Data Science at Argonne Leadership Computing Facility (ALCF)

Elise Jennings,

Venkat Vishwanath, Michael E. Papka

ejennings@anl.gov

ALCF Data Science Program (ADSP)

The ADSP program started in 2016 and is now in its 4th year.

ADSP’s goal is to support “Big Data” science that require the scale and performance of leadership computing.

Successful projects have

- high potential impact
- data scale readiness
- diversity of science domains and algorithms
- can fully exploit the architectural features of Theta

Two-year proposal period. Yearly call for proposals.

ALCF Data Science Program (ADSP)

Two main targets for development

Science applications

Tools

To date the majority of proposals have been science applications.

Tool development and support is becoming a major requirement.

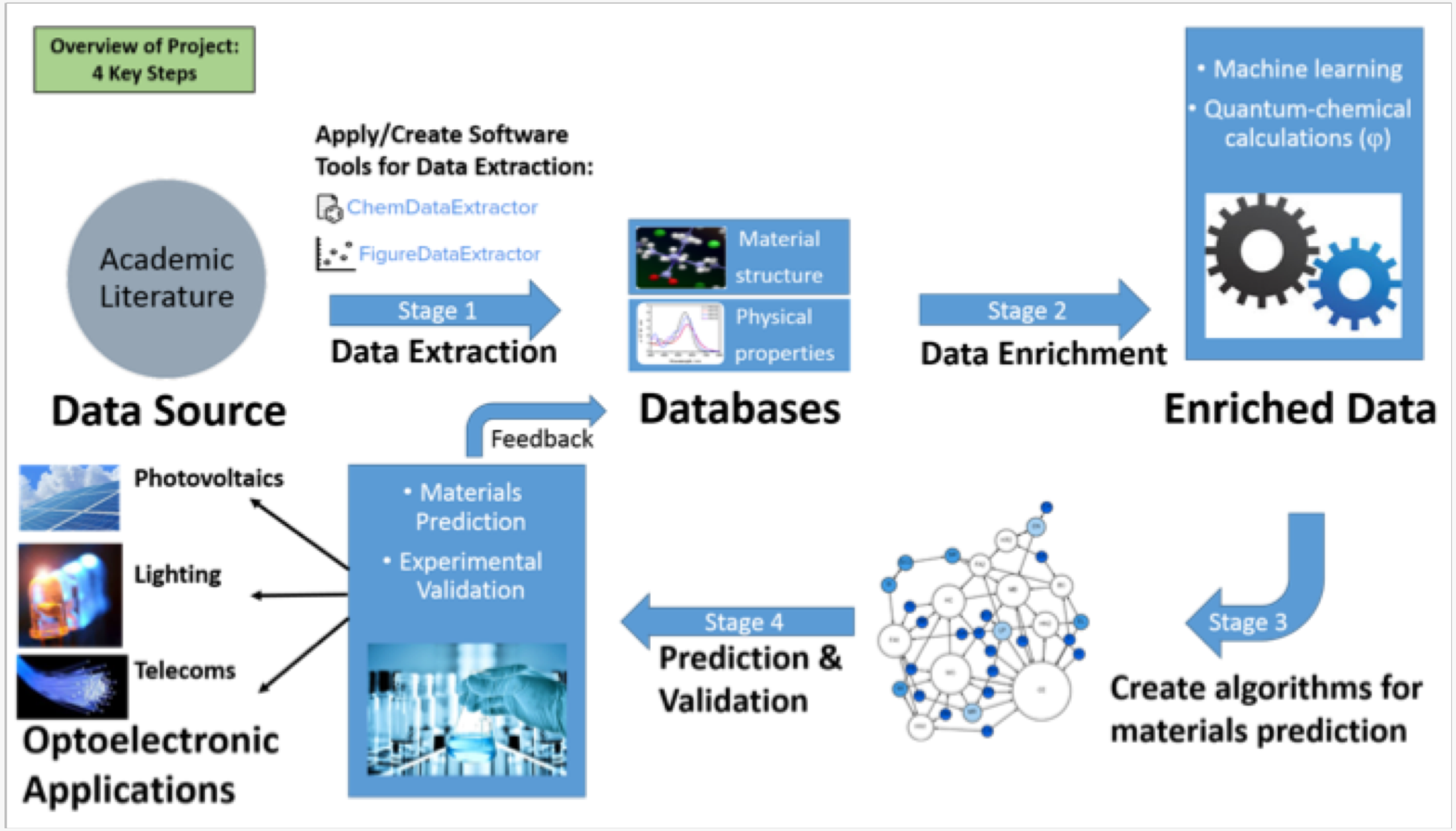
