

nanoAOD-like ntuple for Run 1: status and plans

DESY, 16.5.19, collection of slides from previous DPOA and C&O meetings

DPOA session at C&O week

CERN, 2.4.2019

Achim Geiser, DESY Hamburg

Nur Zulaiha Jomhari, DESY Hamburg

(more people joining)

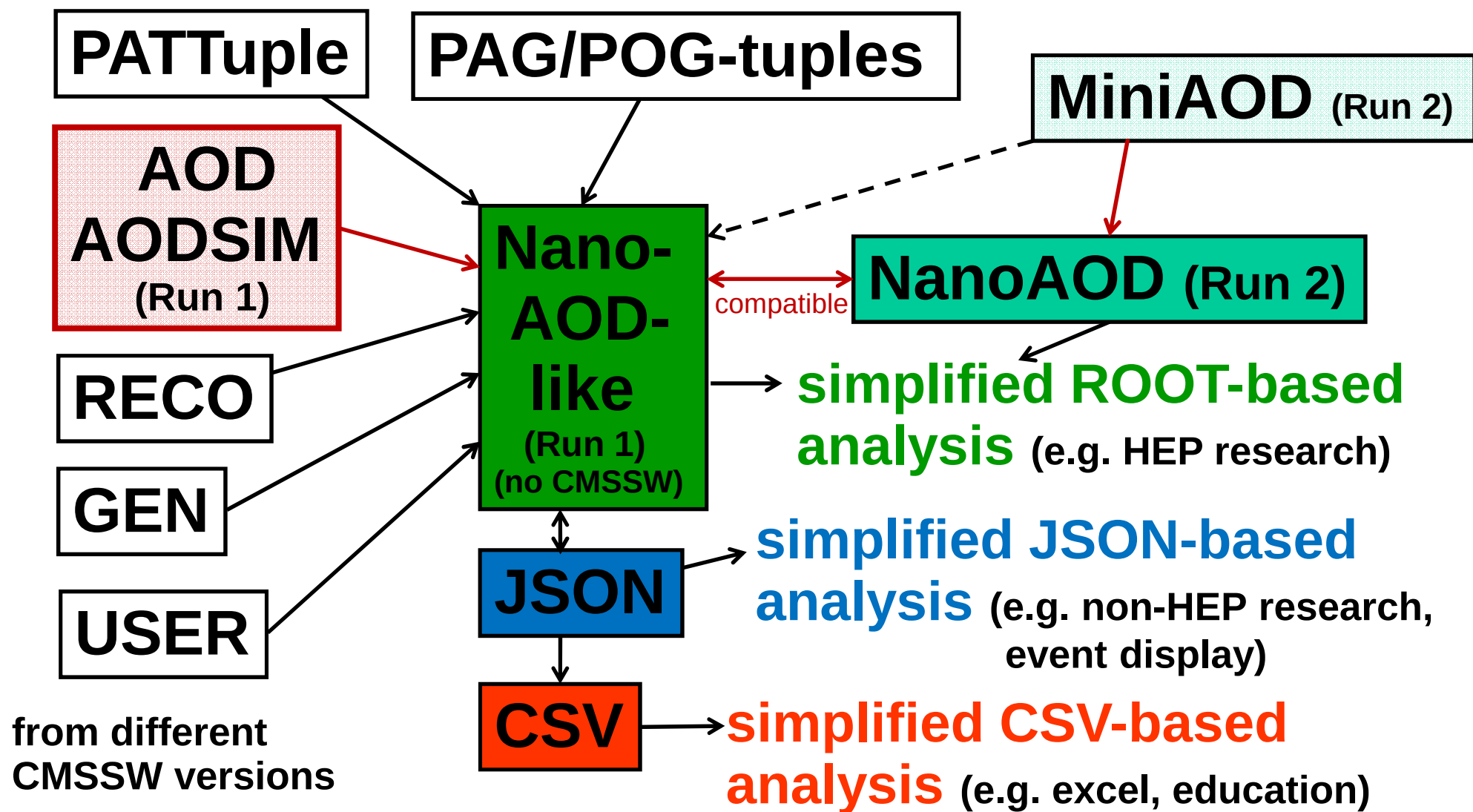


- Reminder of motivation/goals
- Status of technical implementation + validation
- Conclusions and Outlook

