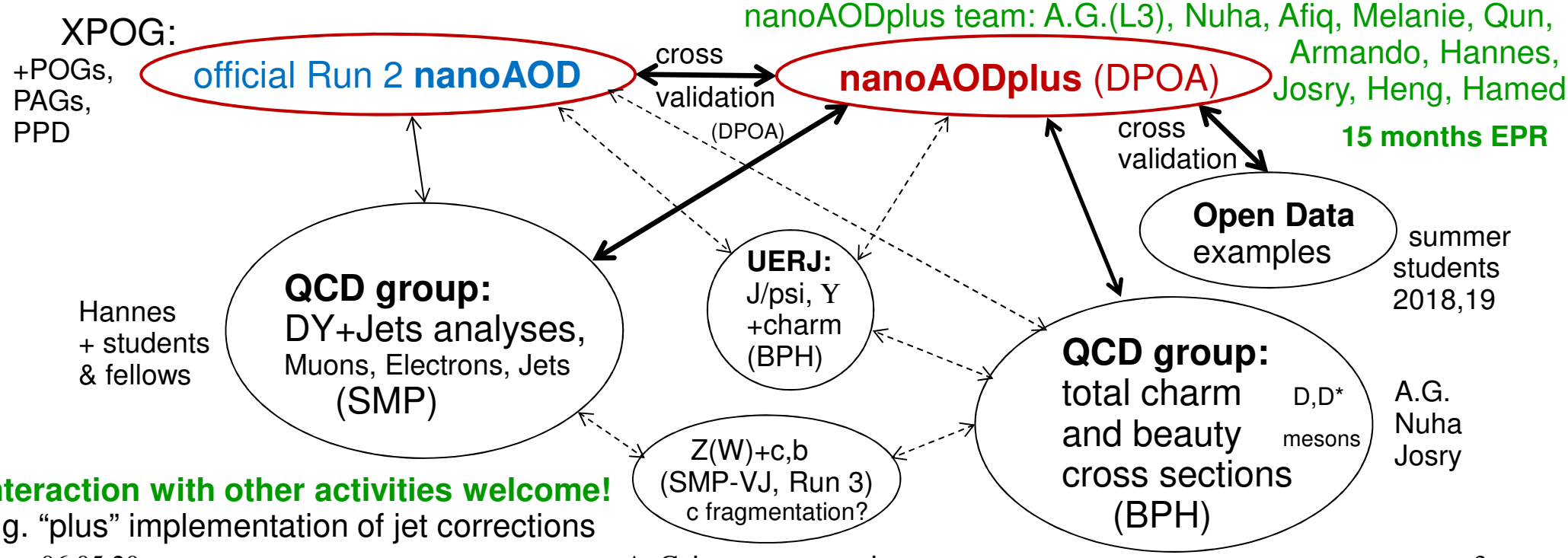
Data preservation and Open Data: nanoAODplus legacy data format for Run 1

Achim Geiser, shown at DESY physics planning meeting, 20.4.20

nanoAOD (created from miniAOD) is **currently advocated** and future default **format for Runs 2 and 3.** ~20% of analyses use nanoAOD so far, many enhanced by additional information **from miniAOD**.

Run1: miniAOD does not exist -> create **nanoAODplus** format directly **from AOD**, functionally consistent with **nanoAOD**; **plus** additional AOD variables and extensions

monthly meetings



interaction with other activities welcome!
e.g. “plus” implementation of jet corrections

NanoAODplus for Run 1, status and plans

reminder of DPOA meeting, 7.4.2020

Achim Geiser, DESY Hamburg, for the nanoAODplus team

- job ~half way done by end 2019, according to plan (design, implementation, validation); work has picked up again for this year.
- goal: finalize 1st practically usable/releasable version by end of this year (2020) (in accordance with original plan). Actual pioneer usage for physics has started. Potential further iteration(s) according to feedback.

