DATA PRESERVATION AND DATA WORKFLOWS SUMMARY

Adam Daskalakis & David Brown

BLUF

- Nuclear science is hard!
- The large nuclear data sets we as a community steward have innumerable benefits to society
- Continued care, thought and resources goes into maintaining these sets over decades, including improvements to workflows & data formats
- Society's needs have (and will continue) to change requiring: more diverse forms of data; more precise data; more careful characterization of uncertainties, corrections, etc.
- Open data, AI/ML and even legacy data present opportunities & concerns. We need to position ourselves appropriately
- However, other fields have faced similar challenges (i.e., high energy physics) and Gov't recognizes these challenges (SANPC, D3, CODA)

WE ARE NOT THE ONLY COMMUNITY WHO HAS THESE ISSUES

- Workshop on Software Infrastructure for Advanced Nuclear Physics Computing (SANPC) – Kyle Godbey
  - Basic nuclear science, both theory and experiment
  - URL https://www.jlab.org/conference/2024SANPC
- DOE Data Days (D3) - Andre Newsom
  - Promotion of data management as a means to higher quality, more efficient research and analysis, and as a critical component of data science.
  - URL https://data-science.llnl.gov/d3
- Conference on Data Analysis (CODA) – Scott Vander Wiel
  - Highlighting data-driven problems of interest to the Department of Energy
  - https://web.cvent.com/event/7845571b-b15d-418c-a24a-14468480c4ff/summary

WE HAVE A NEED FOR GUIDANCE

- We have guideline, policies, and initiatives but a lot of it is under review
- Data access:
  - https://www.energy.gov/datamanagement/doe-requirements-and-guidance-digital-research-data-management
  - https://doi.org/10.11578/2023DOEPublicAccessPlan
  - https://www.osti.gov/stip/about/supporting-documentation/2411C
- Data workflows:
  - https://bidenwhitehouse.archives.gov/wp-content/uploads/2024/12/NSTC-Framework-For-Considering-Data-Infrastructure-and-Interconnectivity.pdf
- AI/ML:
  - https://www.energy.gov/fasst

OUTLINE

- Presentation summaries + takeaways
- Deep dive on key takeaways

PRESENTATION SUMMARIES

OPEN DATA PRESENTATION SUMMARY

- Kyle Godbey – SANPC summary. 5 key areas: Innovation, cross-cutting initiatives, stewardship, data curation and preservation, workforce with primary question: What to do once funded project is over?
- Andre Newsom – Highlighted DOE Data Days (D3) themes and key activities. There's overlap between D3 and WANDA and potential for collaboration opportunities.
- David Brown – Discussed FAIR, FARR, FASST, and PuRe. Digital Object Identifiers (DOIs) - Started for NNDC and NNDC has robust backup plan is backed up with amazon web cloud.
- David Brown (summarizing Cooke) - Summarized Michael Cooke's presentation from ASCAC on Jan. 17th.

OPEN DATA SUMMARY

- Basic requirements (releasing data from plots at publication) “easy”
- Some questions about publication costs impacting poor institutions
- Concerns with cost of long-term data storage and fidelity of information to be stored
  - Ranges from trivial (text files) to significant financial burden
- Open Data without open codes vastly complicates retrieving information in a timely manner
  - Is it realistic to expect institutes to release codes – This is a part of a broader conversation
  - There's diminishing returns in time when data relies on old programming languages
- What is the future of Open Data?

LEGACY DATA PRESENTATION SUMMARY

- Vivian Dimitriou - 13 Data Centers for EXFOR but not all data stored. Discussed cases where there were issues, and the solutions were found.
- Boris Pritychenko – EXFOR contains results from thousands of experiments, which has a lot of real-world value. There's still missing legacy data and there's a want and need to recover. Lastly, there's a need for EXFOR modernization
- Denise Neudecker – There's a need to store data used for evaluations to guarntee reproducability. Templates developed to help fill in experimental uncertainties for evaluation. Uncertainty publication for best use of experimental data.
- Catherine Percher – Presented on some success and challenges retrieving information needed for benchmarks. Talked to information needs of benchmarks going forward.
- Keri Nicoll – Legacy data provided to public to retrieve information. Providing clear instruction was crucial to help aclimate outside workforce to project needs.

