Collaborations in ML: A Study in Scarlet

Joshua Einstein-Curtis

RadiaSoft LLC

March 5, 2024

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Accelerator Research & Development (R&D) and Production (ARDAP), under Award Number(s) DE-SC0024543.

What is the Problem?

Collaborating in machine learning has several challenges to overcome in any domain, but in particular in online systems in closed communities (e.g., particle accelerators, industrial plants)

One particular problem is with code that needs to be *deployed in production*

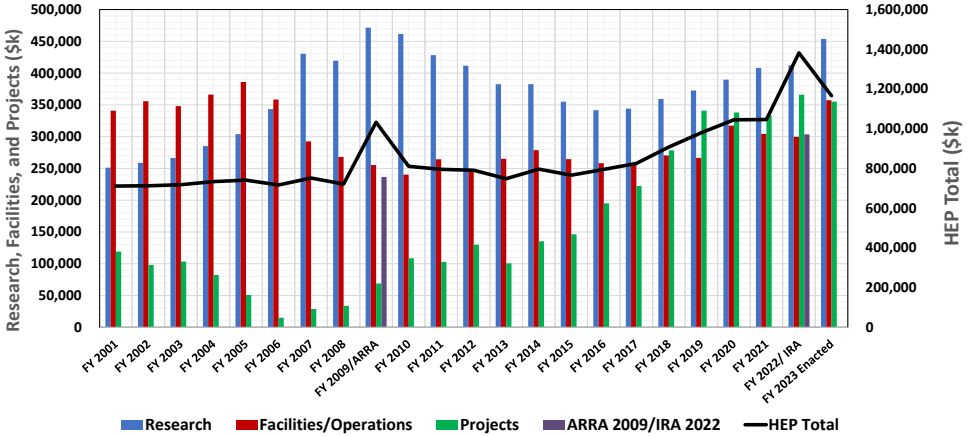
Accelerator operations, in particular, has challenges for collaboration across: regulatory, financial, intellectual property, and, unsurprisingly, people and incentive spheres

Collaborations in Machine Learning: A Project Plan

RadiaSoft has been focusing on collecting data from across the DOE SC labs and industry on their use of, perception of, and plans for machine learning in both operations and wider domains (e.g., EIC, beamlines)

ML tooling is all over the place in research and operations

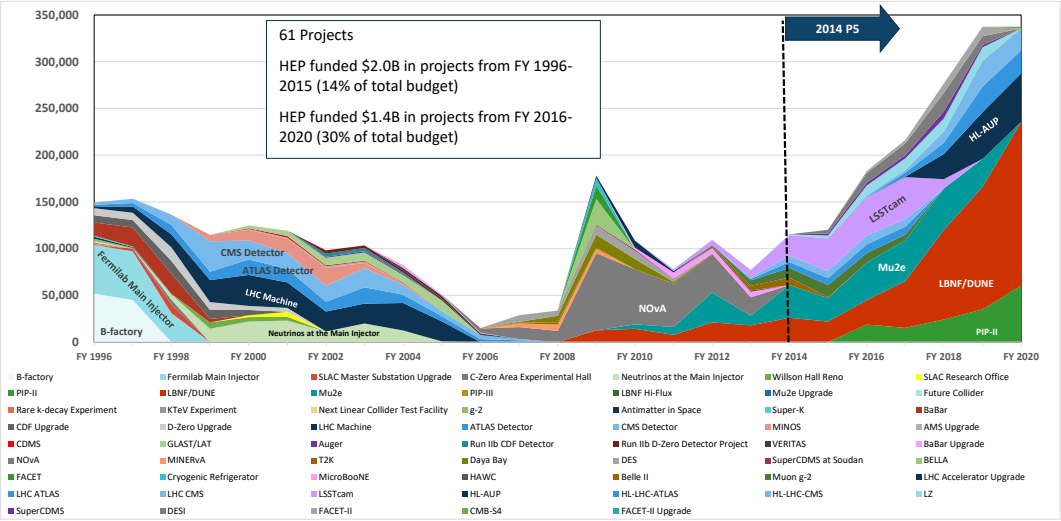
1. No 'standard' in place outside of specific groups
2. Each lab has *part* of a toolchain, but no one has a complete toolchain outside of specific use cases (ie SLAC and loss of personnel, FNAL and own needs, JLab, Argonne)
3. Or in other words: Kubernetes, GeOFF, bespoke, ???
4. Quote: *I think machine learning tools will deprecate even faster [than other software tools]*



ARRA 2009 funds supported Research, Facilities, and Projects
IRA 2022 funds supported Projects only

