

Update on mini-DST project

Shin-ichi Kawada (DESY)

2020/June/17

ILD Analysis/Software Meeting



mini-DST project: Introduction

- Triggered by Snowmass, but also good for beginners.
- We want to involve many newcomers and theorists to the ILC. Theorists typically use Delphes card to evaluate their favorite new physics models.
- Is Delphes reliable? SiD Delphes card gives factor 2 to 10 worse performance. How about ILD Delphes card? (investigated by Marcel and Filip)
- We are planning to provide “minimum” set of data based on the full simulation (or based on SGV). Absolutely reliable.

mini-DST project: Philosophy in Cartoon

- What is “minimum”? This is open question.
 - Usually only interested final products from our great reconstruction
 - No Tracks, No Clusters, ...
 - Have interests about isolated leptons, isolated photons, jets...
(result of High-Level Reconstruction)

result of our great reconstruction
too many information for beginners

compact but useful
read data without using iLCSoft

DST

mini-DST

throw away
unnecessary
collections



add useful
information






Updates

- Created a repository in GitHub for mini-DST project.
 - <https://github.com/shkawada/mini-DST>
 - README.md will be updated frequently to show the status and documentation for future.
- Created 0th version prototype of mini-DST
 - Full information available: <https://github.com/shkawada/mini-DST/blob/master/v0/memo.md>
 - Used all physics validation test samples
 - More things / further treatments will be in next iteration

Details of v0 mini-DST

- In v0 mini-DST, the following collections are kept from DST.
 - PandoraPFOs, BCalRecoParticle (not merged, not simplified anything)
 - MCParticle
 - PrimaryVertex, PrimaryVertex_RP
 - RecoMCTruthLink, MCTruthRecoLink (kept full relation)
- The following collections are added to the mini-DST file.
 - event shape variables (used [ThrustReconstruction](#), [Sphere](#), [Fox](#)): stored at the header of PandoraPFOs
 - IsolatedMuons, IsolatedElectrons (used [IsolatedLeptonTagging](#), not tuned)
 - IsolatedTaus (used [TaJetClustering](#), not tuned)
 - RefinedNJets (N = 2, 3, 4, 5, 6) (used [LCFIPlus](#): JetClustering, JetVertexRefiner, FlavorTag, not tuned)
 - The [ErrorFlow](#) is applied to RefinedNJets to calculate covariance matrix for jets.

Flow Chart (1)

-  stored in mini-DST
-  not stored in mini-DST
(throw away when creating mini-DST)
-  tools