Brief summary, more in previous DPOA and C&O week presentations

Thoughts about simplified DPOA data format: CMS

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



Motivation/goals for nanoAOD-like 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)

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)**
(single code, different configurations, differences between CMSSW versions accounted for via #ifdef flags)
for technicalities see presentation N. Jomhari at DPOA meeting March 13
- **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>

Reminder: implementation of nanoAOD muon variables

see talk Nuha on DPOA meeting 21.8.18

Muon [back to top]

	Object property	Type	Description
→	Muon_charge	Int_t	electric charge
→	Muon_cleanmask	UChar_t	simple cleaning mask with priority to leptons
→	Muon_dxy	Float_t	dxy (with sign) wrt first PV, in cm
→	Muon_dxyErr	Float_t	dxy uncertainty, in cm
→	Muon_dz	Float_t	dz (with sign) wrt first PV, in cm
→	Muon_dzErr	Float_t	dz uncertainty, in cm
→	Muon_eta	Float_t	eta
→	Muon_genPartFlav	UChar_t	Flavour of genParticle for MC matching to status==1 muons: 1 = prompt muon (including gamma*->mu mu), 15 = muon from prompt tau, 5 = muon from b, 4 = muon from c, 3 = muon from light or unknown, 0 = unmatched
	Muon_genPartIdx	Int_t(index to Genpart)	Index into genParticle list for MC matching to status==1 muons
→	Muon_highPtId	UChar_t	high-pT cut-based ID (1 = tracker high pT, 2 = global high pT, which includes tracker high pT)
→	Muon_ip3d	Float_t	3D impact parameter wrt first PV, in cm
→	Muon_isPFcand	Bool_t	muon is PF candidate
→	Muon_jetIdx	Int_t(index to Jet)	index of the associated jet (-1 if none)
→	Muon_mass	Float_t	mass
→	Muon_mediumId	Bool_t	cut-based ID, medium WP
→	Muon_miniPFRelIso_all	Float_t	mini PF relative isolation, total (with scaled rho*EA PU corrections)
→	Muon_miniPFRelIso_chg	Float_t	mini PF relative isolation, charged component
→	Muon_mvaTTH	Float_t	TTH MVA lepton ID score
→	Muon_nStations	Int_t	number of matched stations with default arbitration (segment & track)
→	Muon_nTrackerLayers	Int_t	number of layers in the tracker
→	Muon_pdgId	Int_t	PDG code assigned by the event reconstruction (not by MC truth)
→	Muon_pfRelIso03_all	Float_t	PF relative isolation dR=0.3, total (deltaBeta corrections)
→	Muon_pfRelIso03_chg	Float_t	PF relative isolation dR=0.3, charged component
→	Muon_pfRelIso04_all	Float_t	PF relative isolation dR=0.4, total (deltaBeta corrections)
→	Muon_phi	Float_t	phi
→	Muon_pt	Float_t	pt
→	Muon_ptErr	Float_t	ptError of the muon track
→	Muon_segmentComp	Float_t	muon segment compatibility
→	Muon_sip3d	Float_t	3D impact parameter significance wrt first PV
→	Muon_softId	Bool_t	soft cut-based ID
→	Muon_tightCharge	Int_t	Tight charge criterion using pterr/pt of muonBestTrack (0:fail, 2:pass)
→	Muon_tightId	Bool_t	cut-based ID, tight WP
→	nMuon	UInt_t	slimmedMuons after basic selection (pt > 3 && track.isNonnull && isLooseMuon)