ADSP projects

- span a diverse set of science domains (Materials, Imaging, Neuroscience, Engineering, Combustion/CFD, Cosmology).
- involve large science collaborations (multiple APS, LSST, DESC, LIGO, DES, ATLAS) and smaller research groups developing ML at scale.
- have used nearly 300M core hours on Theta (26% as capability runs)

Data driven science

Molecular Engineering of Solar-Powered Windows

PI: Cole, Cambridge University, ANL



Machine Learning and simulations to aid in the discovery of better light-absorbing dye molecules.

High impact results

- Discovery of a new class of dye-sensitized solar cells (DSSC)
- Fabrication of five DSSC proposed by ML which have been validated experimentally.

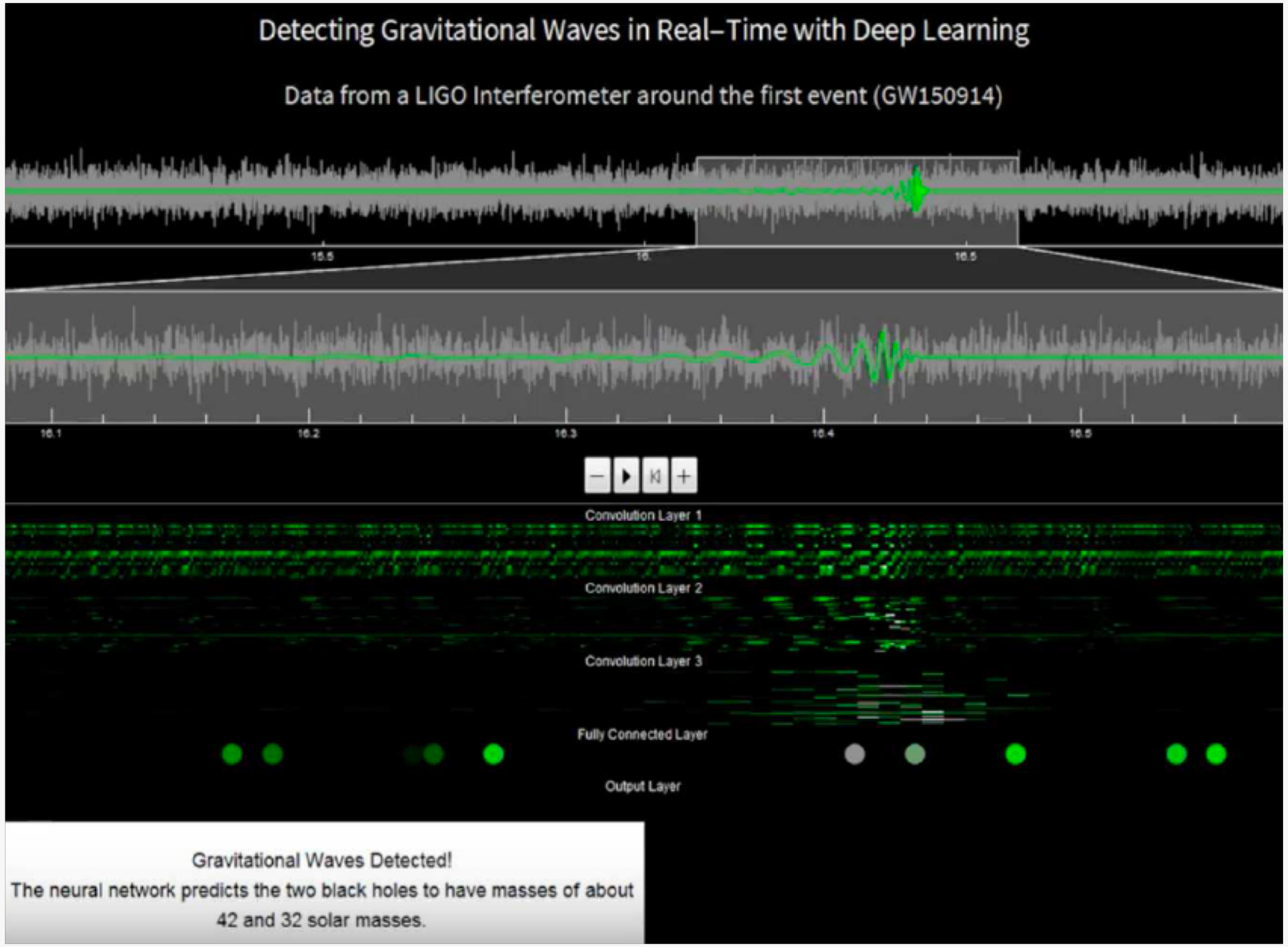
J. M. Cole *et al*, *ACS Appl. Mater. Interfaces*, **2017**, *9* (31),25952