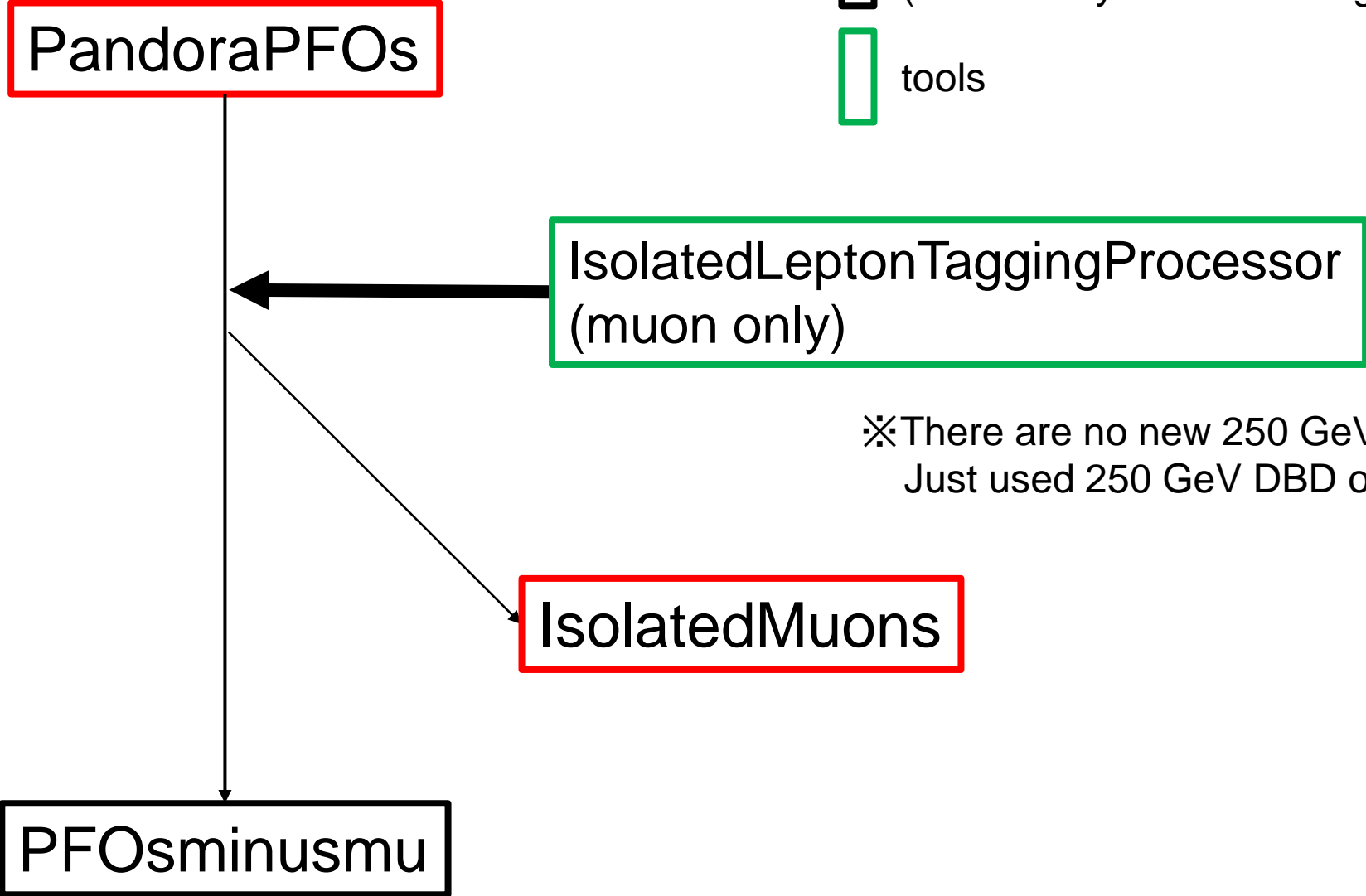
PandoraPFOs

ThrustReconstruction
Sphere
Fox

stored parameters (header of PandoraPFOs)
FoxWolfram_moment(0), Oblateness, aplanarity,
majorThrustAxis, majorThrustValue,
