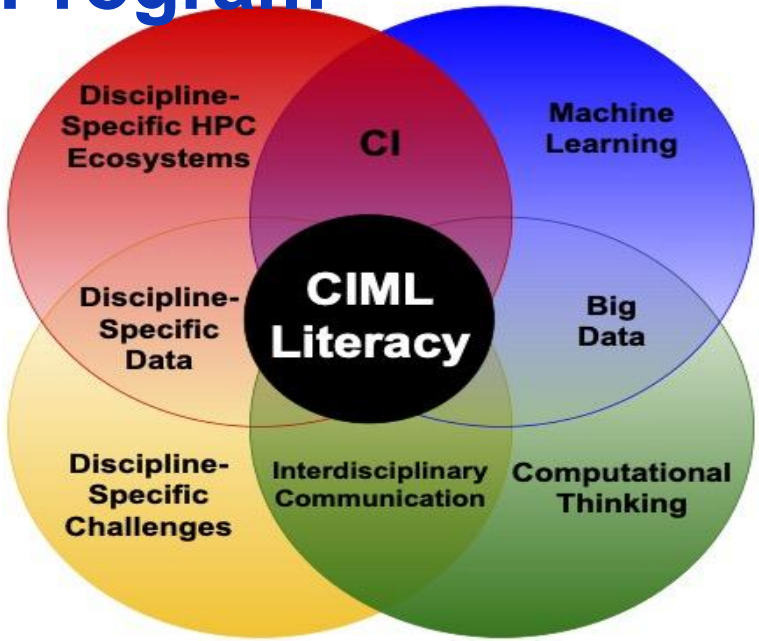
Expanding the AI Curricula: The Cyberinfrastructure-Enabled Machine Learning (CIML) Training Program



Presented by Mary P. Thomas
At the Sixth National Research Platform (6NRP) Workshop

January 29, 2025

NSF Award 1928224

SDSC SAN DIEGO SUPERCOMPUTER CENTER
UC San Diego

About Cyberinfrastructure-Enabled Machine Learning (CIML) Training Program

- **Objectives**: teach skills needed for ***Scalable Machine Learning***
- ***Facilitate*** researchers/educators using ML and big data analytics methods for their domain specific applications or instructional material
- ***Create generalized ML training*** and project materials that run on large-scale NSF funded cyberinfrastructure resources (NRP, ACCESS)
- ***Synthesize*** training material into a domain independent CIML workflow system that can be used for creating applications that run on the NSF HPC ecosystem.
- **Develop a *community*** of users who actively contribute training material & incorporate the materials into their projects and courses → HPC-ED.

Defining Core Competencies

- Parallel computing concepts
- Hardware for AI Computing
- Software Containers
- Conda environments and Jupyter notebooks
- Scalable Machine Learning & Deep Learning

Overview of the CIML Summer Institute

- Annual, four-day event: a preparation day and 3 days of interactive training.
- To date, we have held 3 summer institutes: 2021, 2022, and 2023
- COVID impact: 2021 and 2022 were virtual:
  - feedback from participants indicate that the training was successful and has been continuing to impact their research.
- CIML SI23 was held in-person; ~60 participants/day
  - 2 participants contracted COVID 🙁

Typical CIML Institute Agenda

Day 1: Preparation Day		Day 2: HPC and Parallel Computing Concepts	
1.1	Welcome and Orientation	2.2	Introduction to HPC Cyberinfrastructure
1.2	Accounts, Login, Running Jobs. Portal	2.3	CPU Computing - Hardware, arch, software infr.
1.3	Running Jupyter Notebooks on Expanse	2.4	Data Managment and File Systems
		2.5	GPU Computing - Hardware, arch, software infr.
Day 3: Scalable Machine Learning		**Day 4: Deep Learning**	
3.2	Introduction to Singularity	4.2	Intro to NN/CNN
3.3	CONDA envs&Notebooks-->Scalable & Repro Data Exploration	4.3	Deep Learning
3.4	Machine Learning (ML) Overview	4.4	Deep Learning Layers and Models
3.5	R on HPC Demo	4.5	Deep Learning Transfer Learning
3.6	Spark	4.6	Deep Learning - Other Topics

