Cyberinfrastructure-Enabled Machine Learning Summer Institute

CIML SI25 - Day 2

June 24, 2025

Logistics and Introductions

HPC, Parallel Concepts

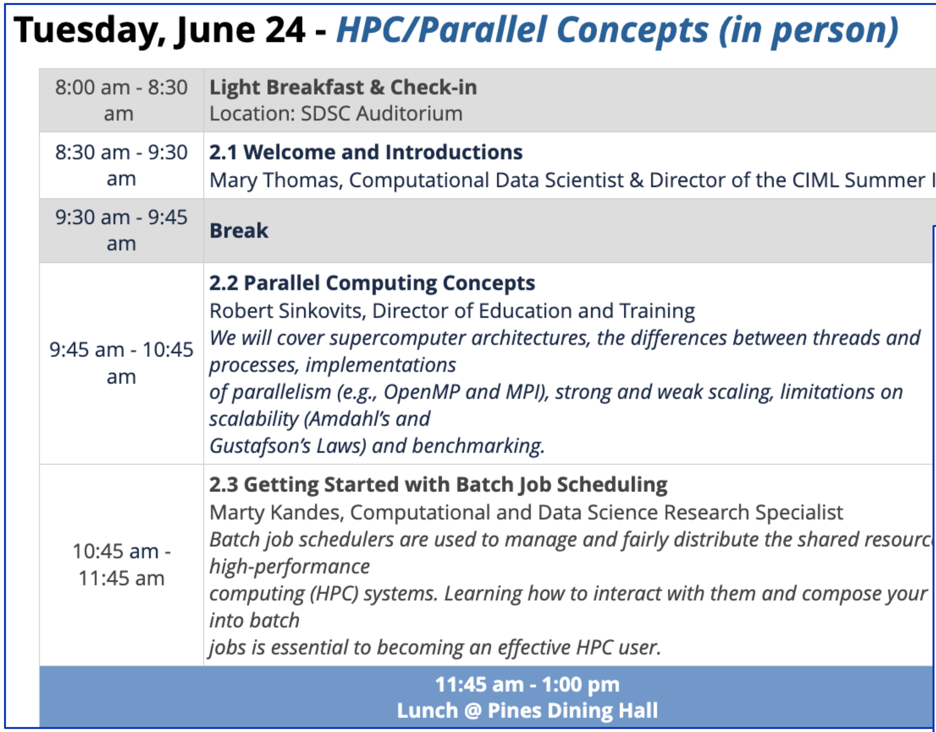
What is CIML?

- NSF CyberTraining Grant: *Developing a Best Practices Training Program in Cyberinfrastructure-Enabled Machine Learning Research (CIML)*
- Objectives: Scalable Machine Learning
    - To create generalized machine learning training and project materials that run on large-scale NSF funded cyberinfrastructure resources such as XSEDE
    - Targeted towards researchers and educators who are using machine learning (ML) and big data analytics methods for their domain specific applications or instructional material
    - To develop a community of machine learning and data analytics CI Users (CIU) and Contributors (CIC) who actively contribute to the training material repository and incorporate the materials into their projects and courses.
    - Synthesize the training material into a domain independent CIML workflow system that can be used for creating applications that run on the NSF HPC ecosystem.

Welcome to the FIFTH CIML Summer Institute!

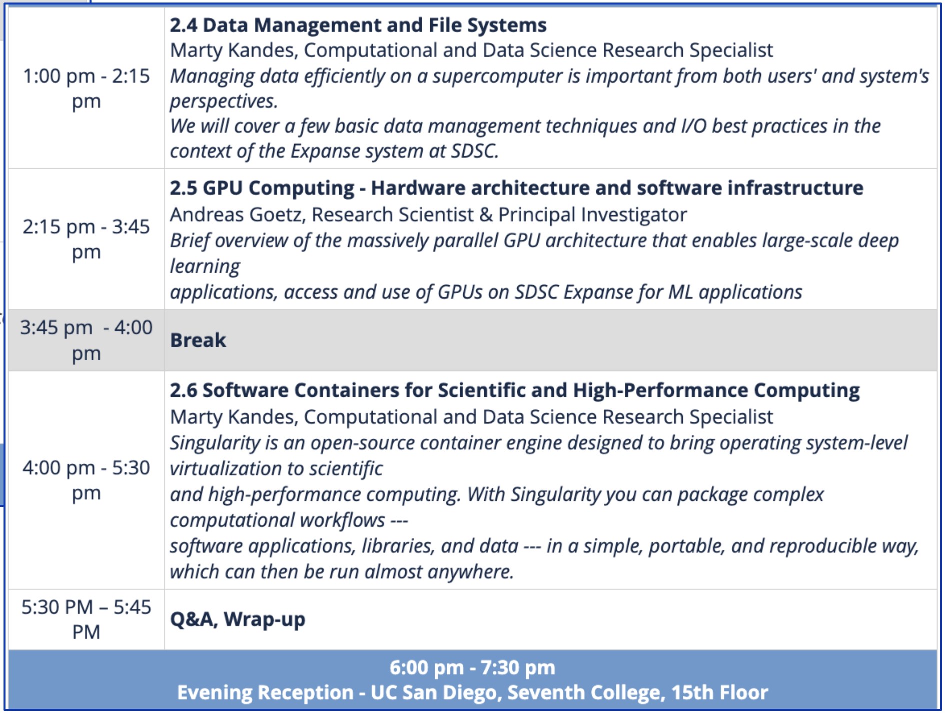
- Focus is on scalable machine learning.
- GitHub: https://github.com/ciml-org/ciml-summer-institute-2024
- Please be on time so we can stay on schedule.