minorThrustAxis, minorThrustValue,
principleThrustAxis, principleThrustValue,
sphericity, sphericity_tensor_eigenvalues

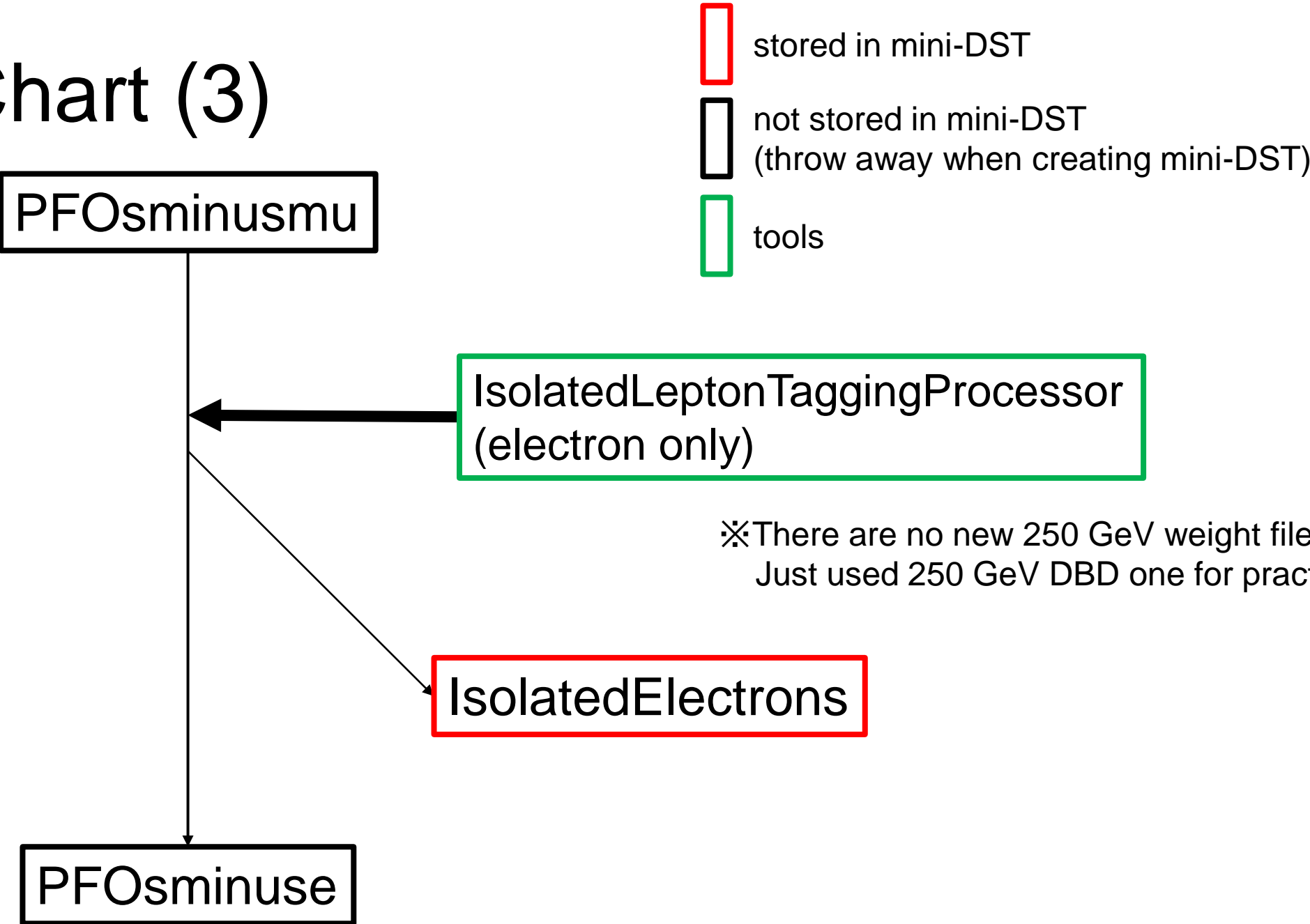
Flow Chart (2)

