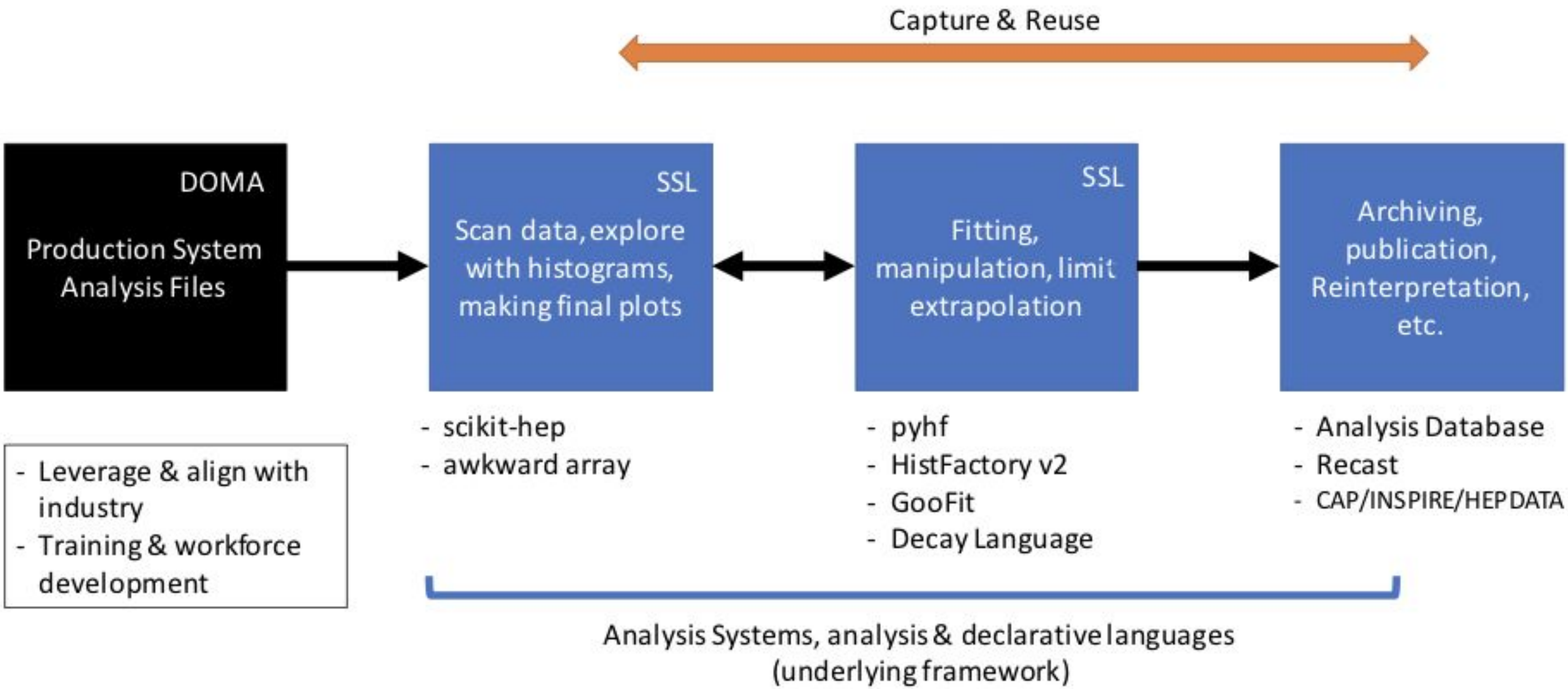
Analysis Systems Kickoff Closeout

Institution(s): Everyone in AS
PI: Everyone in the AS

Context

- Compared to DOMA and IA (which has more targeted reco/trigger goals), the Analysis Systems group is dealing with more “greenfield” area where there is a very heterogeneous set of use cases and relevant components
  - Nature of AS tasks will be more exploratory and “big R”
  - Jim’s analogy about mining and vein
- The AS group is bringing together a few existing groups
  - DASPOS and capture/reproducibility/reuse components of DIANA
  - Scikit-hep and Jim’s efforts on interoperability and query-based systems
  - High-performance statistical analysis tools (eg. GooFit, HistFactory, pyhf, etc.)
- And adding new connecting theme: declarative specifications

Analysis Systems Data Flow & Projects



Capture & Reuse

DOMA

Production System Analysis Files

SSL

Scan data, explore with histograms, making final plots

SSL

Fitting, manipulation, limit extrapolation

Archiving, publication, Reinterpretation, etc.