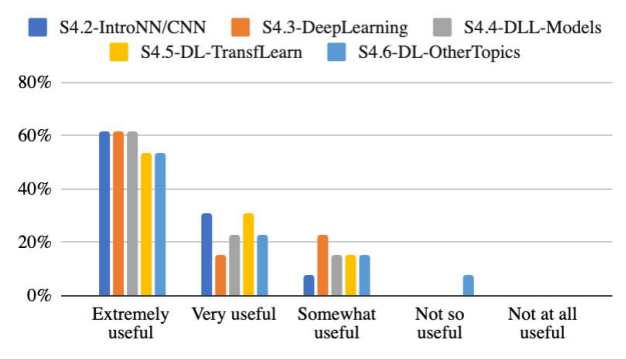
- Prep Day usually held 1 week before –ensures participants can access HPC system, have accounts, etc.

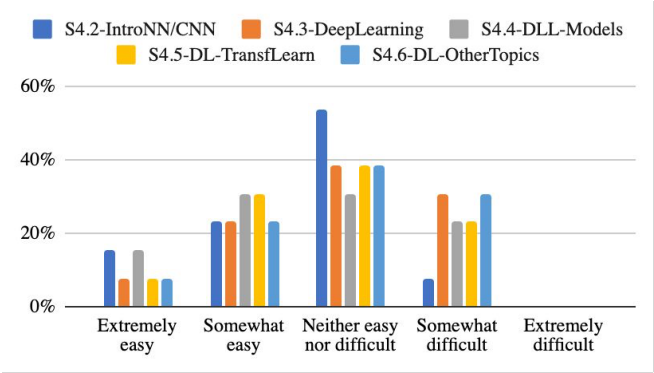
Participant Surveys

- Occur on the day of training, immediately after completion of the institute, and 1 or 2 years later.
- Designed using a 5-point Likert impact scale evaluating user sentiments such as usefulness or satisfaction.
- Less than 10 questions

Session Surveys

- Introduction to NN/CNN
- Over 80% of participants found topic be “very to extremely useful”
- Most found instruction level to be neither easy nor difficult, while about 20% found it to be somewhat difficult.





CIML Impact Survey

- and was sent out to 78 participants from SI21 and SI22. The response rate to date is 24\% and 13\% failed (were rejected by email servers). Of the respondents, 38\% attended CIML SI21, and 63\% attended CIML SI22.
- The overall level of satisfaction (satisfied or very satisfied) for both groups is over 80\% and 93\% were likely or highly likely to recommend CIML to their colleagues.
- This is evidenced by the fact that over 87\% of the respondents use ML frequently in their research and that 59\% are using the scalable AI tools taught at the institute (Pytorch, Spark, SciKitLearn, Tensorflow). This high-level of adoption by our participants indicates the usefulness our training program.

Training Material repository

- CIML training materials are designed to be used to train other users as part of the overall CIML CyberTraining program.\cite{ciml-proj}
- As part of the SDSC HPC Training program, all training materials and event information are collected and hosted on the SDSC HPC Training Material Catalog page. \cite{hpc-training-catalog} As mentioned above, our training program applies the FAIR
- (Findable, Accessible, Interoperable, and Reusable) concepts to increase the accessibility and reproducibility of training materials and research findings. \cite{Garcia2020}
- Training material and project data (build scripts, libraries, code, notebooks, example data sets), and event type (workshops, webinars, hackathons, and meetings), are collected and stored on a searchable project websites and repositories (e.g.GitHub), that are open source and made available to the public. \cite{ciml-repos} Participants will be encouraged to contribute content to CIML.
- Through these efforts, the CIML project contributes to the NSF goal of "democratization of AI" which aims to increase access to and participation in AI research and development across different fields and communities. \cite{Parashar2022}

Recommendations

- As software tools change, become easier to use, and new tools emerge, CIML will continue to evolve its training program, to enable scientist to find good paths for applying machine learning, deep learning, and AI techniques at scale.
    - we will add session on how to use NRP
- It is essential to not only teach introductory AI topics (ML, DL…) we need to find more ways to support domain experts who need to improve/optimize their models
    - more domain specific training proposals (NSF CyberTraining)
- PNRP is a good resource to develop models where optimization and speedup needs to be done, and that prepares

Thank You!


Science/HASS Disciplines
Computational Ethics
CI
Machine Learning
CI/ML Literacy
Big Data
Computational Thinking
Technologies
Policies/Policy