Design-to-Device Approach Affords Panchromatic Co-Sensitized Solar Cells, Chris Cooper et al., Advanced Energy Materials, Dec 6 2018)

Deep Learning driven science

Multimessenger Astrophysics through the NCSA-Argonne Collaboration

PIs: Huerta, Zhao, Haas, Saxton (NCSA)



Detecting Gravitational Waves in Real–Time with Deep Learning

Data from a LIGO Interferometer around the first event (GW150914)

Gravitational Waves Detected!
The neural network predicts the two black holes to have masses of about 42 and 32 solar masses.

Novel data-parallel deep learning fusing HPC and AI for MultiMessenger Astrophysics (MMA).

Huge potential for scientific discovery

- Convergence of all-sky GW observations (LIGO) with deep, high-cadence electromagnetic observations (LSST)
- Novel visualization of Neural Networks

Deep Learning for Multi-Messenger Astrophysics. A Gateway for Discovery in the Big Data Era, Huerta et al., Nature Review Physics

Emerging trends

We now have a mix of applications at scale

- HPC simulations, Big Data analytics and ML

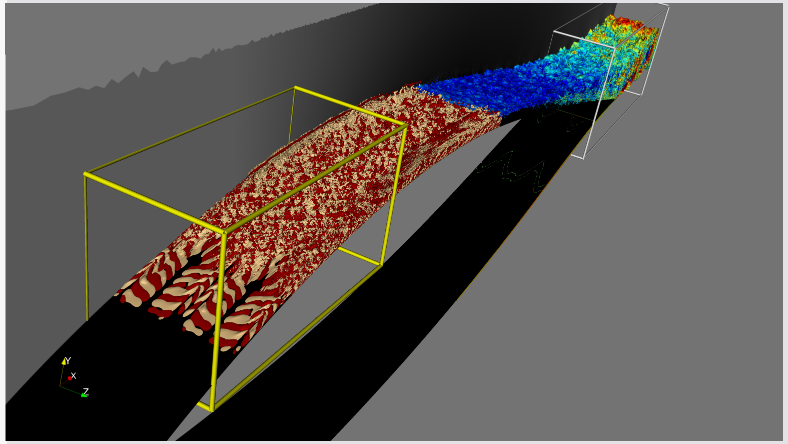
Deep integration of HPC simulations and Machine/Deep Learning

- Augment training data, provide supervised labels for training
- ML model can be embedded into the simulation
- Speed/accuracy trade off in replacing first principal model with ML

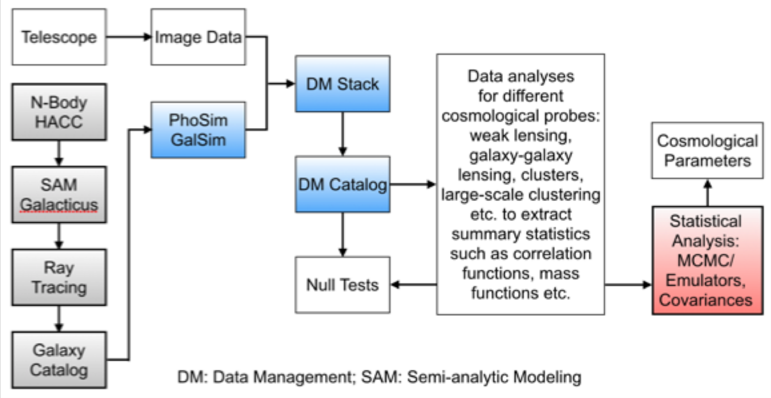
Big drive for

- Scientific techniques (Uncertainty Quantification, reproducibility etc.)
- Image processing, in situ analysis and visualization
- Complex and interactive workflows with performance capabilities
- Smart configuration space sampling
- Tools to couple simulations with analysis and ML