LEGACY DATA SUMMARY

- Integral benchmarks – Benchmarks require expert reviewers, but those expert reviewers are unfunded
  - Some older benchmarks may not have the rigor needed to be considered benchmarks today, but are we also approaching a point where its wandered too far for users
- EXFOR – workforce stretched, but there is so much legacy data to capture
  - Can and should we effectively leverage less experienced workforce? What other options do we have to ensure proper rigor in EXFOR compilations?
  - Moxon left boxes of data
    - Office dumps like this happen infrequently, but are a tremendous opportunity
    - THIS WILL HAPPEN AGAIN
    - How do effectively retrieve this information?
    - How do we prevent this from happening in the future?

DATA WORKFLOWS PRESENTATION SUMMARY

- David Brown (Summarizing Keith Jankowski) - Nuclear Data is a global need, since 2020 making the NNDC a PuRe resource. Database modernization of ENDSF and GNDS.
- Jean-Christophe Sublet – Multiple tools are needed for cross-checks and target comparible metrics with open-source systems. Shouldn't shy away from processes because its archaic.
- Michael Fleming - New members have joined and- having these intercommunicated codes requires a level of hidden knowledges. Succession planning is desirable. Looking to develop workflow that doesn't require all this information to be publicly available and enable a quicker workflow with GitLab-based services.
- Caleb Mattoon – Generalized Nuclear Database Structure (GNDS) stores and exchanging nuclear reaction data. Challenge now is to make transition of secondary tools from ENDF-6 to GNDS. FUDGE and GIDI+ workflows enables ENDF-6/GNDS to transport codes.

DATA WORKFLOWS SUMMARY

- LLNL, BNL, ORNL, LANL, NEA all have data pipelines that take process, check and often validate ENDF files
- Recreating the wheel vs validation and verification
  - Sometimes multiple codes doing the same thing is needed! (NJOY vs PREPRO)
  - It isn't reinventing the wheel if each institution has different needs
- GNDS -
  - What tools are a strategic investment is needed
- Succession planning, relying on retirees is not a long-term strategy
- Our digital ecosystem is constantly changing with updates and new codebases
  - Can we adopt best practices from other disciplines other learn from other groups

DATA & WORKFLOW PRESERVATION GOVERNANCE AND KNOWLEDGE MANAGEMENT PRESENTATION SUMMARY

- Stephen Bell - 50 to 100-year life cycles, how do workflows stay enabled over that time frame? Ned to communicate back and forth between disciplines and risk management with high-cost programs is a must, primarily through long-term R&D tied to programmatic goals. Common output and input across storage system. Searchable centralized databases is essential.
- Adam Daskalakis - Experiments and expensive and goal is to save all pertinent information to reproduce analysis and extract additiional information. Provides an example of how a common file format can be used to by community to more easily process complex analytical work.
- Yi Chen - Reanalyzing data from an archived data set. Allows innovative ways to reuse the data, way beyond the conclusion of the original experiment. Reproduced results from the original publication, then published additional papers using new algorithms.

PRESERVATION & GOVERNANCE SUMMARY

- The issue:
  - We need to think about data & workflow preservation on very long timescales
  - Traditionally, the focus was on the final files only
  - Saving the data & now workflows requires programs with long-term mindset like the existing ND collaborations
  - Will these collaborations need to adapt, and if so, how?

KEY TAKEAWAYS

DOE DATA ACCESS PLAN LIKELY TO REQUIRE RELEASE OF "DATA BEHIND THE DATA" TO ENSURE REPRODUCIBILITY

A laudable goal! But there is push-back!