- stored in mini-DST
- not stored in mini-DST
(throw away when creating mini-DST)
- tools

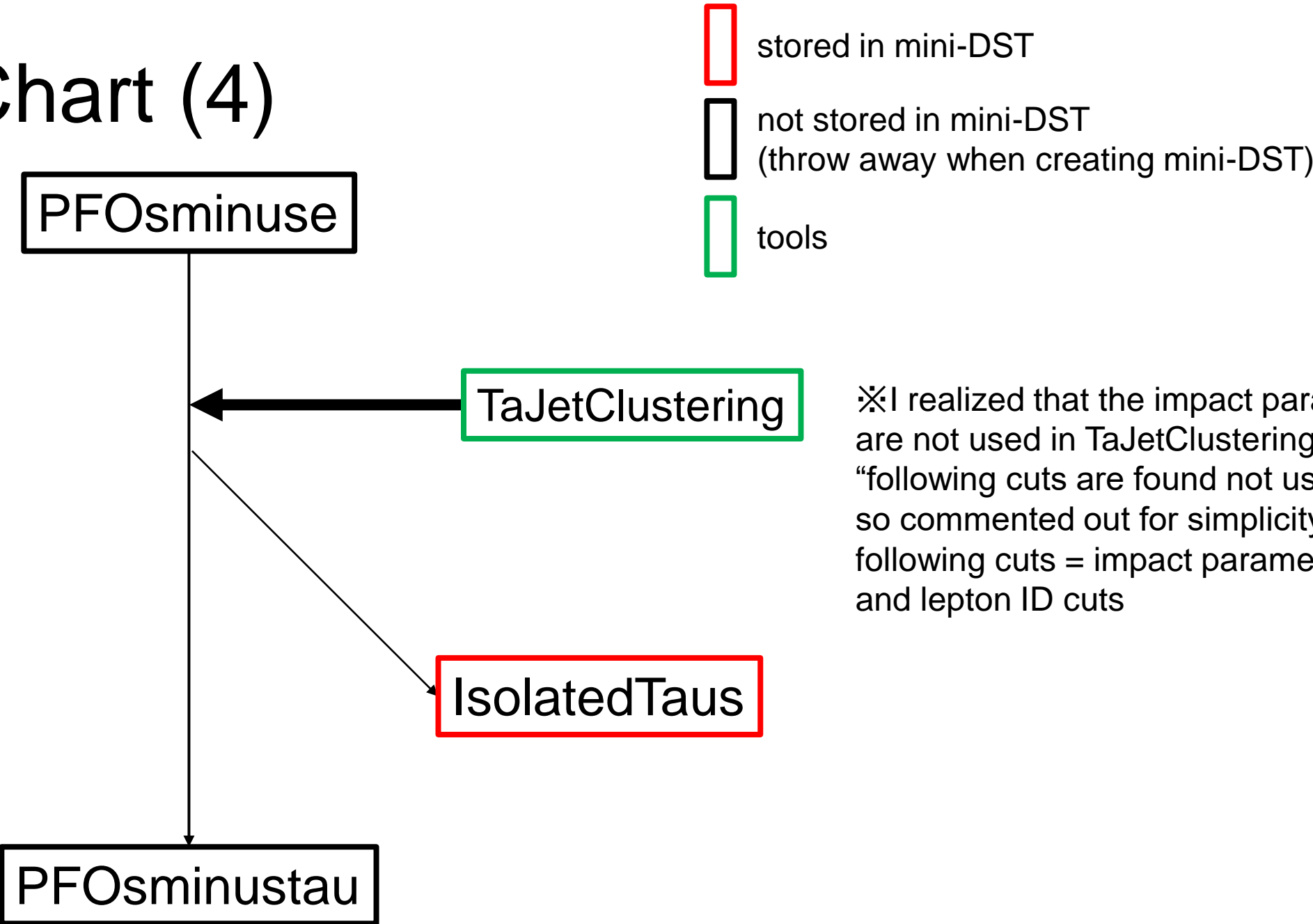


※There are no new 250 GeV weight files.
Just used 250 GeV DBD one for practice.

Flow Chart (3)



Flow Chart (4)



Flow Chart (5)

PFOsminustau



stored in mini-DST



not stored in mini-DST
(throw away when creating mini-DST)



tools

LCFIPlus
(JetClustering, JetVertexRefiner, FlavorTag)

RefinedJets

VertexJets
RefinedJets_rel
RefinedJets_vtx
RefinedJets_vtx_RP
RefinedVertex
RefinedVertex_RP

※ There are no new 250 GeV weight files.
Just used 500 GeV one for practice.

Final outcome is RefinedJets
ErrorFlow is applied to RefinedJets
to calculate covariance matrix of jets

Removed from DST so far (CUTSET)