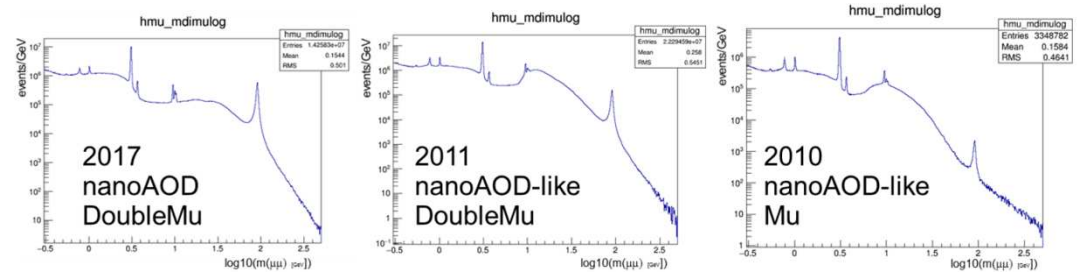
Meanwhile all are implemented

https://cms-nanoaod-integration.web.cern.ch/integration/master/mc94X_doc.html

Validation tools and strategy

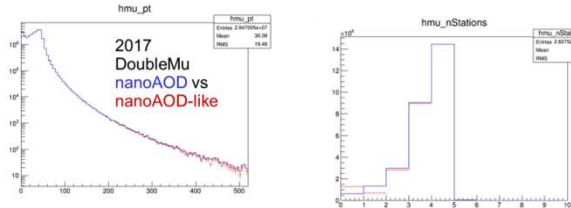
- Indirectly compare some **physics distributions** for different datasets
examples see presentation
at fall C&O meeting:

summer students
(+ DY studies??)



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

all developers



- 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** (some examples today) **all developers**

(currently ~30% of variables implemented, ~20% usefully filled, ~5% fully validated)

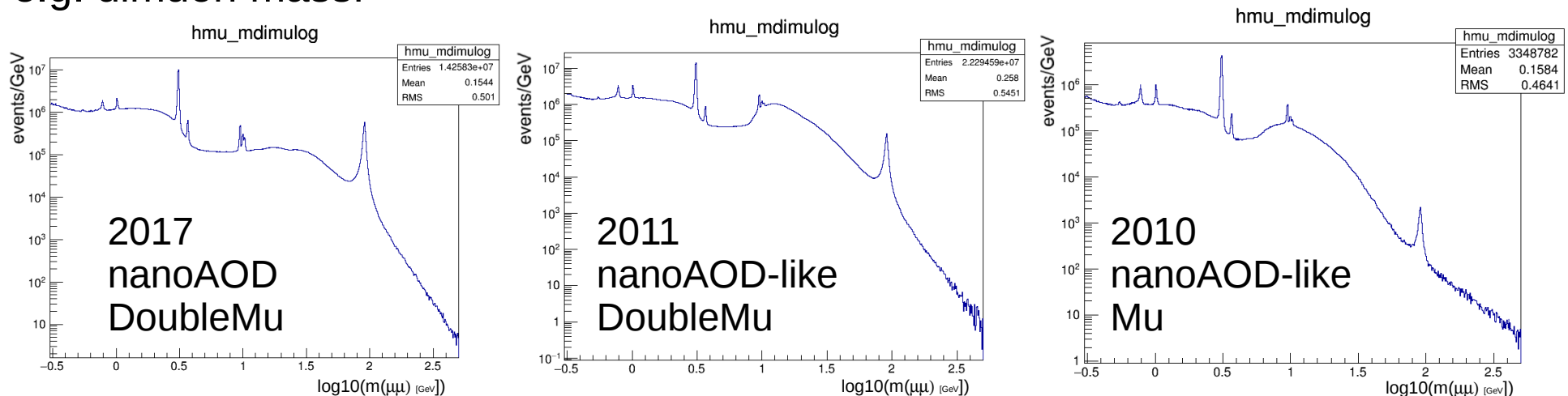
- Exactly reproduce some known/well-validated Run 1 distributions from nanoAOD-like ntuple

summer students
(+ DY studies??)