  E.g. better documentation/readability. More flexibility for future developments.
- ~10 people have already expressed commitment to contribute at various levels this year (significant overlap with team of last year). Many already pledged.
  -> 15 FTE-months covered
- plan:

  give final touches to implementation of muons, primary vertices, and triggers;
  complete implementation of electrons, jets, MET, secondary vertices, and generator information;
  (leave analysis-level implementation of e.g. photons and taus for later iteration)
  add "plus" variables (AOD variables, non-nanoAOD higher level objects) according to need/request
  improve readability and flexibility of code, improve separation of "nanoAOD" and "plus" parts.

Tentative list of tasks/contents for 2020

- nanoAOD ntuple content (9_4 v2): fully done , partially done, already useful/used for analysis, being implemented , other

	variables	implemented	content implemented	content validated	remaining work
run/event/lumis.	3	100%	100%	100%	done
Generator /PSweight	11	-	-	-	~0.5 months,
PV /OtherPV /Pileup	14	70%	70%	50%	~0.25 months, half done
SV	13	100%	10%	-	~2 months,
GenPart	9	100%	50%	20%	~0.25 months, advanced
Muon	35	100%	80%	75%	~0.5 months, almost done
Electron	48	55%	50%	45%	~3 months, half done
Photon	28	25%	25%	-	~2 months, !not yet covered! (use H->gg?)
Tau	38	25%	25%	-	~3 months, !not yet covered! (use H->tau?)
IsoTrack	13	100%	-	-	~0.5 months, -> less
GenDressedLepton + GenVisTau	14	-	-	-	~0.5 months, !not yet covered!
Jet+FatJet +SubJet +SoftActivityJet	79	10%	10%	-	~3 months,
GenJet +GenJetAk8	14	-	-	-	~0.5 months,
MET+TkMET +CaloMET +RawMET+PuppiMET	23	30%	30%	-	~0.5 months ,
TrigObj	11	55%	45%	20%	~1 month, advanced
HLT	569	100%	100%	100%	done
LHEPart +LHEPdfWeight + LHEScaleWeight+LHEWeight_originalXWGTUP	11	-	-	-	~0.5 months -> less
Flag	26	100%	-	-	~0.5 months, !not yet covered!
Various other	10	-	-	-	~1 month, !not yet covered!

- implementation of nanoAOD header — ~0.5 months,
- compatibility with nanoAOD tools — ~0.5 months ,
- coordination + set up & manage twiki + git repository — ~1 month, -> more
- various general technicalities of setup , improve code rmoduklarity/eadability — ~0.5 months, -> more

total

3 months 2018
12 months 2019
15 months this year (2020)

names are assigned

~7 not covered/next year? (tbc)

Discussion of strategy issues

plan to "publish" nanoAODplus ntuples (to DBS or similar) within CMS using Rucio (technically not possible with current setup, access only via xrootd)
-> in contact with Rucio people; currently a bit of delay in Rucio project w.r.t. originally announced schedule.

original concept was Open Data release of nanoAODplus as a "one time legacy action" together with ultralegacy nanoAOD from Run 2.
Currently looks like

- **Run 2 nanoAOD will not be released anytime soon (not before 2022?)**
- **likely(?) no single clear "legacy" nanoAOD version anytime soon, rather continous further development of Run 2 nanoAOD? (well into Run 3?)**
- **practical usage of Run 2 nanoAOD increasingly moving towards external amendments from miniAOD (or even AOD) -> will there be many different "flavours" of nanoAOD?**