What does it mean?

- Experiment: release raw(er) data + analysis package
- Theory: release intermediate data + theory codes
- Evaluation: release assembly scripts + corrections

- Issue 1: **Quality**
  - Intermediate work published, but the main product is not finished!
  - Some non-expert can take my code & data and pump out crumby papers
- Issue 2: **"Unfunded mandate"**
  - Requires lot of work to make useable by non-expert
  - Requires $$ to keep system alive as computers/compilers change
- Issue 3: **Ensuring credit & attribution**
  - Some expert can swoop in and take credit for my work
  - Some non-expert can take my code & data and pump out crumby papers

DOE DATA ACCESS PLAN LIKELY TO REQUIRE RELEASE OF "DATA BEHIND THE DATA" TO ENSURE REPRODUCIBILITY

A laudable goal! But there is push-back!

- Need time to complete work!
- Need resources for release & upkeep
- Need community guidelines/guardrails to ensure proper use

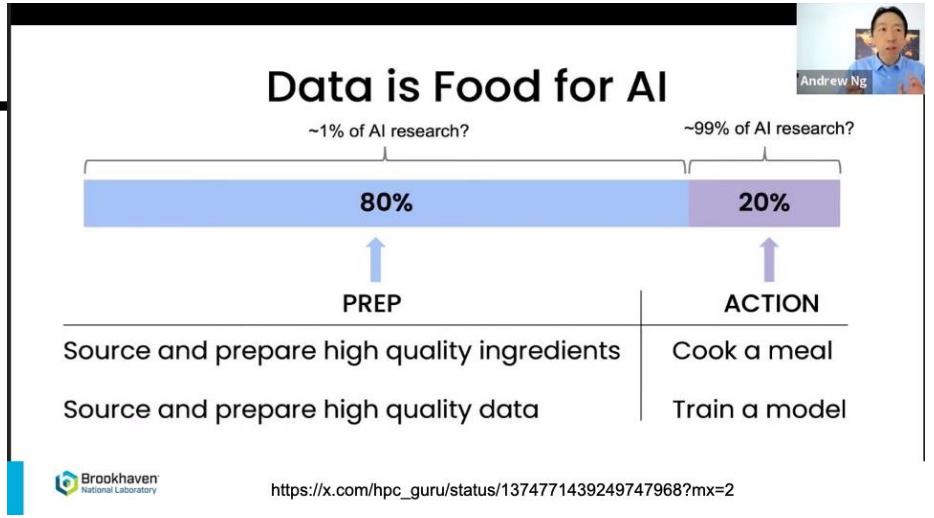
- Some non-expert can take my code & data and pump out crumby papers