Motivation for nanoAOD-like 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)

e.g. dimuon mass:



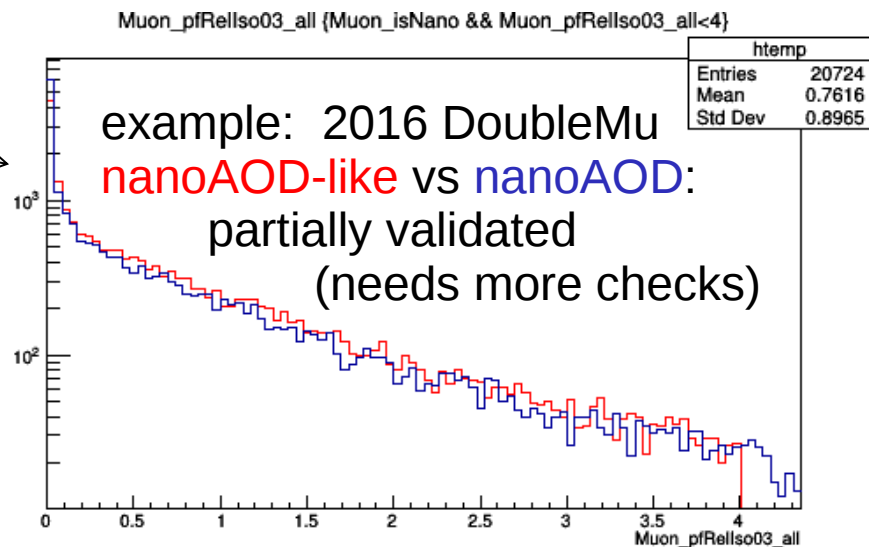
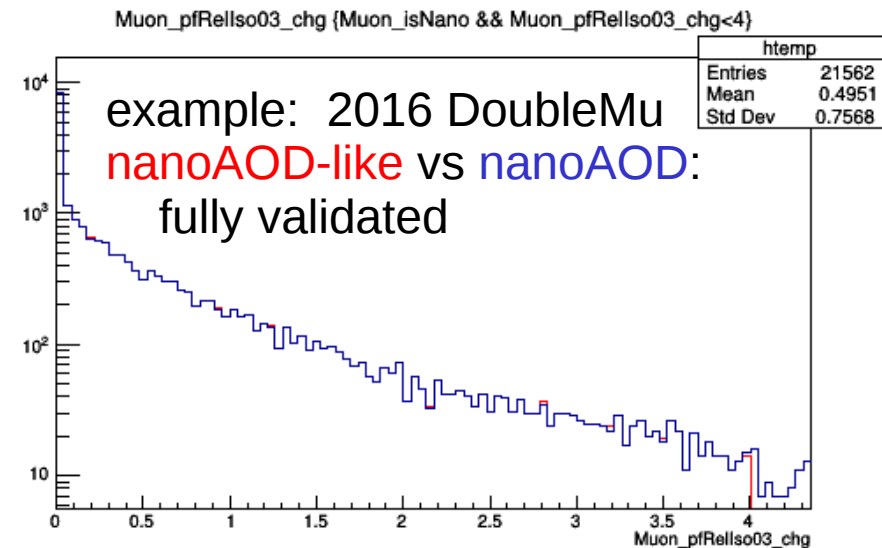
each plot produced within minutes with identical plain ROOT script

Status of nanoAOD-like format

(3 months 2018 + 2 weeks 2019)

