# HSF Data Analysis 2021-2022 kick-off meeting

6 October 2021

**Attending:** Allison Hall, Nicole Skidmore, TJ Khoo, Alexander Held, Eduardo Rodrigues, Gordon Watts, Lindsey Gray, Liz Sexton-Kennedy, Mark Neubauer, Mason Proffitt, Oksana Shadura, Remco de Boer, Nick Manganelli, Jim Pivarski, Stefan Piperov

## Analysis benchmarks

https://github.com/iris-hep/adl-benchmarks-index

L. Sexton-Kennedy- another important consideration for how to choose an analysis system is whether or not it is sustainable and will continue to be supported years in the future

L. Gray: one way to make things last is to be really good at what you do so people come join, or to make your tool useful to many groups (like Awkward Array has done)

- Single experiment scope is too small for core software; need to convince others to use it
- Regarding fitting benchmarks, it's useful to be able to look in and check what is happening at each stage. That might be easier to compare.
- Another metric for fitters is how many other tools it can interface with and how easy that translation/interface is

G. Watts: many current tools will be replaced by cooler tools in the future

- Does this matter if we have good analysis preservation? - this could be a benchmark in itself
- For model building need cross-checks
    - Make this a benchmark - how easy is it to access intermediate values
- Want benchmarks in between individual, isolated analysis tasks and a fully reproducible analysis - can build it up from the simpler ADL benchmarks we have now

J. Pivarski: Do not focus on future proofing software tools but on coding concepts

- We do not know the software tools that will be available in 30 years but we do know the concepts that will followed

T.J Khoo: Should get in the habit of "writing our tests first" before writing the software so we know what the software needs to do first

- Think of benchmarks as unit tests/regression tests/integration tests for the full analysis ecosystem
    - Elementary data extraction operations
    - Workflow structuring and configuration

L. Gray: How many common ideas/concepts are in the tool proposed as a benchmark

- Indicates if it's being built to last, to be upgraded?

N. Skidmore: How do we go about crowdsourcing, ask the experiments, file issues on ADL page?

- G. Watts: suggest to brainstorm in a more free-form structure, gdoc etc

## Preexisting ADL benchmarks (data extraction?)

1. Plot the ETmiss of all events.
2. Plot the pT of all jets.
3. Plot the pT of jets with $|\eta| < 1$.
4. Plot the ETmiss of events that have at least two jets with pT > 40 GeV.
5. Plot the ETmiss of events that have an opposite-charge muon pair with an invariant mass between 60 and 120 GeV.
6. For events with at least three jets, plot the pT of the trijet four-momentum that has the invariant mass closest to 172.5 GeV in each event and plot the maximum b-tagging discriminant value among the jets in this trijet.
7. Plot the scalar sum in each event of the pT of jets with pT > 30 GeV that are not within 0.4 in $\Delta R$ of any light lepton with pT > 10 GeV.
8. For events with at least three light leptons and a same-flavor opposite-charge light lepton pair, find such a pair that has the invariant mass closest to 91.2 GeV in each event and plot the transverse mass of the system consisting of the missing transverse momentum and the highest-pT light lepton not in this pair.

## Fitting benchmarks

- Unit tests that check for correctness
- Benchmark amplitude fit
- How do the fitters interface with the other tools?

- L. Gray: pyhf tests might be a useful starting point
- G. Watts: Since LHCb fits can be quite different than ATLAS and CMS fits, next step might be to share more about the fits to educate the rest of us to try to find common ground
- E. Rodrigues: Good news is that we don’t even need Open Data for this task; can easily use toy data
- L. Gray: we're all minimizing likelihoods, the common ground is definitely there - the inputs and data treatment can be wildly different
- A. Held: Bear in mind we have the unbinned and binned worlds as well. Being able to see implementations would help to understand what needs to be done.

- L. Gray: we should open up a common discussion about what people need from a shared fitting interface between the model building and minimization steps

## Workflow benchmarks

- Make the same histogram of pT 10 times, with different slices of eta from -2.0 to 2.0.
- Two steps… err, no three:
    - Make a plot of Pt for eta between -1.0 and 1.0. Make a second plot of pT from -2.0 to 2.0 with -1.0 and 1.0 excluded
    - Create ratio plot of the two eta regions
    - Plot the eta of the outer region, but weight each entry by the ratio of the first two plots.
- Steps for training a NN:
    - Train a simple (well specified) multi layer ntuple on pT and p to predict eta. Use the odd numbers for training.
    - Run the prediction on even event numbers and show standard plots.
- Produce plot of eta for pT above 30 GeV. Then cycle through sys errors on jet energy, and remake the eta plot, stack them to show differences, and print out a table showing event yields.
    - Advanced exercise: Determine the envelope of differences over all the systematic variations (could be more relevant for theory syst exercise)
    - Quantitative comparison: determine the CPU/event and memory/disk size for all systematics (excellent: some things we should be evaluating these things).
- For 2 highest pT jets above 30 GeV, plot the dijet invar mass. Then show how this looks when sys errors are applied.
    - This might be too close to the above sys error example).
    - For both sys errors: think a bit how the two different experiments do sys errors - will have to be included in the benchmarks somehow.
- Make 100 plots. Add one additional plot - measure the time/complexity/resources for the additional plot (Ask Lindsey for details).
    - What is the incremental cost?
- Define a signal region and select events for it
    - Define a control region that substitutes leptons for MET (e.g.) in the SR, and select events for it
    - Extend to many control regions, validation regions, ...
- Run an ABCD method background estimate
    - Advanced option: a modified ABCD method (where signal leaks into B, C, and D). Explores the interface between histograms and a relatively straightforward fit.
- Force scaling:
    - Dataset on disk (GB)
    - Dataset on cluster (TB)
    - Dataset on ~grid (...)
    - How to access data on the grid (but that might end up being too experiment-specific)

TJK: Can we write down all the exploratory/validation checks requested after the first analysis approval stage for our most recent analysis?

- G. Watts: This is benchmarking the capacity for “analysis evolution”

Need to keep benchmarks as universal as possible

- Quantitative studies with reference data, example implementation, comparisons
- Open tasks to demonstrate capability
- Liz: Snowmass datasets?
    - Can they be realistic enough? Could make these more relevant even if physics output is not necessarily close enough.
    - Equal access intended

# Metadata document

https://docs.google.com/document/d/1zT5tPCtiNfuRm8ywKNbaNGvXGtCZYaO-GOj77pV2BEY/edit

L. Gray: Is there something approaching these standards we can try out? Can we try interfacing with CMS database? Should have some discussions and see what is possible before we make any concrete plans. Will get back to DAWG.

TJK: Define requirements first then talk to technical experts for implementation. This is a statement of intent. Can start writing down metadata tasks - “benchmarks”.

AH: Just because it's difficult with current tools we should not be prevented from specifying “ideal” scenario

TJK: Will rewrite final part and re-circulate

E. Rodrigues: Game changer is going beyond this document (which should be advertised widely) and finding cross-experiment effort/solutions