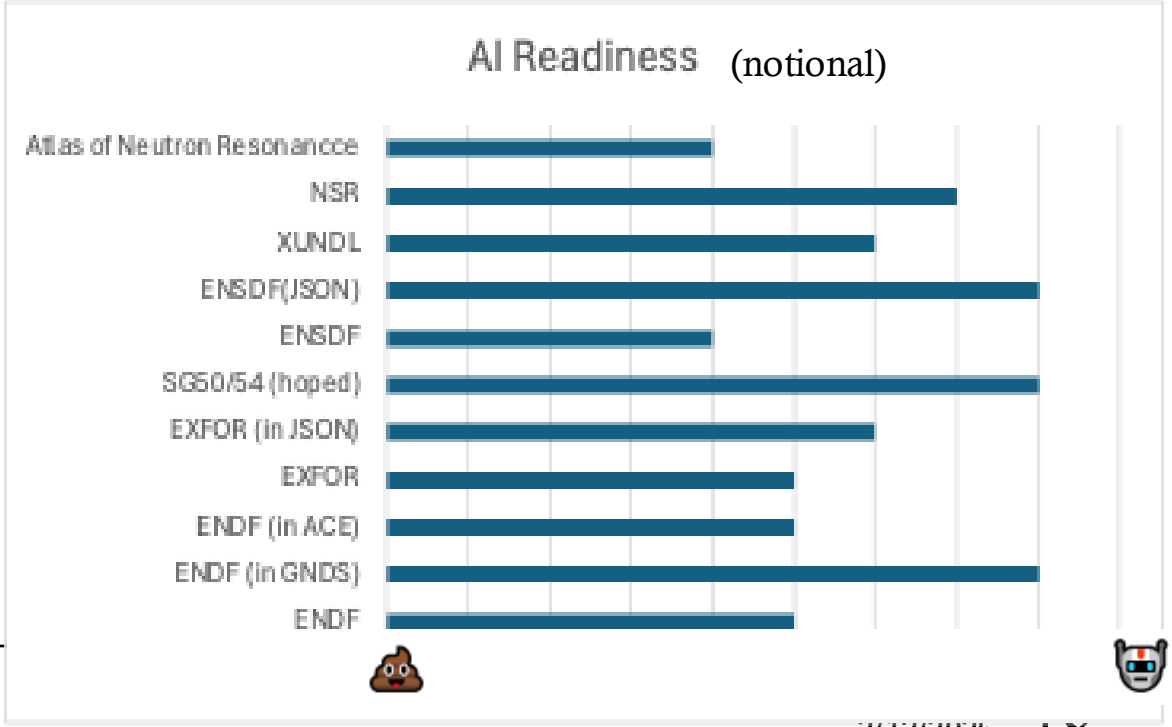
- Issue 2: **"Unfunded mandate"**
  - Requires lot of work to make useable

- Issue 3: **Ensuring credit & attribution**
  - Some expert can swoop in and take credit for my work
  - Some non-expert can take my code & data and pump out crumby papers

AI READINESS

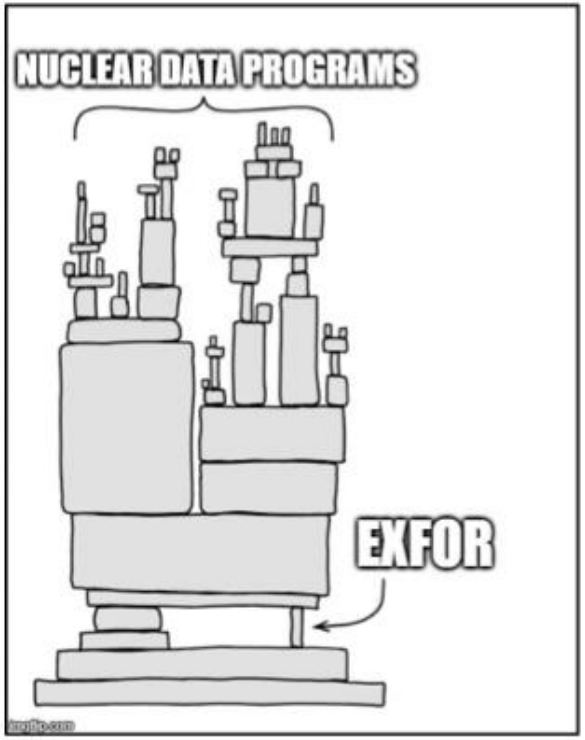
- We need to be ready to respond to future needs.
- Various modernization projects are getting our data ready
- Where in our pipeline can AI support data extraction?





EXFOR NEEDS HELP!

