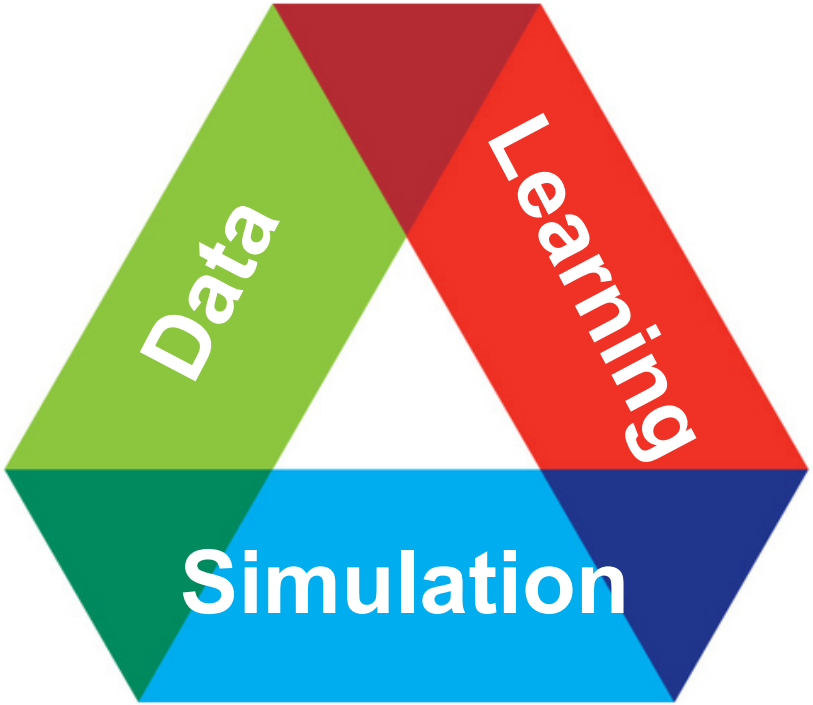
ALCF Data Science Program Overview

Elise Jennings
Argonne Leadership Computing Facility
Data Sciences Group
Argonne National Laboratory
ejennings@anl.gov

Argonne
NATIONAL
LABORATORY

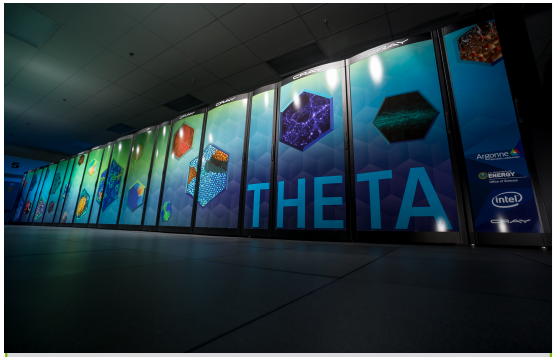
ALCF DATA SCIENCE PROGRAM (ADSP)

- “Big Data” science that require the scale and performance of leadership computing
- Projects cover a wide variety of application domains that span computational, experimental and observational sciences
- Focus on data science techniques including but not limited to statistics, machine learning, deep learning, UQ, image processing, graph analytics, complex and interactive workflows



ALCF DATA SCIENCE PROGRAM (ADSP)

- Two-year proposal period. PIs will be required to fill out a renewal application for each allocation period of the award.
- Proposals will target **science** and **software technology** scaling for data science
- Review process
  - potential impact, data scale readiness, diversity of science domains and algorithms
  - emphasis on projects that can use the architectural features of Theta



Theta Intel/Cray
3,624 nodes
231,935 cores
56 TB MCDRAM
679 TB DDR4
453 TB SSD
Peak flop rate: 9.65 PF

Mira IBM BG/Q
49,152 nodes
786,432 cores
786 TB RAM
Peak flop rate: 10 PF

ALCF DATA SCIENCE PROGRAM (ADSP)

- ADSP Resources
  - **Staff and Postdoc Support:** The chosen ADSP projects will receive part-time support from ALCF staff. Tier-1 projects will be supported in part with postdoctoral scholars.
  - **Training and Hardware Access:** Targeted training for the ADSP projects. Detailed introduction to the hardware and software stack, access to early hardware, deep dives on specific hardware features, and customized tutorials.
  - **Computing and Storage Resources:** Compute time and storage space on Theta, Mira, as well as visualization and analytics clusters. Range from 50M -150M core-hours. Storage may be up to 100 TB

ADSP Program
Call for Proposals in April 2018

https://www.alcf.anl.gov/alcf-data-science-program

Priority research directions (PRD) from the DOE ASCR ML workshop Feb 2018

Grand challenges and priority research directions

Explainability & model selection	• Interpretable ML
Exploiting *a priori* scientific information & constraints	• Effective features for scientific ML • Leveraging domain knowledge & constraints in ML formulation
Performance, training, high dimensions, & big data	• ML in high dimensions • ML for enhancing data collection & use on DOE facilities • ML for inverse problems and inverse problems for ML • Addressing the complexity of DOE applications & modern architectures • ML-Enabled Adaptive Scientific Computing
Quantifying ML limits, validity, and reproducibility	• Reproducibility (stability) of ML • Quantifying the Discrepancy in QoI Derived Using ML (accuracy)

DOE Scientific Machine Learning

Research Themes / meta-ML types	Inner ML: Scientific Inference & Scientific Data Analysis	Coupled ML: ML-Hybrid Algorithms & Models	Outer ML: Automated Decision-Support, Adaptivity, Resilience, Control
Explainability & model selection	Interpretable ML		
	Effective features for scientific ML		
Exploiting *a priori* scientific information & constraints	Leveraging domain knowledge & constraints in ML formulation		
	ML in high dimensions		
	ML for enhancing data collection & use on DOE facilities		
Training, high dimensions & big data	ML for inverse problems and inverse problems for ML		
	Addressing the complexity of DOE applications & modern architectures		
	ML-Enabled Adaptive Scientific Computing		
Quantifying ML limits, validity, reproducibility	Reproducibility (stability) of ML		
	Quantifying the Discrepancy in QoI Derived Using ML (accuracy)		

Thank you !

Upcoming Program Deadlines

Aurora Early Science Program for
Learning and Data
Call for Proposals in January 2018

ADSP Program
Call for Proposals in April 2018