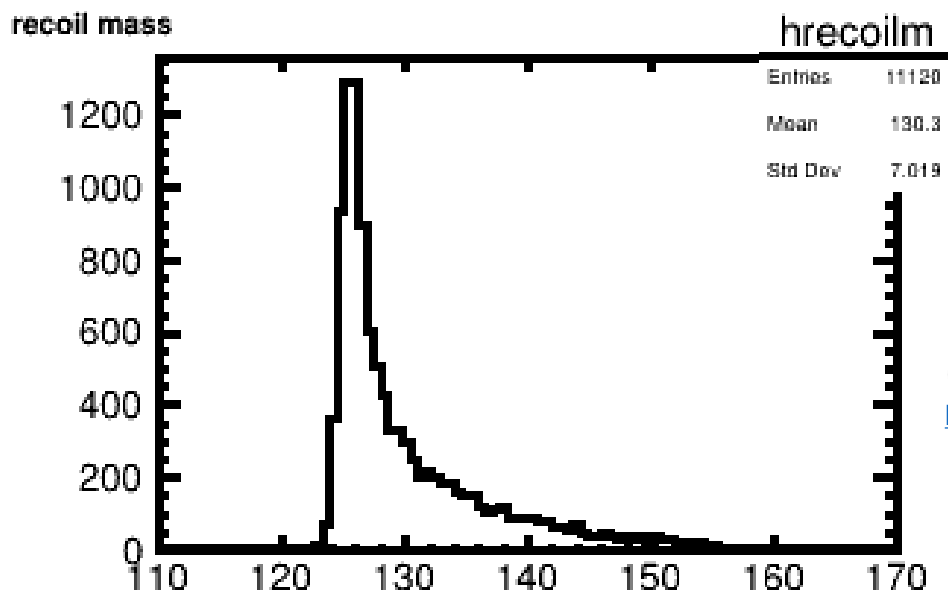
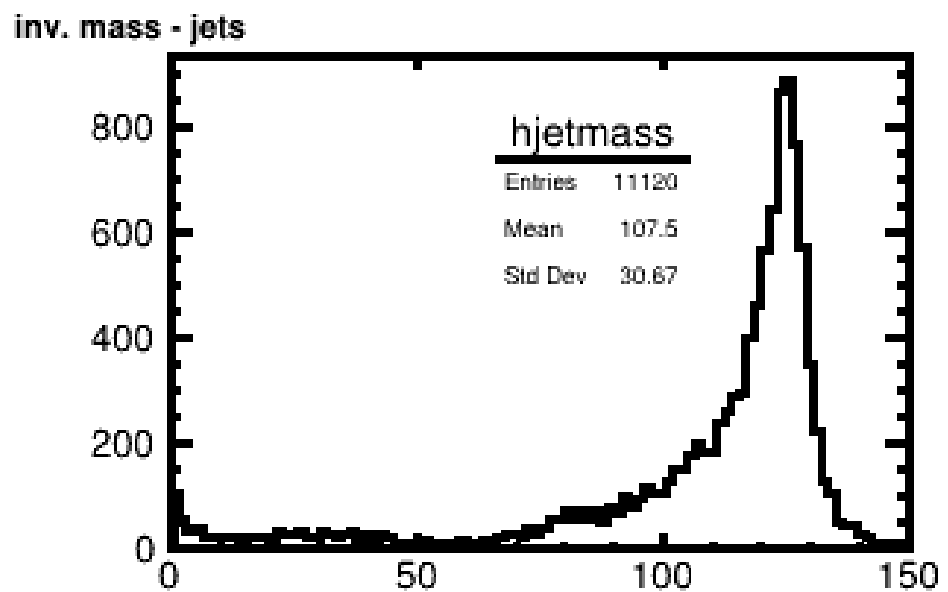
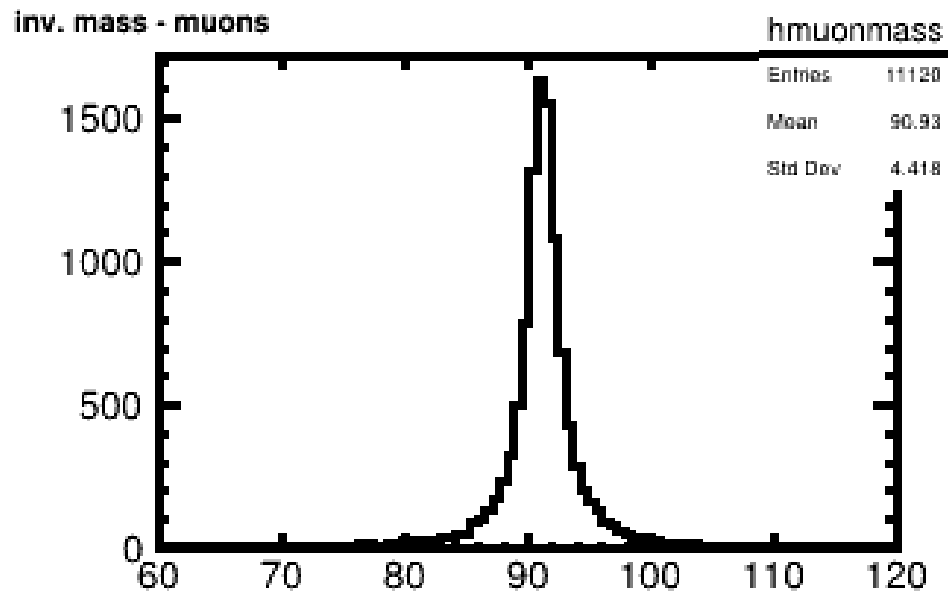
- Collection Types
 - Track
 - Cluster
- Collection Names
 - ClusterMCTruthLink
 - MCTruthClusterLink
 - MCTruthMarlinTrkTracksLink
 - MarlinTrkTracksMCTruthLink
 - DistilledPFOs
 - GammaGammaCandidateEtaPrimes
 - GammaGammaCandidateEtas
 - GammaGammaCandidatePi0s
 - GammaGammaParticles
 - V0RecoParticles
 - V0Vertices
 - ProngRecoParticles
 - ProngVertices
 - KinkRecoParticles
 - KinkVertices
 - SplitRecoParticles
 - SplitVertices
 - BuildUpVertex_RP
 - BuildUpVertex
 - BuildUpVertex_V0_RP
 - BuildUpVertex_V0

File size (measured with “wc -c filename”)

		e2e2h									qqh_zz	2f_z_l	4f_ww_sl	4f_sw_sl	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
no cut	events	12500	12500	12500	11000	10500	10500	10500	10500	8000	5500	4500	10000	10000	10000
	no cut (MB)	663	663	660	584	761	762	762	764	584	680	551	91	595	603
	no cut / event (MB)	0.053	0.053	0.0528	0.0531	0.0725	0.0726	0.0726	0.0728	0.073	0.1236	0.1224	0.009	0.0595	0.0603
case1	remove CUTSET	267	267	265	234	302	302	303	303	232	265	214	50	248	256
	(MB) / event	0.0214	0.0214	0.0212	0.0213	0.0288	0.0288	0.0289	0.0289	0.029	0.0482	0.0476	0.005	0.0248	0.0256
case2	case1 + following event shape numbers IsolatedM/E/T JCandFT (N=2,3,4,5,6)	285	285	284	251	319	319	319	320	245	275	223	56	264	271
	(MB) / event	0.0228	0.0228	0.0227	0.0228	0.0304	0.0304	0.0304	0.0305	0.0306	0.05	0.0496	0.006	0.0264	0.0271
	only have event shape numbers PandoraPFOs (w/ shape) IsolatedM/E/T JCandFT (N=2,3,4,5,6) PrimaryVertex PrimaryVertex_RP all others dropped	162	162	162	143	173	173	173	173	132	137	111	36	139	139
	(MB) / event	0.013	0.013	0.013	0.013	0.0165	0.0165	0.0165	0.0165	0.0165	0.0249	0.0247	0.004	0.0139	0.0139

✂case3: test for file size, and useful sets if you only interested in final ReconstructedParticle/Vertex collections.

