CompF7: Reinterpretation and long-term preservation of data and code

Conveners: Stephen Bailey, Kyle Cranmer

- Report overview and recommendations
- Comments by white paper contributors
  - Data and Analysis Preservation
  - Analysis Description Language
  - Data Preservation for Cosmology
- Open Discussion
CompF7 Scope

Preservation of **data, software, and simulations**, both raw and processed (and also LQCD, theory, etc);

Preservation of **analysis data**, metadata, and other inputs/outputs like detector acceptances and derived likelihoods;

Preservation of **analysis logic** and code;

...beyond the **lifetime** of individual projects and analyses,

...in a way that makes these data/results **useful** for reuse & reanalysis

**Why?**

- **Maximizes the return on investment**
  - more science with the same data / results

- **Diversity, Equity, and Inclusion**
  - opens resources to people and institutions lacking the resources to participate in original work
  - training the next generation
CompF7 Report

Draft report v0

Comments doc
- https://docs.google.com/document/d/17M1iY3upfdDpp1vXATqy2OvC4GMSA4-mzp6RaWREVwM/edit?usp=sharing
- Same doc used for live notes from this session

What’s next?
- Comment via live notes for this session and async comments into the above doc
- We’ll make a v1 and distribute for further comment / refinement / co-authorship
  — will be posted at https://snowmass21.org/computational/start#report_drafts (& advertised)
Contributed White Papers

Data and Analysis Preservation, Recasting, and Reinterpretation
- arxiv:2203.10057

Analysis Description Language: A DSL for HEP Analysis
- arxiv:2203.09886

Data Preservation for Cosmology
- arxiv:2203.08113

Not directly covered by the white papers
- Neutrinos (though see “Data Preservation at Minerva LOI” arxiv:2009.04548)
- Lattice QCD, theory
- Simulations not directly tied to an individual experiment (e.g. cosmo sims)
- Dark Matter

Each of these will present highlights / priorities after this talk

We especially welcome comments from these in the discussion time today
Related Topics / Sessions

CompF5 End User Analysis, Monday
- End User Analysis Preservation discussed at length

CompF4 Storage and Processing Resources, Tuesday
- Underlying infrastructure

Cosmic Frontier 6, Complementarity of Probes and New Facilities
- Recommendation: “... We advocate for the creation of clear pathways to support cross-survey analyses as part of the core mission of the HEP Cosmic Frontier”
<table>
<thead>
<tr>
<th>Topic</th>
<th>Current Technologies / Solutions</th>
<th>Key Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experimental Data</td>
<td>CERN Open Data Portal; ESCAPE Project; NSF/NASA data centers</td>
<td>Smaller &amp; non-accelerator projects; preservation beyond lifetime of projects</td>
</tr>
<tr>
<td>Analysis Data</td>
<td>HEPData, Zenodo</td>
<td>Buy in; motivation/reward structure; what gets saved</td>
</tr>
<tr>
<td>Analysis Logic</td>
<td>Analysis Description Languages, code+containers, papers, ...</td>
<td>Preserving analyses in a useful way that makes this pragmatically possible</td>
</tr>
<tr>
<td>Recasting/Reinterpretation</td>
<td>RECAST / CheckMATE / Rivet / ... bespoke</td>
<td>Documentation; Computing resources to use the data</td>
</tr>
<tr>
<td>Diversity, Equity, Inclusion</td>
<td>Outreach programs; Good will</td>
<td></td>
</tr>
<tr>
<td>Experimental Code</td>
<td>Containers, Open Source GitHub+, CERN Open Data Portal</td>
<td>Will it still run N&gt;&gt;1 years from now?</td>
</tr>
<tr>
<td>Analysis Code</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. Ensure that all current and future experiments have a strategy and resourced program for the long term preservation of data and analysis capabilities, including beyond the lifetime of the individual experiments.

2. Invest in shared cyberinfrastructure to preserve these data and support a comprehensive analysis from various experiments and surveys — both active and completed — in order to realize their full scientific impact. The infrastructure should support the requisite theoretical inputs and computational requirements for analysis as well as metadata and APIs to track provenance and incentivize participation. Specifically,

(a) DOE Cosmic Frontier should fund a data archive organization (analogous to those supported by NASA and NSF) to preserve Cosmic Frontier datasets and simulations, and facilitate their joint analysis across the existing multiple computing centers.

(b) US funding agencies should coordinate with international partners such as the CERN Open Data Portal and fund additional resources as needed to ensure that all US-supported projects have data and analysis preservation support, including post-operations and including non-collider programs.
Commentary from White Papers

Data and Analysis Preservation, Recasting, and Reinterpretation
- arxiv:2203.10057

Analysis Description Language: A DSL for HEP Analysis
- arxiv:2203.09886

Data Preservation for Cosmology
- arxiv:2203.08113

... then open discussion