- Leverage & align with industry
- Training & workforce development

- scikit-hep
- awkward array

- pyhf
- HistFactory v2
- GooFit
- Decay Language

- Analysis Database
- Recast
- CAP/INSPIRE/HEPDATA

Analysis Systems, analysis & declarative languages
(underlying framework)

Outcomes of the workshop

- Improve the narrative around AS (clear to group, but not articulated well)
  - Will organize dedicated AS organizational meeting soon
- Brian’s suggestion: have end-to-end demonstrators of AS
  - We can go from AOD -> limit plots using preserved analysis now using CAP/REANA like system now. This can serve as an ATLAS starting point
  - Identify or build CMS and LHCb equivalents
- Add milestones associated with scalability tests
  - Blueprint activity to orchistrate AS -> SSL -> ATLAS/CMS Ops

Outcomes of the workshop

- Two different target communities:
  - Average End-user focused (not power users)
  - Activities aimed at power-users and those in the community beyond IRIS that will be developing analysis systems
    - Jim’s analogy about software dev. Company / mining and hitting a vein
  - Thought:
- Other takeaways for PEP:
  - Think about surveys, training, etc.
  - May add milestones connected to collaborating with external communities
  - Rethink granularity of Risks

Backup

Overall R&D goal for Analysis Systems

- Develop sustainable analysis tools to extend the physics reach of the HL-LHC experiments by creating greater functionality, reducing time-to-insight, lowering the barriers for smaller teams, and streamlining analysis preservation, reproducibility, and reuse.

Goals for 4 focus areas

- **WBS2.1** Establish declarative specifications for analysis tasks and workflows that will enable the technical development of analysis systems to be decoupled from the user- facing semantics of physics analysis.
- **WBS2.2** Leverage and align with developments from industry and the broader scientific software community to enhance sustainability of the analysis systems.
- **WBS2.3** Develop high-throughput, low-latency systems for analysis for HEP.
- **WBS2.4** Integrate analysis capture and reuse as first class concepts and capabilities into the analysis systems.

For Reference

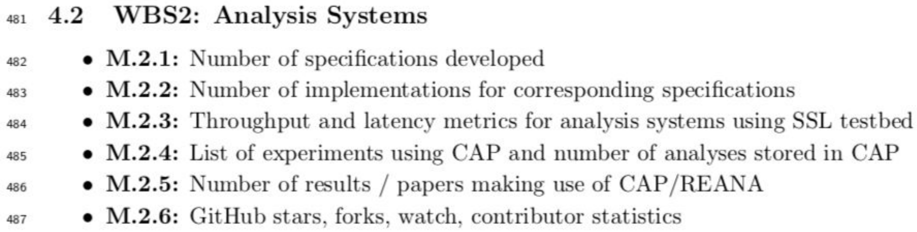
WBS2 (Analysis Systems): Develop sustainable analysis tools to extend the physics reach of the HL-LHC experiments by creating greater functionality, reducing time-to-insight, lowering the barriers for smaller teams, and streamlining analysis preservation, reproducibility, and reuse.

– WBS2.1 Establish declarative specifications for analysis tasks and workflows that will enable the technical development of analysis systems to be decoupled from the user-facing semantics of physics analysis.

– WBS2.2 Leverage and align with developments from industry and the broader scientific software community to enhance sustainability of the analysis systems.