Historical Chart of HEP Projects

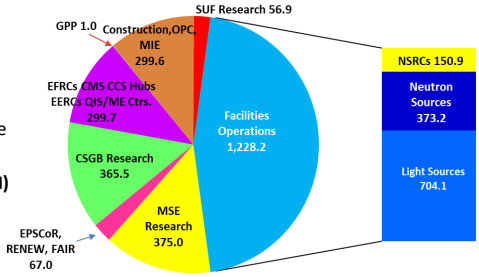
FY 1996 – FY 2020



FY 2024 Request: $2,693M (+$159M or 6.3% above FY 2023 Enacted)

Research programs Δ = +$56.0M

- Continued investments in research for clean energy, manufacturing, microelectronics, critical materials and minerals, BRaVE, and RENEW (+$12M)
- Computational Materials and Chemical Sciences, Energy Innovation Hubs, and National QIS Research Centers continue ($119.7M)
- **Establish Microelectronics Science Research Centers (+$25M)**
- Energy Frontier Research Centers continue ($130M)
- **Expanded investments in SC Energy Earthshots initiative (+$35M)**



Scientific user facilities Δ = +$165.9M

- Operations of 12 facilities supported at ~90% of funding required for re-baselined, normal operations ($1,228.2M)
- Facilities research ($56.9M, +$7M): Accelerator & Detectors; AI/ML; BRaVE

Construction/MIE Δ = -$63.1M (includes OPC)

- LCLS-II-HE ($120M); ALS-U ($57.3M); PPU ($15.8M); STS ($52M); CRMF ($10M)
- **New starts: HFIR Pressure Vessel Replacement ($13M); NEXT-III ($6.6M)**
- MIEs: NSRC Recap ($5M); NEXT-II ($20M)

What is a Collaboration?

A collaboration, in this project's definition, is not the same as a business

A collaboration needs a science goal or mission to be resilient in the long-term

Accelerator operations supports multiple science missions, possibly at the same facility; this *needs to be clear* to funding agencies and external partners

What is a *Sustainable* Collaboration?

"A sustainable collaboration is one that continues to grow."
– *Andy Götz, Tango Controls*

But:

- What does this mean for collaborations grown out of project-based funding (e.g., Ecascale Computing Project, MLExchange)?
- What about projects with no evangelist? – especially true with open source

People

People are the drivers of success. If they don't (or won't) work together or follow processes or agreements, a collaboration isn't possible.

We are fighting against silos, *cant* (*argot*, jargon), and 'experts' that know better themselves when attempting to change a culture

What We've Done

Interviewed a number of scientists, engineers, and managers across the DOE accelerator space, including 2 industry partners, 1 DOE program, and 1 lab tech transfer office

Research artifacts from 64 collaborations, from across the landscape of research institutions, FFRDCs, standards organizations, consortiums, and open source organizations, have been collected, including: IP, legal, and governance documents

Analyzing Collaborations

We selected 13 of 62 select public and private *collaborations* based on initial research:

1. Software preferences among accelerator facilities are varied, having been born and developed by the facilities themselves – I'm looking at you, EPICS
2. What is the financial incentive/motivation of all stakeholders in the collaboration – quid pro quo?
3. Do quality assurance standards differ between public and private organizations - standards?
4. What is the nature of IP to be shared – software, know-how, etc. for facilities operations and what is the potential loss to the different types of collaboration members?

How governance structures differ

Formal v Informal governance, Multi-party vs Bilateral, ...

Epics	Formal governance, minimal organization buy-in to central committee, relies on contributors to open source
Tango Controls	Formal governance, buy-in required for voting rights, formal agreements in place among consortia members
Exascale Computing Project	Funding just ended, future of individual projects uncertain

Machine Learning in Operations: An Overview

Need for higher reliability and closed off networks

Need for tooling with long-term support – but ML/AI (and other emergent technologies) require fast iteration

Different concerns for knowledge sharing, IP law, and ownership – especially as operations operates under different funding from experiments

Types of Machine Learning in Operations

Real-time image processing, trigger algorithms, fast control

1 Hz – <1kHz data

Requires offline processing: experimental data, large workloads, physics simulations

Operational domains and collaboration concerns

Most funding seems to be available for 'offline' AI, such as LLMs and image processing

The (US) National AI Research Resource Pilot report[1] is of particular interest as it does not introduce mechanisms for handling ML and AI usage in infrastructure

In addition, there are particular inter- and intra-agency concerns with handling data transfer between infrastructure owned by different groups at the same facilities

[1] https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf

Machine Learning in Operations: ML/AI Centers

Several labs have become key, unofficial "centers" for machine learning in accelerator operations. These include: CERN, SLAC, DESY, and Argonne

Even with each institution having their own simulation codes, LBNL has been actively providing support for several, including WarpX

So many simulation codes...[2]

[2] https://en.wikipedia.org/wiki/Accelerator_physics_codes

Machine Learning in Operations: What is Currently Used?