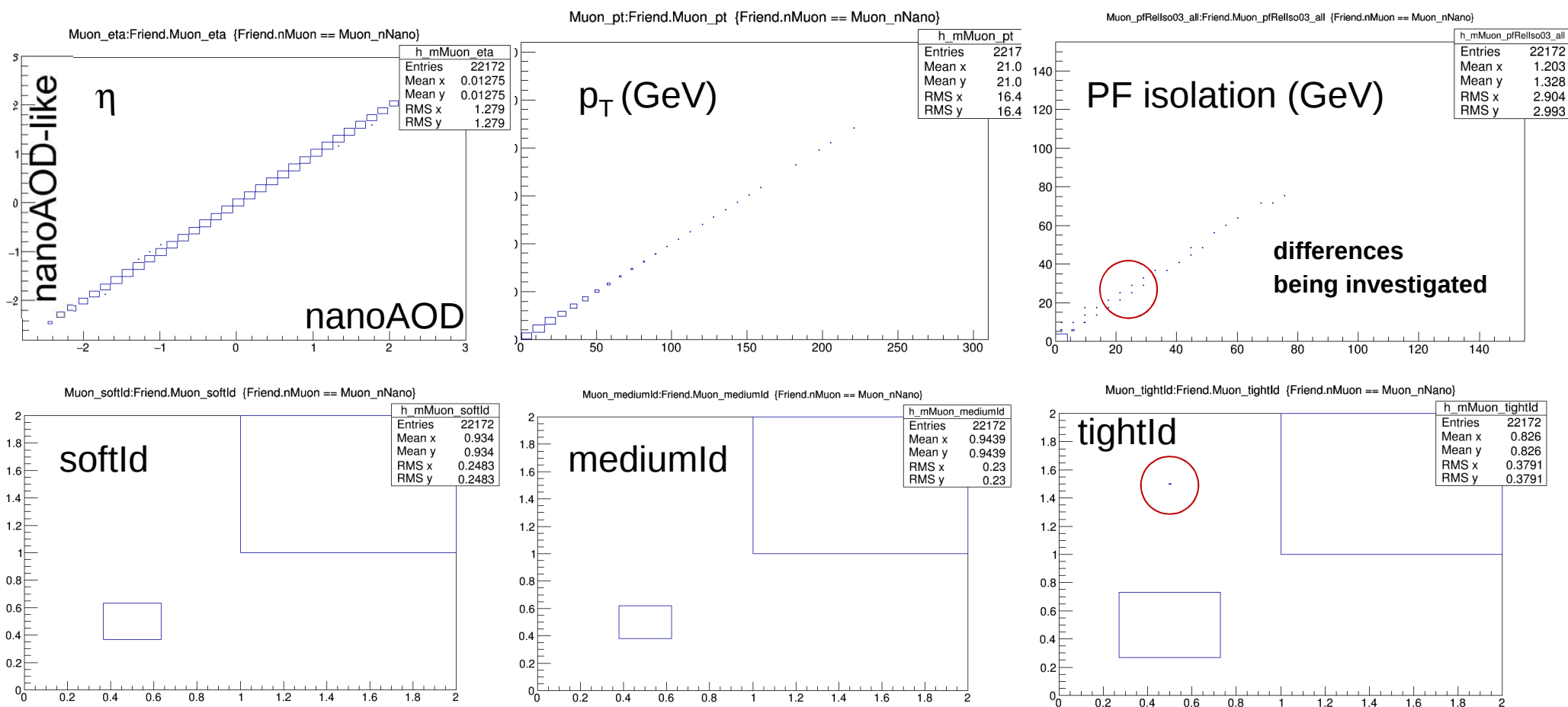
- **100% of run/event, PV and Muon nanoAOD variables implemented on CMSSW 4_2_8, 5_3_32 (Run 1), 7_6_1, 8_0_20, 9_4_8 and 10_2_26 (new) (Run 2 for checks/validation only)**
- **85% filled with useful content** according to workbook recipes
- **40% fully validated** against **nanoAOD version 9_4_8**
- **25% partially validated** (needs more checks)
- **20% not yet validated**
- **15% not yet filled usefully**
- **Run 1 revalidation against public examples (from AOD) started**