✂I could not remove PandoraPFOs due to runtime error. But PandoraPFOs will be stored in mini-DST.



one e2e2h test sample
 just require 2 jets (Refined2Jets)
 and 2 muons (IsolatedMuons)
 long tail of 2-jet invariant mass
 would be from $h \rightarrow \tau\tau$, WW^* , ZZ^* ...

(used higgs_recoil_plots.C in
https://github.com/LCSOft/LCIO/blob/master/examples/cpp/delphes2lcio/examples/higgs_recoil_plots.C)

Summary & Future Plan (1)

- The 0th prototype of mini-DST files are created.
- Planning to compare mini-DST and SGV-mini-DST
- Similar file from Delphes (delphes2lcio)

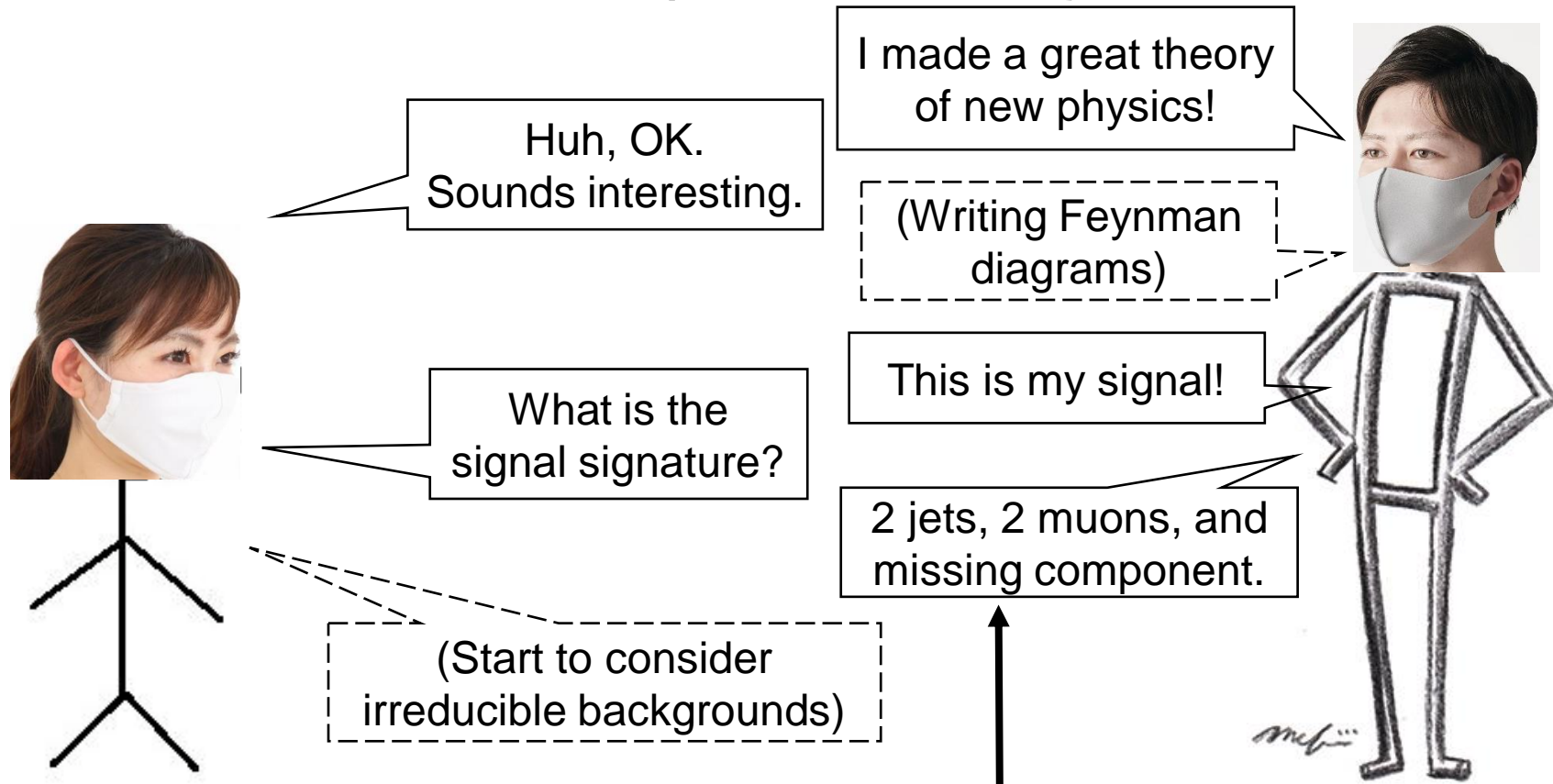
Future Plan (2)