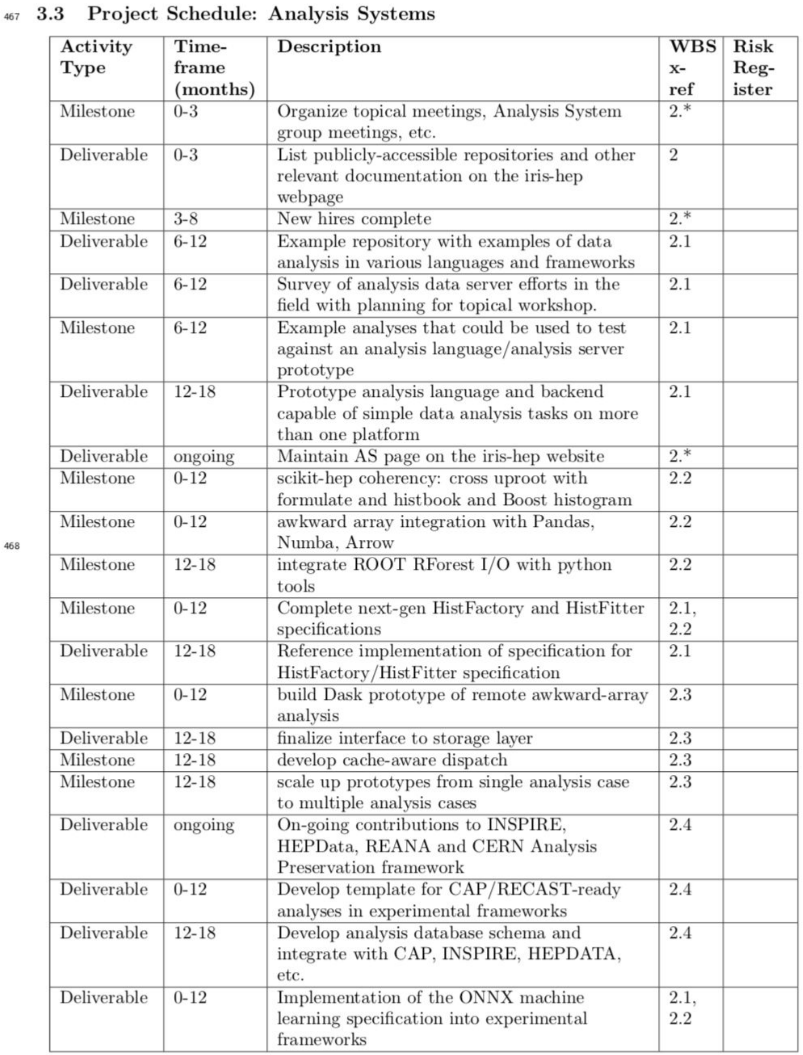
– WBS2.3 Develop high-throughput, low-latency systems for analysis for HEP.

– WBS2.4 Integrate analysis capture and reuse as first class concepts and capabilities into the analysis systems.



4.2 WBS2: Analysis Systems

- **M.2.1:** Number of specifications developed
- **M.2.2:** Number of implementations for corresponding specifications
- **M.2.3:** Throughput and latency metrics for analysis systems using SSL testbed
- **M.2.4:** List of experiments using CAP and number of analyses stored in CAP
- **M.2.5:** Number of results / papers making use of CAP/REANA
- **M.2.6:** GitHub stars, forks, watch, contributor statistics



3.3 Project Schedule: Analysis Systems

Activity Type	Time-frame (months)	Description	WBS x-ref	Risk Register
Milestone	0-3	Organize topical meetings, Analysis System group meetings, etc.	2.*	
Deliverable	0-3	List publicly-accessible repositories and other relevant documentation on the iris-hep webpage	2	
Milestone	3-8	New hires complete	2.*	
Deliverable	6-12	Example repository with examples of data analysis in various languages and frameworks	2.1	
Deliverable	6-12	Survey of analysis data server efforts in the field with planning for topical workshop.	2.1	
Milestone	6-12	Example analyses that could be used to test against an analysis language/analysis server prototype	2.1	
Deliverable	12-18	Prototype analysis language and backend capable of simple data analysis tasks on more than one platform	2.1	
Deliverable	ongoing	Maintain AS page on the iris-hep website	2.*	
Milestone	0-12	scikit-hep coherency: cross uproot with formulate and histbook and Boost histogram	2.2	
Milestone	0-12	awkward array integration with Pandas, Numba, Arrow	2.2	
Milestone	12-18	integrate ROOT RForest I/O with python tools	2.2	
Milestone	0-12	Complete next-gen HistFactory and HistFitter specifications	2.1, 2.2	
Deliverable	12-18	Reference implementation of specification for HistFactory/HistFitter specification	2.1	
Milestone	0-12	build Dask prototype of remote awkward-array analysis	2.3	
Deliverable	12-18	finalize interface to storage layer	2.3	
Milestone	12-18	develop cache-aware dispatch	2.3	
Milestone	12-18	scale up prototypes from single analysis case to multiple analysis cases	2.3	
Deliverable	ongoing	On-going contributions to INSPIRE, HEPData, REANA and CERN Analysis Preservation framework	2.4	
Deliverable	0-12	Develop template for CAP/RECAST-ready analyses in experimental frameworks	2.4	
Deliverable	12-18	Develop analysis database schema and integrate with CAP, INSPIRE, HEPDATA, etc.	2.4	
Deliverable	0-12	Implementation of the ONNX machine learning specification into experimental frameworks	2.1, 2.2	