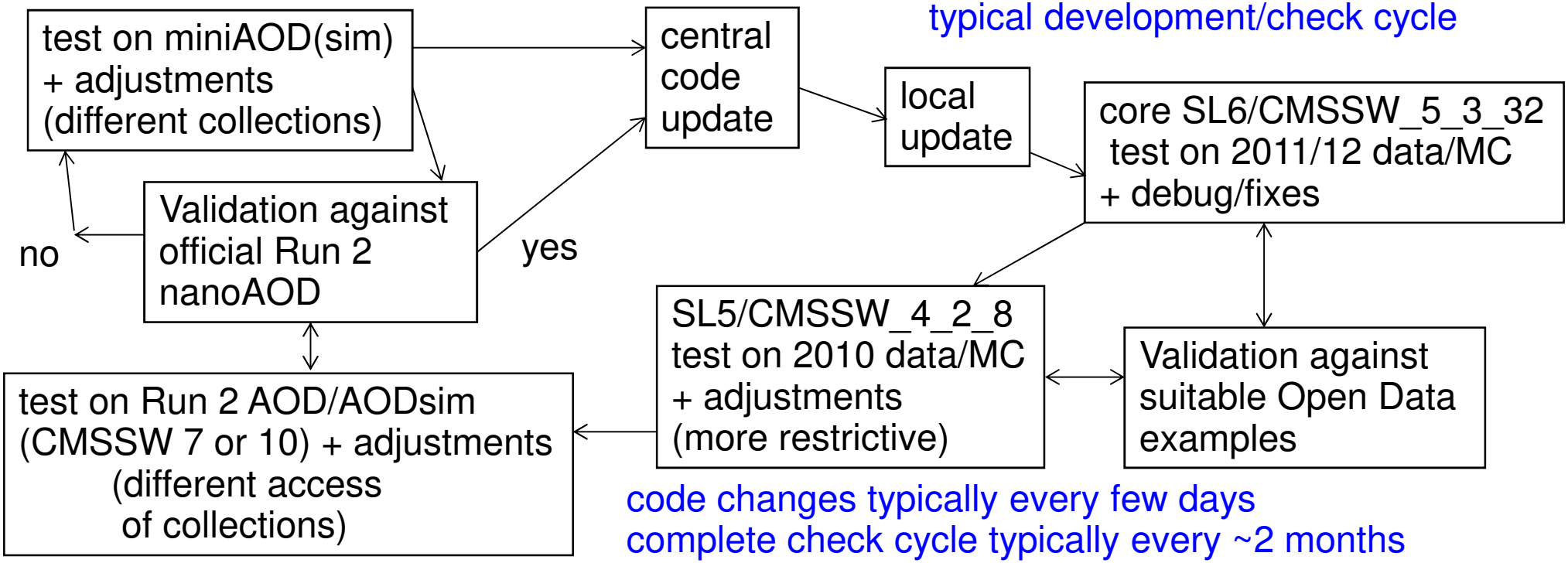
-> nanoAODplus might need to be organized more flexibly than planned so far
-> delay some of the content developments in favour of such a more flexible approach?
-> aim for release of more than one version of nanoAODplus in time, following evolution of nanoAOD?

(latest news (s.Donato): extensions of "legacy" CMSSW versions might technically not be excluded a priori after all, even though currently not politically favoured)

some technicalities of nanoAODplus implementation and checking cycle

single set of C++ and preprocessor core code for all input data types (AOD, AODsim, miniAOD, miniAODsim), legacy CMSSW versions (4,5,7,10), compilers (SL5,SL6,EL7) and run environments (interactive, batch, VM/container, crab)
-> core directory in git repository (exists already since last year)
-> any code change will automatically be propagated to all use cases (important for central development)

different CMSSW environments (cmsenv) and different configuration file setups for different data/MC sets and running environments (development, validation, production).
-> need multiple parallel setups, orthogonal to code updates (not obvious, concept of implementation in git still under development, greatly simplify for actual release)



Conclusions and Outlook

nanoAODplus data format for Run 1 making progress

subtasks defined (see slide) and person power (EPR) for this year being assigned (team of ~10 people part time), increase emphasis on flexibility for future developments

-> complete usable nanoAODplus ntuple for Run 1 by end of 2020, (first version), in parallel to Run 2 super-legacy processings