- Current compiler workforce is aging out & not being replaced fast enough
- Need more detailed information & expanded scope
- Need to rethink compilation approach!
- Rate determining step in ND Pipeline



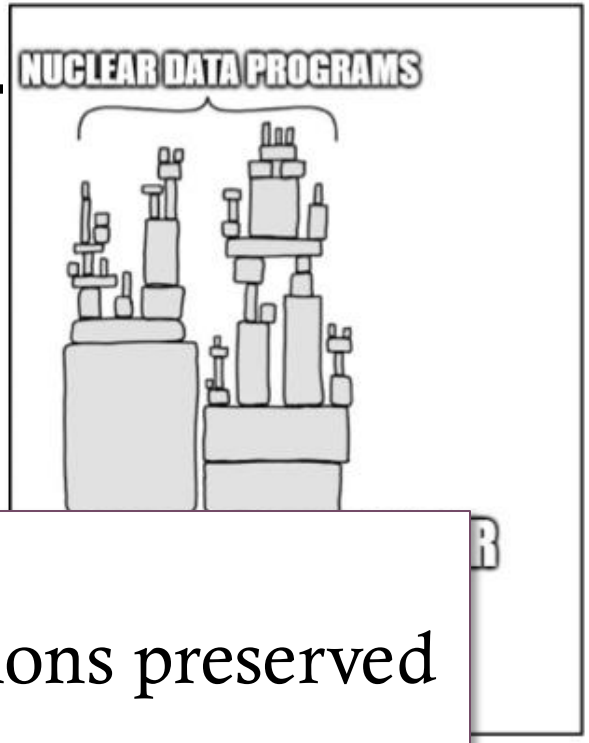
Source: XKCD 2347

Options:

1. **Students** (Model: Nattrass (UTK) & HepData compilation): potential to grow workforce; *needs simpler workflow*, patience and education
2. **More contractors**: *EXFOR compilation currently not enjoyable* so smells like exploitation
3. **Citizen science**: very unskilled so *need extremely simple workflow*
4. **AI**: not ready and needs a lot of development
5. **Experimenters themselves** (HEPData model): *need simpler workflow*
6. **Fail**

EXFOR NEEDS HELP!

- Current compiler workforce is aging out & not being replaced fast enough
- Need more detailed information & expanded scope
- Ne
- Ra



2347

SG-50/54

- Expand information & corrections preserved
- Make EXFOR AI ready
- Try out new compilation approaches
- Need a simpler EXFOR workflow
- *NEEDS RESOURCES*

Optio

1. **S**
   H
   w
   p
2. **More contractors**: *EXFOR compilation currently not enjoyable* so smells like exploitation

ed

a

model): *need simpler workflow*

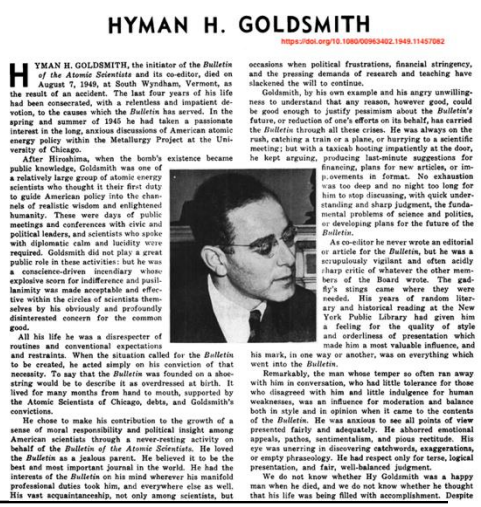
6. **Fail**

DATA ARCHEOLOGY & OFFICE DUMPS

- This has happened
- This will continue to happen
- Arguably it is a failure of succession planning
- When it happens,
  - Saving the data takes time and resources
  - The data may be damaged
  - The data must be triaged
- But is clearly a tremendous opportunity

SUCCESSION PLANNING

- The ND community is built of several long-running (50+ years) collaborations who deliver data to users
- Every step in the pipeline requires domain expertise
- Need to ensure expertise preserved (& grows to meet new needs!)
  - Rule of Two (Sith rule) is not enough
  - Mentoring does not begin just before retirement, takes YEARS
  - Junior staff need leadership roles



HYMAN H. GOLDSMITH

https://doi.org/10.1080/00963402.1949.11457082

HYMAN H. GOLDSMITH, the initiator of the *Bulletin of the Atomic Scientists* and its co-editor, died on August 7, 1949, at South Wyndham, Vermont, as the result of an accident. The last four years of his life had been consecrated, with a relentless and impatient devotion, to the causes which the *Bulletin* has served. In the spring and summer of 1945 he had taken a passionate interest in the long, anxious discussions of American atomic energy policy within the Metallurgy Project at the University of Chicago.