(47 variables out of ~400)



muon variables: comparison with official nanoAOD

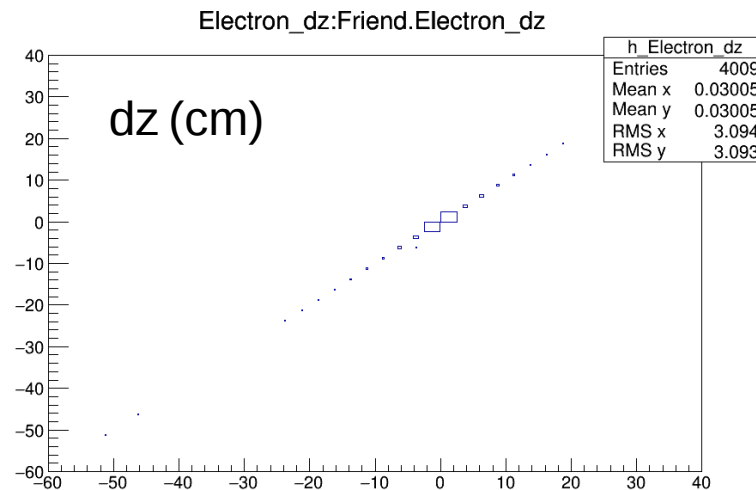
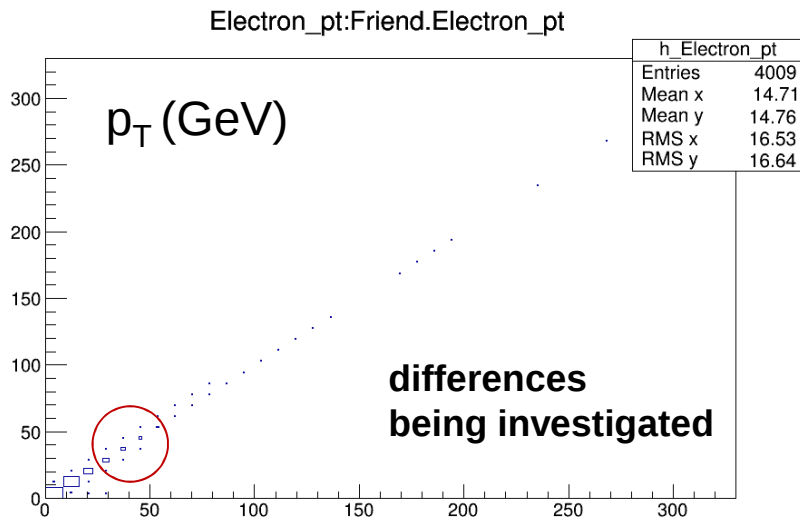
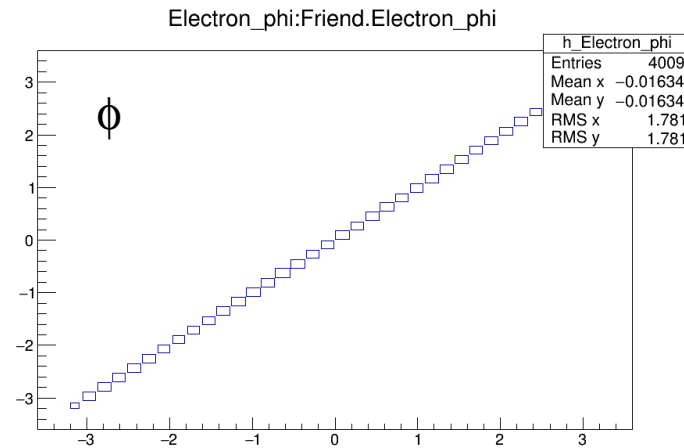
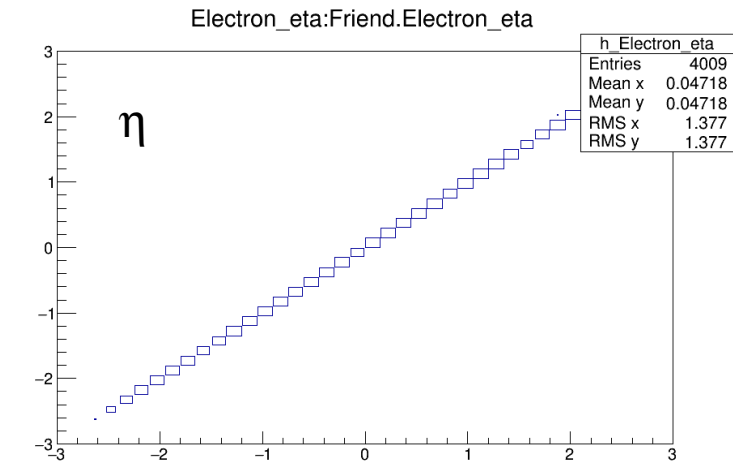
fraction of 2016 DoubleMu dataset: nanoAOD-like from 8_X legacy AOD vs official 9_X nanoAOD
(reimplementation of variables according to workbook)
-> hope to be ready for “blessing” of variable subset by muon POG soon