-> all legacy data should be analysable in nanoAOD(plus) format with the same CMSSW-independent Root analysis code, and (where possible) with the same variable content

eventually available as Open Data in addition to AOD
-> easier for outsiders to do physics analysis compared to current Run 1 AOD

Tentative list of tasks/contents for 2020, to be updated

- nanoAOD ntuple content (9_4 v2): fully done , partially done, already useful/used for analysis, being implemented , other

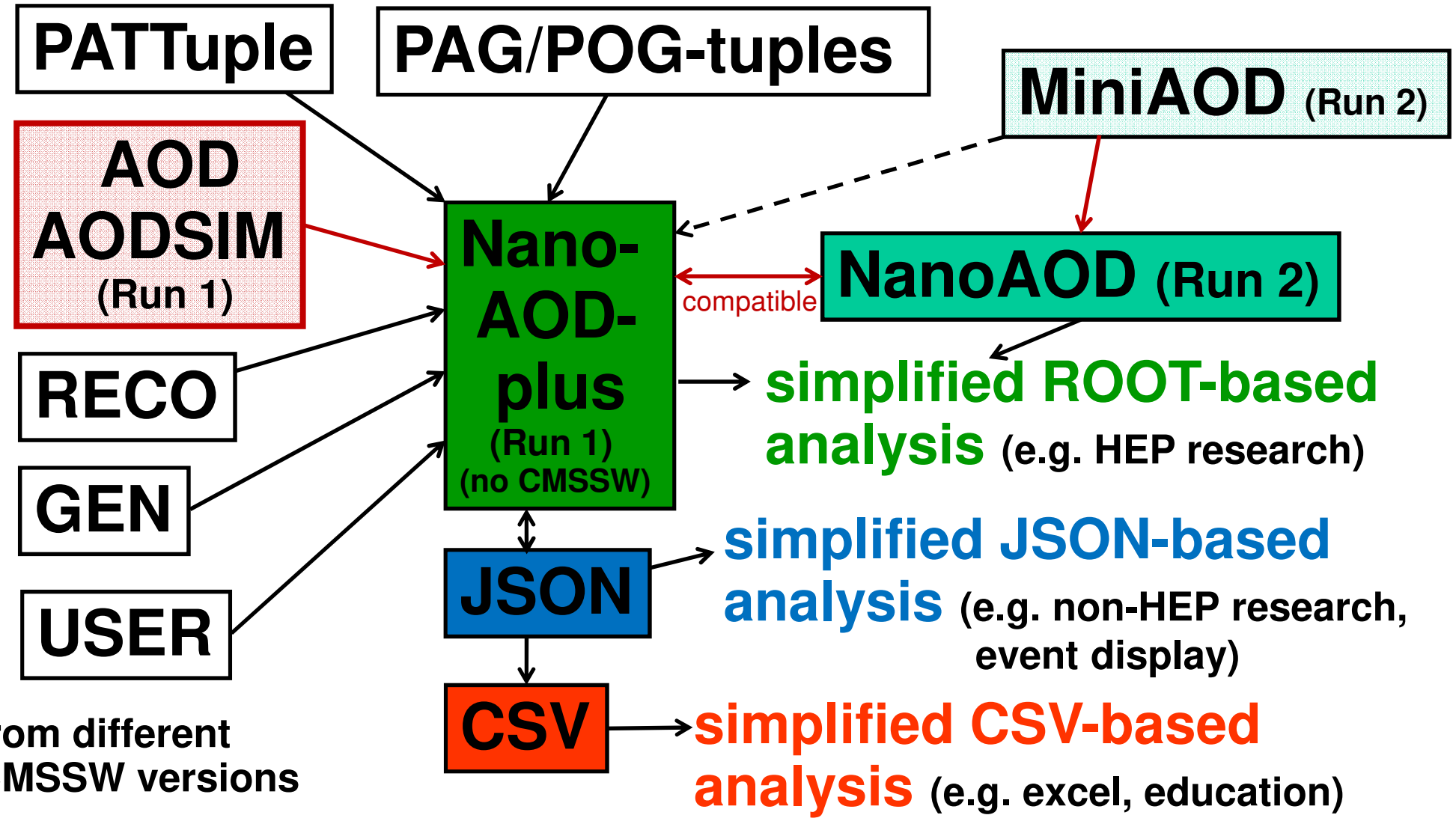
	variables	implemented	content implemented	content validated	remaining work
run/event/lumis.	3	100%	100%	100%	done
Generator /PSweight	11	-	-	-	~0.5 months, Melanie? (Hannes??)
PV /OtherPV /Pileup	14	70%	70%	50%	~0.25 months, Achim half done
SV	13	100%	10%	-	~0.5+0.5 months, Josry, Achim
GenPart	9	100%	50%	20%	~0.25 months, Achim advanced
Muon	35	100%	80%	75%	~0.5 months, Achim almost done
Electron	48	55%	50%	45%	~2.5(+0.5) months, Melanie (+Nuha) half done
Photon	28	25%	25%	-	~2 months, !not yet covered! (use H->gg?)
Tau	38	25%	25%	-	~3 months, !not yet covered! (use H->tau?)
IsoTrack	13	100%	-	-	~0.5 months, Achim
GenDressedLepton + GenVisTau	14	-	-	-	~0.5 months, N.N.
Jet+FatJet +SubJet +SoftActivityJet	79	10%	10%	-	~3.5 months, Armando? Heng? Josry? AG?
GenJet +GenJetAk8	14	-	-	-	~0.5 months, N.N.
MET+TkMET +CaloMET +RawMET+PuppiMET	23	30%	30%	-	~0.5 months , N.N.
TrigObj advanced	11	55%	45%	20%	~0.5+0.5 month, Achim, Qun?
HLT	569	100%	100%	100%	done
LHEPart +LHEPdfWeight + LHEScaleWeight+LHEWeight_originalXWGTUP	11	-	-	-	~0.5 months Josry ? (Hannes??)
Flag	26	100%	-	-	~0.5 months, !not yet covered!
Various other	10	-	-	-	~1 month , !not yet covered!