- PandoraIsolatedPhotonFinder developed by Yan
 - This processor actually behaves as ISR finder, and using many other external functionalities
 - Planning to...
 - develop new general isolated photon finder
 - attach photon finding option in IsolatedleptonTagging
 - In any case, weight files are necessary which are not produced

BACKUP



Typical Conversation (Cartoon)



✘ This signature can be any number of anything.

EXPERIMENTALIST

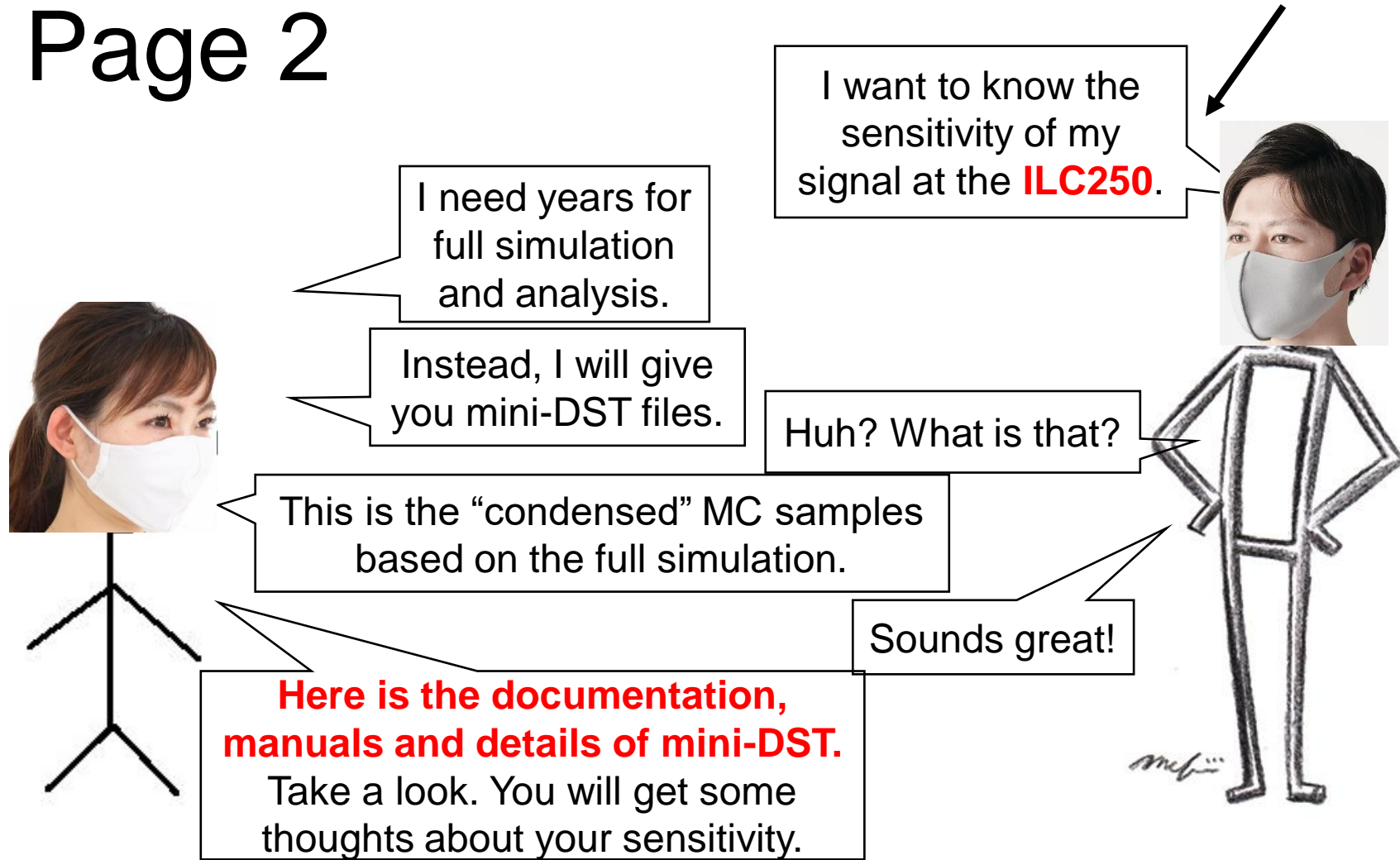
THEORIST

at least 1.5 m



✘ This machine setting can be anything.