Day 2 Agenda: *HPC, Parallel Concepts*



Tuesday, June 24 - *HPC/Parallel Concepts (in person)*

8:00 am - 8:30 am	**Light Breakfast & Check-in** Location: SDSC Auditorium
8:30 am - 9:30 am	**2.1 Welcome and Introductions** Mary Thomas, Computational Data Scientist & Director of the CIML Summer Institute
9:30 am - 9:45 am	**Break**
9:45 am - 10:45 am	**2.2 Parallel Computing Concepts** Robert Sinkovits, Director of Education and Training *We will cover supercomputer architectures, the differences between threads and processes, implementations of parallelism (e.g., OpenMP and MPI), strong and weak scaling, limitations on scalability (Amdahl's and Gustafson's Laws) and benchmarking.*
10:45 am - 11:45 am	**2.3 Getting Started with Batch Job Scheduling** Marty Kandes, Computational and Data Science Research Specialist *Batch job schedulers are used to manage and fairly distribute the shared resourc high-performance computing (HPC) systems. Learning how to interact with them and compose your into batch jobs is essential to becoming an effective HPC user.*
11:45 am - 1:00 pm **Lunch @ Pines Dining Hall**	



1:00 pm - 2:15 pm	**2.4 Data Management and File Systems** Marty Kandes, Computational and Data Science Research Specialist *Managing data efficiently on a supercomputer is important from both users' and system's perspectives.* *We will cover a few basic data management techniques and I/O best practices in the context of the Expanse system at SDSC.*
2:15 pm - 3:45 pm	**2.5 GPU Computing - Hardware architecture and software infrastructure** Andreas Goetz, Research Scientist & Principal Investigator *Brief overview of the massively parallel GPU architecture that enables large-scale deep learning applications, access and use of GPUs on SDSC Expanse for ML applications*
3:45 pm - 4:00 pm	**Break**
4:00 pm - 5:30 pm	**2.6 Software Containers for Scientific and High-Performance Computing** Marty Kandes, Computational and Data Science Research Specialist *Singularity is an open-source container engine designed to bring operating system-level virtualization to scientific and high-performance computing. With Singularity you can package complex computational workflows --- software applications, libraries, and data --- in a simple, portable, and reproducible way, which can then be run almost anywhere.*
5:30 PM – 5:45 PM	**Q&A, Wrap-up**
6:00 pm - 7:30 pm **Evening Reception - UC San Diego, Seventh College, 15th Floor**	

Basic Information

- Today is Day 1: Welcome-Orientation Day:
  - https://github.com/ciml-org/ciml-summer-institute-2025/tree/main/1.1_welcome_and_orientation
- Check out the 0_Prep doc – contains important information:
  - https://github.com/ciml-org/ciml-summer-institute-2025/tree/main/0_preparation
- Expanse User Guide:
  - https://www.sdsc.edu/support/user_guides/expanse.html
- You need to have an Expanse account in order to access the system. There are a few ways to do this:
  - You have been assigned a training accounts: they expire, save your data.
  - Request an Expanse trial account, send email to consult@sdsc.edu
  - Reach out to a PI with an active allocation can add you to their allocation.
  - Submit a proposal through the ACCESS allocation request system. https://allocations.access-ci.org
- Online Expanse training repo and information:
  - https://github.com/sdsc-hpc-training-org/expanse-101
  - https://hpc-training.sdsc.edu/expanse-101/

Other Resources

- GitHub Repo for this webinar: clone code examples for this tutorial – clone example code:
  - https://github.com/ciml-org/ciml-summer-institute-2025
- SDSC Training Resources
  - https://www.sdsc.edu/support/user_guides/expanse.html
  - https://education.sdsc.edu/training/interactive/
  - https://www.sdsc.edu/events/index.html
  - https://github.com/sdsc-hpc-training-org/hpctr-examples
- ACCESS Training Resources
  - https://support.access-ci.org/events

CIML Instructors

Andreas Goetz, Ph.D.
Director of Computational Chemistry Laboratory

Marty Kandes, Ph.D.
Computational and Data Science Research Specialist

Mai Nguyen, Ph.D.
Lead for Data Analytics

Paul Rodriguez, Ph.D.
Computational Data Scientist

Robert Sinkovits, Ph.D.
Director of Education and Training

Mary Thomas, Ph.D.
Computational Data Scientists, HPC Trainer

Let’s get to know each other

1. Name
2. Institution/Company & Department
3. How do you like to spend your time when not at work?
4. What have you binged watched or read?

We hope you all have a great workshop!