- implementation of nanoAOD header — ~0.5 months, Hamed?
- nanoAOD tools — ~0.5 months , Nuha
- coordination + set up & manage twiki + git repository — ~1 month, Achim
- various general technicalities of setup — ~0.5 months, Afiq

total — ~26 months 2019/20
~15 this year
~6.5 not covered/next year (tbc)

Thoughts about simplified DPOA data format: CMS

Design common flat ntuple format for all datasets (remove CMSSW dependence)



Motivation/goals for nanoAODplus format for Run 1

- **independence from `old' CMSSW versions** (or CMSSW in general)

  -> analysis in non-CMS environment,

  no need for virtual machines or container encapsulation

- **CMS members can run Run 2 nanoAOD-based analyses also on Run 1 legacy data and vice versa with same code**
  (also outsiders once Run 2 data will be released as Open Data)

  -> **identical** nanoAOD variable **names**

  -> **same variable content** **(as much as possible)**

  -> task:

  recode Run 2 algorithms for nanoAOD content directly from basic AOD variables, such that they work for CMSSW 4_2_8, 5_3_32 (Run 1 legacy), as well as 7_X (2015, no nanoAOD so far) and 8_X/9_X/10_X (for cross-validation with official Run 2 nanoAOD)

- **Add specific Run 1 variables ("plus")**

Technical implementation of ntuple production

- **EDanalyzer (NanoAnalyzer) which compiles/runs**
  in VMs or containers (for SL5)
  on DESY Tier 2 farm (for SL6/SL7)
  or with CRAB (via containers for SL5)

  for technicalities
  see presentation N. Jomhari
  at DPOA meeting March 13

  (single code, different configurations, differences between CMSSW versions accounted for via #ifdef flags)