electron variables: comparison with official nanoAOD

fraction of 2016 DoubleMu dataset: nanoAOD-like from 8_X legacy AOD vs official 9_X nanoAOD
(reimplementation of variables according to workbook)

-> implementation started, making progress



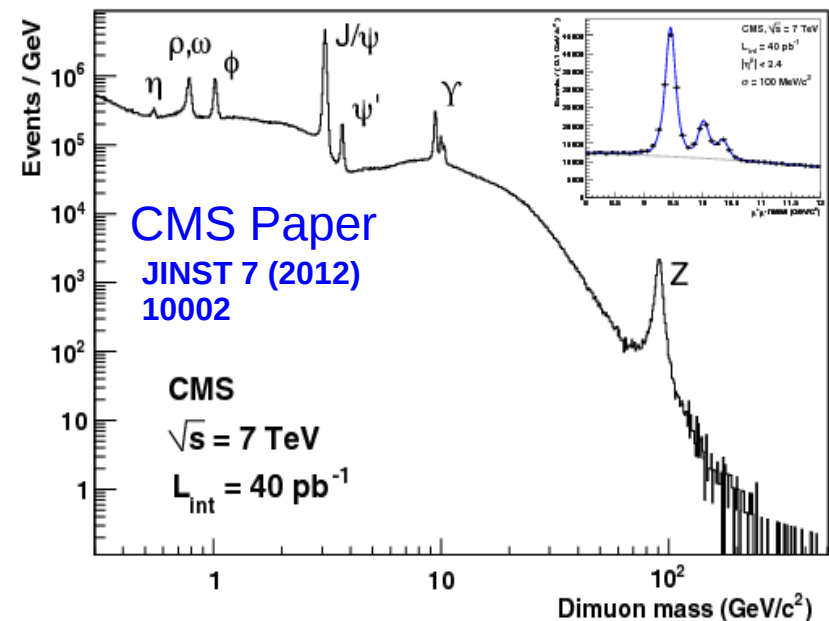
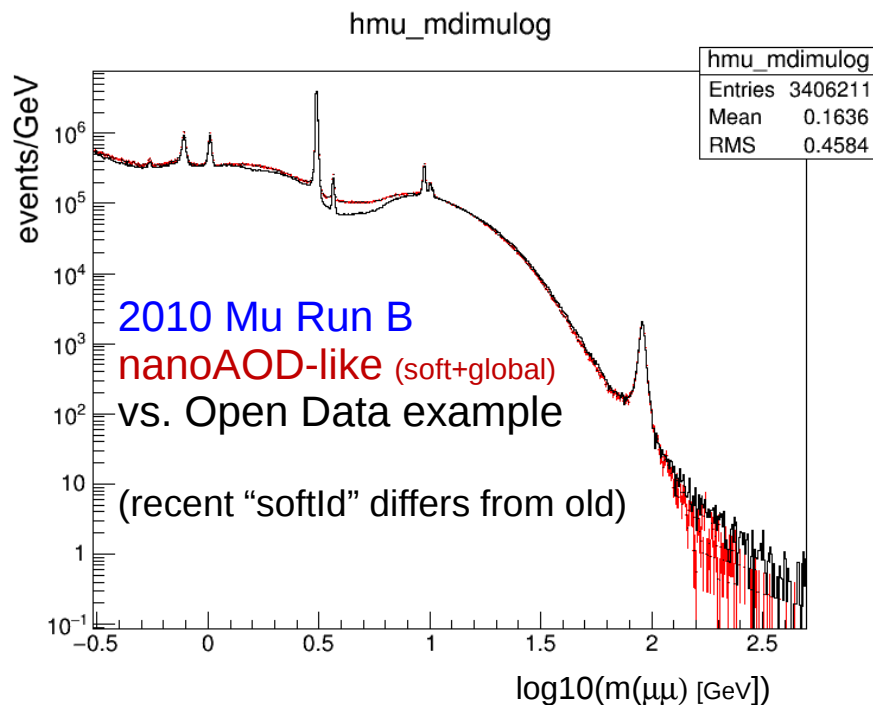
16.05.19