Enabling Multi-Scale Physics for Industrial Design using Deep Learning Networks.
PI: Bhaskaran, GE Global Research



Realistic Simulations of the The Large Synoptic Survey Telescope (LSST) Survey at Scale
PI: Heitmann, Argonne with SLAC, U.Penn, Duke, LANL



Telescope

Image Data

N-Body HACC

PhoSim GalSim

DM Stack

SAM Galacticus

DM Catalog

Ray Tracing

Galaxy Catalog

Null Tests

Data analyses for different cosmological probes: weak lensing, galaxy-galaxy lensing, clusters, large-scale clustering etc. to extract summary statistics such as correlation functions, mass functions etc.

Cosmological Parameters

Statistical Analysis: MCMC/ Emulators, Covariances

DM: Data Management; SAM: Semi-analytic Modeling

Emerging needs

Huge growth in scientific data and learning projects

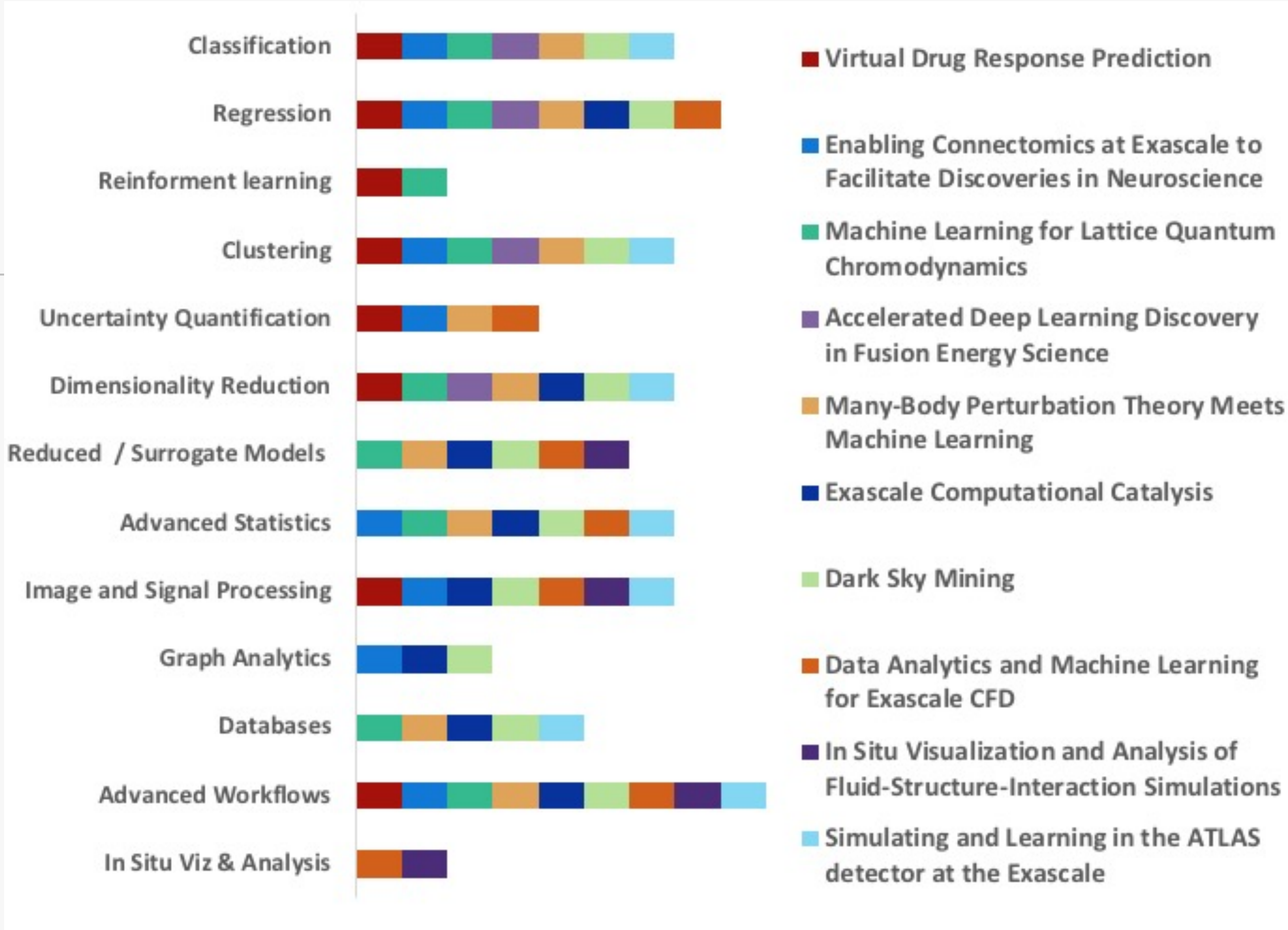
- Director’s Discretionary, ADSP, ESP allow projects to scale up.
- INCITE 2018 included Data and Learning proposals. What we learned in ADSP informed review process.
- Early Science program now has Data and Learning component.
- ALCF workshops, CORAL2 benchmarks, MLPERF etc.