- **Input is AOD (working on miniAOD interface for debugging)**

- **Implement Run 2 nanoAOD algorithms (according to workbook) on Run 1 AOD whenever technically possible**

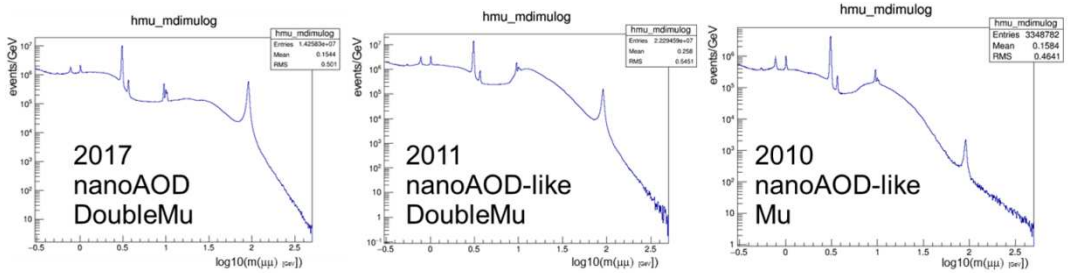
- **In addition, implement legacy Run 1 algorithms (extra variables, according to legacy workbooks) whenever useful (plus some further variables)**

- **Output is flat Root ntuple with nanoAOD variables, currently accessible on DESY dcache via XRootD (working on DBS publication option)**

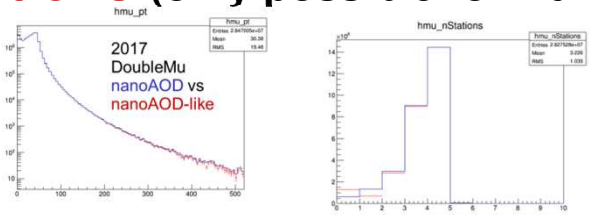
- **Twiki Documentation (under development) :**
  https://twiki.cern.ch/twiki/bin/viewauth/CMS/DPOANanoAODlike

Validation tools and strategy

- **Indirectly compare some physics distributions for different datasets**
  examples see presentation at fall C&O meeting:



- **Directly compare technical distributions (only possible for Run 2)**
  examples see presentation at fall C&O meeting:



- **New: Use BuildIndex and Friend functions of Root to compare nanoAOD and nanoAOD-like variables event-by-event, even if input event sets only partially overlap and events occur in different order (only possible for Run 2)**
  (thanks to A. Ricci and J. Metwally for support!)

  -> can validate and debug exactly

- **Exactly reproduce some known/well-validated Run 1 distributions from nanoAOD-like ntuple**

Conclusions and Outlook

nanoAOD-like data format for Run 1 making progress,
now organised through dedicated DPOA tasks
-> strengthen interaction with XPOG, POGs/PAGs, and PPD

tasks defined (see backup) and person power (EPR) for this year being assigned
(team of ~10 people part time) ***** today's meeting! *****

-> hope to complete nanoAODplus ntuple for Run 1 by end of 2020,
in parallel to Run 2 super-legacy processings

-> all legacy data should be analysable in nanoAOD(like) format
with the same CMSSW-independent Root analysis code,
and (as much as possible) with the same variable content

eventually available as Open Data together with AOD/miniAOD
-> easier for outsiders to do analysis compared to current Run 1 AOD

plans

nanoAOD-like data format for Run 1 making progress,

first actual applications in sight

-> hope to complete for Run 1 within next two years,
in parallel to Run 2 super-legacy processings
-> all legacy data should be analysable in nanoAOD(like) format

current situation:

Table:	0.9	2.76	5	7	8	13 TeV
ep	ZEUS*					
pp	2010/17	2010/13	2015/17	2010/11	2012	2015
						2016/17/18
pPb			2012/13/15		2016	
PbPb		2010/11	2015			

*external project in preparation

AOD
nanoAOD available
miniAOD available
RECO only