Data Automation at Light Sources: Experiments and Lessons Learned

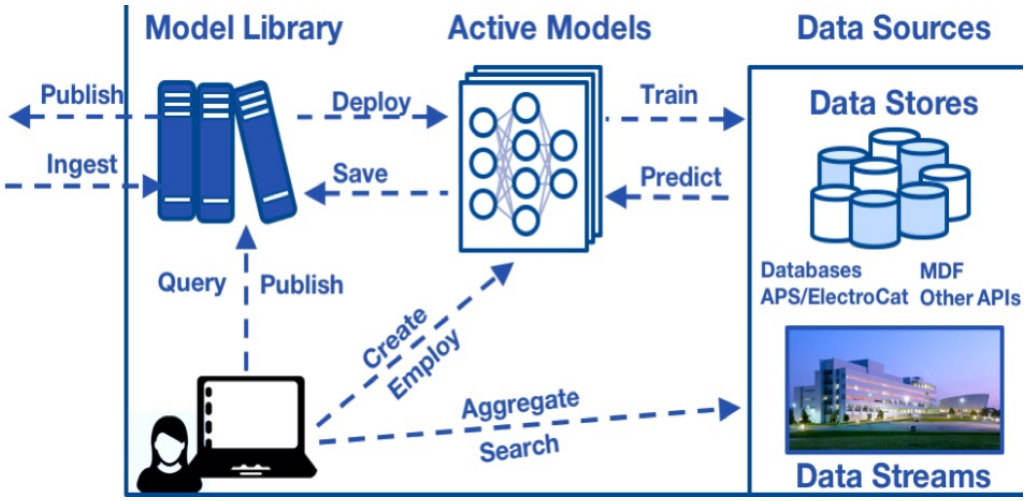
Ian Foster[1,2]

[1]*Argonne National Laboratory, USA*

[2]*University of Chicago, USA*

foster@anl.gov

Rapidly growing data volumes at light sources demand increasingly automated data collection, distribution, and analysis processes, in order to enable new scientific discoveries while not overwhelming finite human capabilities. I present here three projects that use cloud-hosted data automation and enrichment services, institutional computing resources, and high-performance computing facilities to provide cost-effective, scalable, and reliable implementations of such processes. In the first, Globus cloud-hosted data automation services [1] are used to implement data capture, distribution, and analysis workflows for Advanced Photon Source and Advanced Light Source beamlines, leveraging institutional storage and computing [3]. In the second, such services are combined with cloud-hosted data indexing and institutional storage to create a collaborative data publication, indexing, and discovery service, the Materials Data Facility (MDF) [2], built to support a host of informatics applications in materials science. The third integrates components of the previous two projects with machine learning capabilities provided by the Deep Learning Hub (DLHub) to enable on-demand access to machine learning models from light source data capture and analysis workflows, and provides simplified interfaces to train new models on data from sources such as MDF on leadership scale computing resources. I review lessons learned in these projects and draw conclusions regarding best practices for light source data automation systems.



The Deep Learning Hub (DLHub) being developed at Argonne National Laboratory provides on-demand access to best-of-breed machine learning models for analysis of data from many sources, including light source beamlines.

References

[1] R. Ananthakrishnan et al. "Globus platform-as-a-service for collaborative science applications." *Concurrency and Computation: Practice and Experience* 27, no. 2 (2015): 290-305.

[2] B. Blaiszik et al. "The Materials Data Facility: Data services to advance materials science research." *JOM* 68, no. 8 (2016): 2045-2052.

[3] I. Foster et al. "Networking materials data: Accelerating discovery at an experimental facility." *Big Data and High Performance Computing* 26 (2015): 117-132.