Learning Systems for Deep Science

A white paper for the November 2018 Big Data and Extreme-scale Computing workshop

Ian Foster, Argonne National Laboratory and University of Chicago; foster@anl.gov

We are entering an era of what we may call *deep science*, in which machine learning (ML) and in particular deep learning (DL) methods are used increasingly to automate many elements of research workflows.

Motivations from molecular sciences: Materials science and chemistry illustrate many relevant issues. ML/DL methods are being used extensively [6], for example to extract knowledge from the scientific literature [15], process data from experiments [13], synthesize software to study specific problems [5], estimate properties of unfamiliar compounds [16], select the next compounds and materials to study [11], and design experiments and computations [14]. New approaches are being used to capture and organize large quantities of heterogeneous data [4] and associated models [8].

New challenges for computational infrastructure: New methods such as those just reviewed present major challenges for the computational technologies, methods, and infrastructure on which science has long relied. The following are just a few of the issues. High-end computing is no longer restricted to a few niche researchers and their esoteric applications, and big data processing is no longer the exclusive domain of big data specialists. Scientists, engineers, and technicians need massive computing to train ML/DL models and (in the aggregate) to serve such models. They need access to large quantities of training data, which must be collected at many locations and integrated and organized for effective use. They require access to specialized software for defining, training, applying, and interpreting models. The software lifecycle changes also, as for example when the "applications" used by scientists are models created by automated processes, deployed in various forms on different platforms (e.g., at edge devices in field experiments and laboratories [2]), and updated dynamically in response to new data. These new scenarios pose challenges for provenance and reproducibility.

The need for learning systems: Addressing these new demands will require significant evolution of scientific infrastructure at every level, from processors (e.g., new DL-optimized chips), computers, data systems, systems software, libraries [10], and networks to the design of scientific facilities (e.g., to deliver data and to support automated operations). In some cases (e.g., processor design), innovations will come primarily from industry; in others, science will need to innovate to address unique requirements. The end result will likely be new *learning systems* that integrate large-scale computing, storage, and networks into research environments in ways designed to satisfy voracious new demands for both large-scale and timely data and computation; deliver new methods to new communities via new services and APIs, for example for on-demand inference; support the resulting new workloads, for example via the use of serverless computing [12]; distribute and connect data and computation in new ways; and track and explicate increasingly complex computational results.

The vital role of cloud services. The value of cloud services as a convenient source of elastic computing and storage is well known. Less appreciated, but equally important, is their ability to host powerful automation services that can manage the complex workflows that underpin modern data-driven science [9]. The Globus service [7] illustrates the latter use: its cloud-hosted services are used by tens of thousands to manage a wide range of processes relating to authentication and authorization, data transfer and synchronization, and data publication, and are now being extended to manage data lifecycle issues such as data publication [1] and processing of data from experimental facilities [3].

References

[1] R. Ananthakrishnan, B. Blaiszik, K. Chard, R. Chard, B. McCollam, J. Pruyne, S. Rosen, S. Tuecke, and I. Foster. Globus platform services for data publication. In *Practice and Experience on Advanced Research Computing*, page 14. ACM, 2018.

[2] P. Beckman, R. Sankaran, C. Catlett, N. Ferrier, R. Jacob, and M. Papka. Waggle: An open sensor platform for edge computing. In *IEEE SENSORS*, pages 1–3. IEEE, 2016.

[3] B. Blaiszik, K. Chard, R. Chard, I. Foster, and L. Ward. Data automation at light sources. In *13th International Conference on Synchrotron Radiation Instrumentation*. 2018.

[4] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, and I. Foster. The Materials Data Facility: Data services to advance materials science research. *Journal of Materials*, 68(8):2045–2052, 2016.

[5] V. Botu, R. Batra, J. Chapman, and R. Ramprasad. Machine learning force fields: Construction, validation, and outlook. *arXiv.org*, Oct. 2016.

[6] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, July 2018.

[7] K. Chard, S. Tuecke, and I. Foster. Globus: Recent enhancements and future plans. In *XSEDE16 Conference on Diversity, Big Data, and Science at Scale*, 2016.

[8] R. Chard, Z. Li, K. Chard, L. Ward, Y. Babuj, A. Woodard, S. Tuecke, B. Blaiszik, M. J. Franklin, and I. Foster. DLHub: Model and data serving for science. 2018.

[9] I. Foster and D. B. Gannon. *Cloud Computing for Science and Engineering*. MIT Press, 2017.

[10] A. Haidar, A. Abdelfattah, M. Zounon, P. Wu, S. Pranesh, S. Tomov, and J. Dongarra. The design of fast and energy-efficient linear solvers: On the potential of half-precision arithmetic and iterative refinement techniques. In *International Conference on Computational Science*, pages 586–600, 2018.

[11] W. Jin, R. Barzilay, and T. Jaakkola. Junction tree variational autoencoder for molecular graph generation. *arXiv.org*, Feb. 2018.

[12] H. Lee, K. Satyam, and G. Fox. Evaluation of production serverless computing environments. In *11th International Conference on Cloud Computing*, pages 442–450. IEEE, 2018.

[13] D. Pelt, K. Batenburg, and J. Sethian. Improving tomographic reconstruction from limited data using mixed-scale dense convolutional neural networks. *Journal of Imaging*, 4(11):128–20, Nov. 2018.

[14] F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers, and A. Mehta. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science Advances*, 4(4):eaaq1566, 2018.

[15] M. C. Swain and J. M. Cole. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904, 2016.

[16] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2:16028, 2016.