A. Geiser, nano meeting

10

Motivation for nanoAOD-like format for Run 1

- **simplified format for Open Data, same for all open data sets**



- **optionally extend content**
such that it becomes useful for 'almost all' analyses (e.g. PF candidate list equivalent to miniAOD, specific meson candidates, possibility to revertex)
-> could replace Run 1 AOD in the long term

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 already tentatively assigned (soon team of ~6 people part time) ***** today's meeting! *****

**-> hope to complete nanoAOD-like 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

Backup

Tentative list of tasks/contents

- nanoAOD ntuple content (9_4 v2):

	variables	implemented	content implemented	content validated	remaining work
•	run/event/lumis.	3	100%	100%	done
•	Generator /PSweight	11	-	-	~0.5 months, Hannes?
•	PV /OtherPV /Pileup	14	70%	70%	~0.5 months, Achim
•	SV	13	100%	10%	~1 month, Achim
•	GenPart	9	100%	50%	~0.25 months, Achim
•	Muon	35	100%	80%	~1 month, Achim + Nuha?
•	Electron	48	20%	20%	~3 months, Nuha?, Qun?
•	Photon	28	25%	25%	~2 months, !not yet covered!
•	Tau	38	25%	25%	~3 months, !not yet covered!
•	IsoTrack	13	100%	-	~0.5 months, Achim
•	GenDressedLepton	14	-	-	~0.5 months, Hannes?
	+ GenVisTau				
•	Jet+FatJet +SubJet	79	5%	5%	~5 months, 2 covered , Armando?
	+SoftActivityJet+SoftActivityJetHTX				
•	GenJet +GenJetAk8	14	-	-	~0.5 months, Hannes?
•	MET+TkMET	23	30%	30%	~2 months, 0.5 covered, !1.5 not yet!
	+CaloMET +RawMET+PuppiMET				Stefan??
•	TrigObj	11	-	-	~1 month, Qun??
•	HLT	569 (!?)	-	-	concept to be discussed ...
•	LHEPart	11	-	-	~0.5 months !not yet covered! Hannes??
	+LHEPdfWeight + LHEScaleWeight+LHEWeight_originalXWGTUP				
•	Flag	26	-	-	~2 months, !not yet covered!
•	Various other	10	-	-	~1 month
•	coordination + set up & manage twiki + git repository				~1 month, Achim + Nuha?
•	various general technicalities of setup				~0.75 months, Nuha
total					~26 months 2019/20
					~13 this year
					~13 next year (tbc)

plans

nanoAOD-like data format for Run 1 making progress,
need to strengthen interaction with XPOG, POGs/PAGs, and PPD

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			AOD nanoAOD available miniAOD available RECO only	

*external project in preparation

Motivation for nanoAOD-like format for Run 1

- **output size much smaller than AOD**

current implementation of

AOD->nanoAOD-like Run 1

has similar performance as

miniAOD->nanoAOD Run 2

~0.02 s/event (including extensions)

~1 kB/event (including extensions)