After Hiroshima, when the bomb's existence became public knowledge, Goldsmith was one of a relatively large group of atomic energy scientists who thought it their first duty to guide American policy into the channels of realistic wisdom and enlightened humanity. These were days of public meetings and conferences with civic and political leaders, and scientists who spoke with diplomatic calm and lucidity were required. Goldsmith did not play a great public role in these activities: but he was a conscience-driven incendiary whose explosive scorn for indifference and pusillanimity was made acceptable and effective within the circles of scientists themselves by his obviously and profoundly disinterested concern for the common good.

All his life he was a disrespecter of routines and conventional expectations and restraints. When the situation called for the *Bulletin* to be created, he acted simply on his conviction of that necessity. To say that the *Bulletin* was founded on a shoestring would be to describe it as overdressed at birth. It lived for many months from hand to mouth, supported by the Atomic Scientists of Chicago, debts, and Goldsmith's convictions.

He chose to make his contribution to the growth of a sense of moral responsibility and political insight among American scientists through a never-resting activity on behalf of the *Bulletin of the Atomic Scientists*. He loved the *Bulletin* as a jealous parent. He believed it to be the best and most important journal in the world. He had the interests of the *Bulletin* on his mind wherever his manifold professional duties took him, and everywhere else as well. His vast acquaintanceship, not only among scientists, but occasions when political frustrations, financial stringency, and the pressing demands of research and teaching have slackened the will to continue.

Goldsmith, by his own example and his angry unwillingness to understand that any reason, however good, could be good enough to justify pessimism about the *Bulletin's* future, or reduction of one's efforts on its behalf, has carried the *Bulletin* through all these crises. He was always on the rush, catching a train or a plane, or hurrying to a scientific meeting; but with a taxicab hooting impatiently at the door, he kept arguing, producing last-minute suggestions for financing, plans for new articles, or improvements in format. No exhaustion was too deep and no night too long for him to stop discussing, with quick understanding and sharp judgment, the fundamental problems of science and politics, or developing plans for the future of the *Bulletin*.

As co-editor he never wrote an editorial or article for the *Bulletin*, but he was a scrupulously vigilant and often acidly sharp critic of whatever the other members of the Board wrote. The gadfly's stings came where they were needed. His years of random literary and historical reading at the New York Public Library had given him a feeling for the quality of style and orderliness of presentation which made him a most valuable influence, and his mark, in one way or another, was on everything which went into the *Bulletin*.

Remarkably, the man whose temper so often ran away with him in conversation, who had little tolerance for those who disagreed with him and little indulgence for human weaknesses, was an influence for moderation and balance both in style and in opinion when it came to the contents of the *Bulletin*. He was anxious to see all points of view presented fairly and adequately. He abhorred emotional appeals, pathos, sentimentalism, and pious rectitude. His eye was unerring in discovering catchwords, exaggerations, or empty phraseology. He had respect only for terse, logical presentation, and fair, well-balanced judgment.

We do not know whether Hy Goldsmith was a happy man when he died, and we do not know whether he thought that his life was being filled with accomplishment. Despite

COMMUNICATING VALUE OF ND & HELPING USERS/FUNDERS PRIORITIZE

- Significant resources invested in current ND sets (EXFOR ~$25B > GDP of many countries)
- Difficult to demonstrate $$ impact of a specific data improvement (technically hard + often use is proprietary), **BUT** workflow improvements can help with prioritization
- User problems:
  - don't know role of ND in their problem
  - work around ND shortcomings – need to get them to communicate!
- Need for continued & expanded outreach!