PSAAP III Research Topic:
Machine Learning and Scientific Applications

K. Lewis (LLNL) and M. Schraad (LANL)

March 2018
Summary
Machine learning techniques are being integrated into a vast array of applications across domains. The ASC program has introduced a coordinated effort to integrate these techniques into the scientific computing ecosystem. This program can significantly increase efficiency, improve models to better match experimental data, and improve the integration of multi-scale and multi-dimensional models. To better realize this potential, we must explore many different types of applications to grow our understanding of where different machine learning techniques will be most successful in the ASC ecosystem, from in-situ analysis (scalars to images) to building surrogate models from ensembles of high-resolution calculations. Additionally, validation concerns must be addressed. Three primary concerns for validation are (1) how can we quantify confidence in these techniques, (2) can we use the results of machine learning to better understand physics, and (3) how can we accurately train models in an environment that is heavily dominated by simulation data rather than experimental data.

Applications
Applications that have been identified as potential targets for machine learning can be divided into three major categories of benefits: increased efficiency, improved matching to experimental data, and improved integration between multi-scale and multi-dimensional models.

Increased Efficiency
We often refer to the potential increase in efficiency as a “resource multiplier”, where our resources include our workforce, our simulations, and our compute resources. Examples of improvements to the efficiency of our workforce include automatic identification of “important” features during a simulation and automatic steering of simulation controls such as Arbitrary Lagrangian Eulerian (ALE) and Automated Mesh Refinement (AMR) settings. These are situations where our workforce develops intuition over many years or decades of apprenticeship and experience. An example of increasing the efficiency of our simulations includes building surrogate models from ensembles of high-fidelity models to allow quick, lower fidelity exploration of simulation parameters. This allows more expensive, high-fidelity simulations to focus on parameters that are pre-adjusted for the desired outcome, much like simulations are used to focus experiments. This methodology can also allow tabular data to be replaced by models that more closely represent the physics, without significantly increasing computational costs. Examples of increasing the efficiency of compute resources include identifying when to move work to accelerators/co-processors or invoke dynamic load balancing. By allocating work efficiently, we can effectively multiply our compute resources to increase the number of calculations that can be performed. These types of applications are expected to be amenable to machine learning, which is more experience based than rule based.

Improved Matching to Experimental Data
A primary concept in machine learning techniques is that the model is generated through data and adapts with the introduction of new data. This fits into scientific computing nicely (with the exceptions that are highlighted under the validation section), because it allows coefficients to change in different regimes. We see a promising application of machine learning for physics models that evolve. Instead of accommodating step functions through phase changes, coefficients can be calculated more accurately by using specific data points in those regimes. Examples of this include turbulence modeling and equations of state models.

**Improved Integration Between Multi-Scale and Multi-Dimensional Models**

ASC modeling can vary widely in scale (both spatially and temporally) and in dimension. Models at the sub-atomic scale can interact with models at the macro-scale. Additionally, modeling at lower dimensions can be valuable for some calculations, and untenable for others. The integration of different spatio-temporal scales and different dimensions can be complex. Similar to creating surrogate models from high-fidelity simulations to explore parameter space, we also see the use of surrogate models to integrate models of varying scales and dimensions. The coupling of the continuum model to finer scale models can then be reduced to a function call, simplifying the integration. The surrogate model must account for the complexity of time dependent effects, which is addressed by Long Short-Term Memory (LSTM) models in machine learning.

**Validation**

Although we see much promise with machine learning in the ASC program, we must also address concerns with validation of these techniques and understand when new errors are introduced. This is not unique to ASC and is recognized as a major area to be addressed for the prevalence of machine learning. Three primary concerns are (1) how can we quantify confidence in these techniques, (2) can we use the results of machine learning to better understand physics, and (3) how can we accurately train models in an environment that is heavily dominated by simulation data rather than experimental data.

**Quantifying Confidence**

We would like to define a rigorous mathematical model to quantify the confidence in the result of the machine learning results. We envision a model that accounts for (a) the quality of the fit of the model to the data, (b) the amount of data that is close to the query, and (c) the variation of the data that is close to the query. Furthermore, it is necessary to have a model that suggests new data that could improve the confidence of the model. This provides an opportunity for additional guidance to experimental efforts, thus enabling more predictive capabilities by targeting areas of higher uncertainty. Without this capability, the use of machine learning is limited to applications that do not affect the results of a calculation, such as load balancing or transferring computing from the CPU to the GPU.

**Better Understanding of Physics**
Machine learning relies on correlations to predict the results of input parameters. More traditional, feature-based machine learning gives more transparency into how specific features correlate to outcomes. Deep neural networks have been shown to be more powerful and adapt better to new data, yet they have become more opaque with regard to the features that affect the results. Although we want models that correlate well to data and adapt well to new data, we also want a better understanding of causation to improve predictive capabilities in scientific computing. We would like to develop techniques for analyzing machine learning models that allow us to understand how the data affects the results. This can include the assessment of machine learning techniques alongside standard Bayesian statistical analysis, which are less of a “black box”.

**Data Poor Environments**

Much of our data is simulation data instead of experimental data. Although it is useful to create machine learning models using simulations, these models will inherit all of the approximations that were included in the simulation. Additionally, simulations with reduced dimensionality, which are common, introduce another layer of approximations that need to be understood when interpreting the results of machine learning models. We would like to identify ways for machine learning models to be built with a variety of data (experimental/simulation, 2D/3D) while recognizing that some data is superior to others. Simple weighting schemes may prove to be useful for this task, yet more complicated models that take into account the differences between the data may be necessary.

**Research Opportunities**

Advances in these areas require both research in Physics, Machine Learning, and Statistics and reduction to practice for Machine Learning in the area of Physics. Although we are encouraged by the successes of standard Machine Learning tools for current applications, we also foresee that we will run into challenges with standard tools at the scale and performance requirements that we require for simulation codes. Addressing these challenges will likely require research in Machine Learning and potentially in Statistics. Additionally, the validation section encompasses open research issues, to which experts commonly refer as “Explainable AI”. Research in these areas is attractive to newer researchers and opens up many possibilities for publications and collaborations across disciplines.