Cartoon Page 2



EXPERIMENTALIST

THEORIST

at least 1.5 m



ReconstructedParticle	Vertex	Name of collection	Type	fromType ---> toType
		MCParticle	MCParticle	
PandoraPFOs		MCTruthRecoLink	LCRelation	MCParticle ---> ReconstructedParticle
BCalRecoParticle		RecoMCTruthLink	LCRelation	ReconstructedParticle ---> MCParticle
V0RecoParticles	V0Vertices	RefinedJets_rel	LCRelation	ReconstructedParticle ---> Vertex

- ProngRecoParticles
- KinkRecoParticles
- SplitRecoParticles
- PrimaryVertex_RP
- BuildUpVertex_RP
- BuildUpVertex_V0_RP

- RefinedVertex_RP
- RefinedJets_vtx_RP
- VertexJets

RefinedJets

- IsolatedMuons
 - IsolatedElectrons
 - IsolatedTaus
 - (IsolatedPhotons)
- isolated objects

products of LCFIPlus

These are the kept collections after removing CUTSET (newly added) so far.

Open questions:

Which collections should be dropped/kept?

Red: must be stored

Green: dropped in v0

Which collections can be merged? (TQ)

List of Open Questions (1)

- What is the format/content of mini-DST file?
- Do we need to add more information? e.g.: number of isolated objects, jets, event shape numbers...
- How to add parameters at the event header? (TQ)
- Do we need other HLR processors? For what?
- How to add # tracks/clusters at the event header? (TQ)

List of Open Questions (2)

- How to handle particle ID information?
 - currently 5 particle hypotheses (e, mu, pi, p, K) are tested with 4-5 PID algorithms ---> too complex
 - store only results (all 5 hypotheses) of one PID algorithm as a PIDObject? ---> which? pure dE/dx ? Or the combined likelihood PID? If the latter, who tunes it?
 - or even go further and store only the best fitting / highest likelihood assignment, in the "PDG" field of the ReconstructedParticle, possibly only if it is significantly better than the other hypotheses (TQ)
- What to do with the V0's, Kink's, Prong's etc from (pre-)Pandora? (and where's the second V0 collection from LCFI ?) Can we just drop them? Or can they be merged into one "speciality" collection?
 - Drop them for v0
- Can the logic of always having an extra Vertex_RP collection for each Vertex collection be simplified? (TQ)

List of Open Questions (3)

- Should/Can the MCParticles be skimmed further? If yes, how? (TQ)
 - MCParticle ---> at least should be MCParticlesSkimmed (TQ)
- Can we simplify the contents of RecoMCTruthLink/MCTruthRecoLink? (TQ)
 - only highest weight? (TQ) keep full for v0
 - you probably also would like to re-run the RecoMCTruthLink in order to revert it to its "simplified" version, as we had it for the DBD (just highest weight relation stored, which of course does not always give a useful answer in cases with significant confusion, but the study of confusion is beyond the scope of the mini-DST, I'd say). (TQ)

List of Open Questions (4)

- How to run TrueJet? How could "simplified" output of TrueJet be stored? (TQ)
- Can we "fake" a RecoMCTruthLink for the BeamCalPFOs ?
- How many jet collections to be stored?
- Parameter tunings for all processors?
- Prepare macro/library to read mini-DST?
- Documentation: short summary and full detail document?

My Thoughts in Text (1)

- You have your own signal signature, typically number of leptons/photons/jets at first
 - Providing NJets collections are much more useful than the jet collection with fixed Y_{cut} value? Y_{cut} has physical meaning but also has higher degrees of freedom compare to NJets ($N = 2, 3, 4, 5, 6$).
 - Putting $Y_{N \rightarrow N-1}$ in the event header is extremely useful in my opinion if we decide to provide NJets collections.

My Thoughts in Text (2)

- Applying Durham clustering to ILC250 MC samples is fine. But how about higher energy samples?
 - More overlay particles: kT clustering can apply but need to choose the value of R.
 - I think the main target of Snowmass would be ILC250, but what we can do for higher energies?
 - Similar to this: tagging for isolated electrons/muons/taus/photons, need to choose any number of any parameters

My Thoughts in Text (3)

- Documentation!
 - Whatever the contents of mini-DST are, I think this is most important task.
 - Even we provide perfect mini-DST, it is not guaranteed to be used/interpreted properly/correctly. Misunderstanding can always happen.
 - To reduce risk, I think we need following two documents.
 - mini-DST in 10 minutes (overview, simple manual of how to use it, how to interpret your result)
 - full documentation of mini-DST