Optimization routines, both for design and minimizing tuning and downtime, are currently the 'bread and butter' of machine learning in this space

There is an increasing need for simplified models and high-speed parallel simulations, as well as increasing interest in online fault detection and analysis

Is there really a need to minimize the involvement of subject matter experts? Or do we want the "job protection" to keep building a workforce?

Our Findings

We currently believe several possible models **can** exist for a collaboration in ML; but many of these models would require careful consideration of interrelated funding mechanisms and how to meet the requirements of users

There is a particular problem for *formal* collaborations with international partners, including: Export control, international treaties, IP law

Best Paths Forward

We currently believe that one of these two models is a likely best path forward for such a collaboration:

- External NGO/Tango model
  - Allows for multi-party governance and dedicated project resources
- Linux Foundation non-profit/Institute
  - Allows for dedicated resources with formalized governance and agreements
  - A new foundation within the Linux Foundation framework has recently been established for HPC projects[3]; can something similar be done within or for the accelerator community?

[3] `https://hpsf.io/`

Most Likely Paths Forward

We currently believe that it is more likely that one of these models is what is actually feasible:

- NAIRR NSF ML sub-project or research center
  - NAIRR Pilot Program (`https://nairrpilot.org/`) and NAIRR Secure (`https://nairrpilot.org/nairr-secure`)
  - More complicated funding requirements and management with a need for clear mission goals
- New ASCR SciDAC Institute
  - Requires ASCR buy-in and possible redefinition of how critical computing resources would fall under their purview
- Status quo
  - Easiest path forward
  - Each lab independent outside of taking advantage of personal connections made at conferences

Challenging Models

We believe these models are unlikely to be workable due to either IP, political, or funding reasons:

- Single, lab-based center
  - Restricts project resources long-term to a single lab managing project direction
  - Very transactional relationship for parties invovled
  - Governance complicated by Contractor/sub-contractor relationships
- Single, industry-based center
- Industry- or Subcontractor-led consortium

Conclusions

A successful collaboration requires:

1. Strong leadership with directed goals
2. An ability to share data and ideas in a standard manner; especially true if working with industry
3. An agreement on how to govern the collaboration and any resultant products

Just having an agreement on data storage and transport standards would be highly valuable

Whether this is driven by meetings like this one or more formally (*de facto*) is up to us to decide[4]

[4] See extra slides for interview notes worth mention

Disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

IP Complexity

IP agreements and licensing concerns (ie copy-left v reserving copyright)

```
Copyright and License Terms
---------------------------

Copyright (c) 2008-2016 Mark L. Rivers
Copyright (c) 2008-2016 The University of Chicago
Copyright (c) 2008-2016 UChicago Argonne LLC, as Operator of Argonne
                        National Laboratory.
Copyright (c) 2008-2016 Brookhaven Science Associates, as Operator
                        of Brookhaven National Laboratory
Copyright (c) 2008-2016 Diamond Light Source Limited,
                        (DLS) Didcot, United Kingdom
Copyright (c) 2013-2016 UT-Battelle LLC
```

Epics areaDetector code copyrights

Interview Notes I

The following present some examples of key information provided during interviews:

- Operators are in their own world – most science groups seem to treat them as resources to use, but not as a part of the dev process inherently
- 'Seems clear we are not going to bring in professional AI/ML developers as staff'
- Some labs have increasingly siloed off different parts of their computing resources, either due to access controls or regulations, leading to some significant challenges for collaboration (e.g., BNL Indico)
- On a new project: 'So the people building it have ideas on what they want to do...' but haven't decided on implementation
- 'DOE - everything is aimed at commissioning. They don't care about operations.'
- 'Not sure where ML fits in the funding profile with DOE'
- No one knows the ownership or licensing rules for their ML artifacts outside of tech transfer – and even then, it is case-by-case

Interview Notes II

- About infrastructure investment: DOE way behind in scaling when compared to industry (Amazon, Google, DOD) "That horse is behind the barn and out to pasture"
- For online control, don't need to pass data, just need to pass techniques
- Don't underestimate the social aspect of implementing ML and techniques
- From industry about lab subcontractors: 'DOD changes their subcontractors they way they change shoes'