Clear need within the scientific community for

- Hardware and system architectures which support Data and Learning.
- Integrated computing, acceleration and storage with a common software stack.
- Improved scheduling policies, remote data streaming and workflows for large scale campaigns.

https://www.alcf.anl.gov/training

AURORA ESP Data and Learning Methods



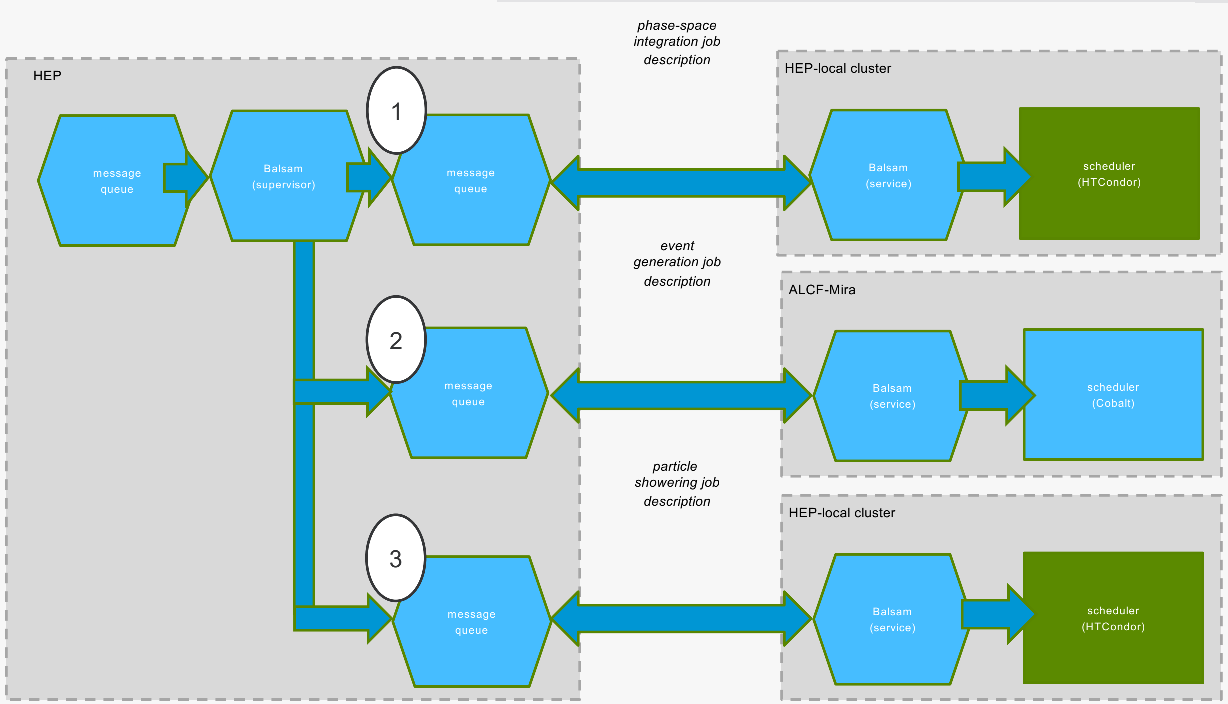
Learning

Data

High impact Data Science software

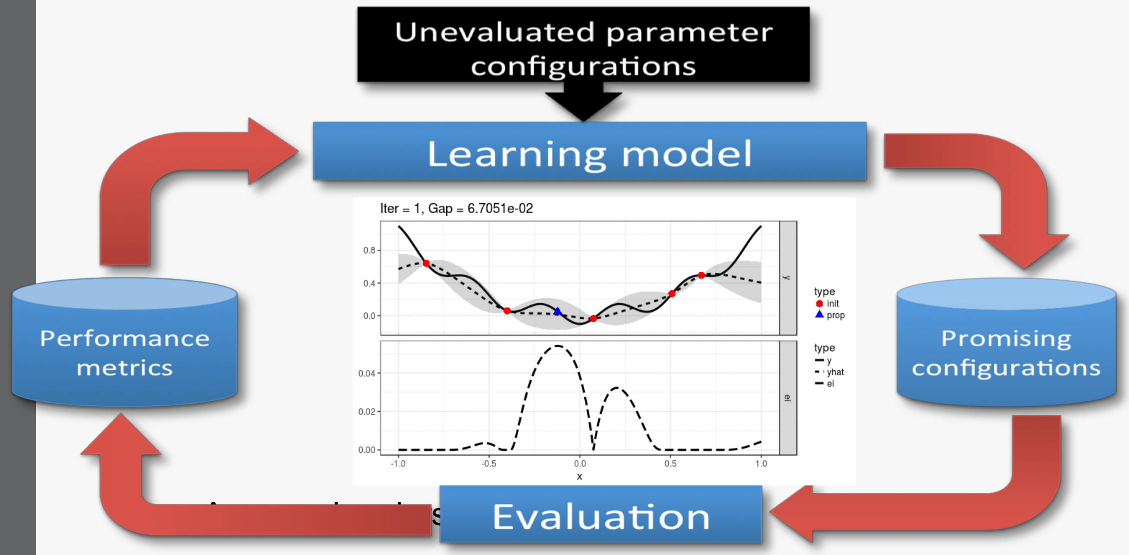
Balsam workflow manager

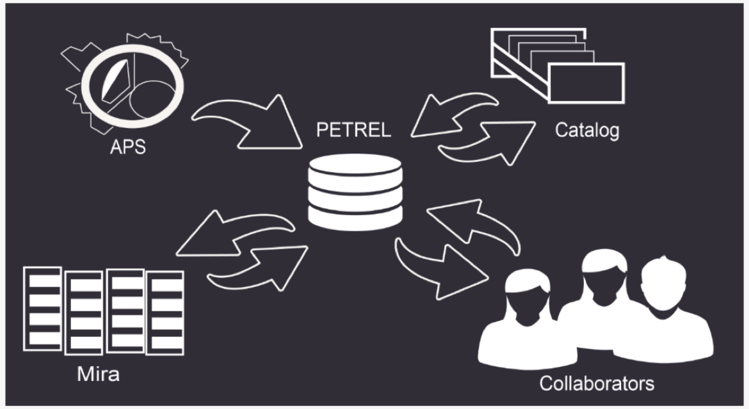
ATLAS experiment used Balsam to run ~100’s million compute hours of jobs across ALCF and NERSC systems. Balsam is used by ADSP, ESP, ALCC and ECP applications.
https://www.alcf.anl.gov/balsam



Deep Hyper

ALCF is currently conducting hyperparameter optimization for DL on thousands of nodes of Theta
https://github.com/deephyper/deephyper





PETREL
Data Management and Sharing Pilot

Petrel

Petrel leverages storage and infrastructure provided by ALCF and Globus Transfer and Sharing services. 100TB allocation per project.

http://petrel.alcf.anl.gov

Keeping up with the pace of Machine Learning is challenging

The pace of Machine Learning is very different to traditional HPC.

ML/DL software

- Updates occur every few weeks (TensorFlow, Keras, PyTorch, Horovod, etc.)
- Stack must enable performance libraries (Intel MKL, MKL-DNN, LibXSMM)
- Must work seamlessly with simulation and data frameworks
- Development does not always prioritize backwards compatibility

Keeping up requires

- Dedicated team members to track and update software regularly
- Containers which can provide portable customized software stack
- Regular training/workshops to update the scientific community

Thank you !

datascience@alcf.anl.gov