Automated Laboratories Coupled with AI Agents and Smart Data Stores

Ben Blaiszik and Logan Ward - Argonne National Laboratory, University of Chicago

Summary: We envision (1) distributed and automated AI-driven laboratories where experiments are chosen to maximize collection of the most new information; (2) shared AI capabilities that may request experiments from available computational resources or the automated labs; and (3) smart data stores to serve as an evolving shared knowledge base accessible to researchers and other AIs.

Automation, software, and data services have become an increasingly important driver of scientific progress, and the labs of the future will leverage this progress and accelerate these trends dramatically. The common roles of automation at the moment are limited. In labs, simple tasks, e.g., maintaining sample positions and state (e.g., temperature) are automated and tracked by data collection routines, and output data are often organized by humans in ad hoc file formats and spreadsheets from heterogeneous output file types and shapes. We anticipate the role of automation, shared data services, and digital curation of data to become even more commonplace due to the revolution in artificial intelligence. AI lets humans write software more complicated than possible only a decade ago and the desire of scientists to use this capability, as we will discuss, will drive fundamental changes in laboratories.

Automated Distributed Experimentation, Coupled with AI: Future laboratories will be filled with smart devices; i.e., devices that can autonomously perform key experiments, be controlled by software, and communicate bidirectionally with a host of data services. Where current laboratory software follows detailed and prescribed scripts, the future laboratory software will plan and perform complicated studies with broad directives such as “Optimize the microstructure of a FeCrAl alloy for high temperature application”. An entire laboratory full of equipment could be driven by an AI agent that chooses how and when to perform certain processing, synthesis, and characterization steps. Each of those steps could be performed on smart equipment that themselves are run by dedicated AI agents. In order to achieve this, laboratories themselves will need to largely be made accessible by API.

Smart Data Stores: Smart data resources will also be needed to support these laboratories and associated AIs. Such data stores will serve foremost as a shared and queryable repository of training data for the agents. As more data are acquired, the latest data will be indexed in appropriate ways and made accessible via API. These data repositories will also have their own embedded intelligence: (1) to identify anomalies within data, (2) to request new information for areas devoid of data, (3) to automatically extract metadata from heterogeneous data collections, (4) to synthesize and arrange these data in a way that other AIs and humans can access the data more intuitively, (5) hold theories and monitor whether they still hold as new data are acquired, (6) and communicate these findings back to bench scientists for feedback. The data store will serve as a source of knowledge and a font of ideas for potential new scientific ideas to be explored.

Central Repository of AI Agents: The nature of AI tools is constantly changing, which will require software infrastructure to keep laboratories stocked with the state-of-the-art AI agents. We envision a centralized repository of AI software (i.e., trained models and their accompanying applications) categorized by their intended application and with known application bounds. As current models are retrained based on new data in smart data stores or new techniques for certain applications are invented, scientists can easily update their laboratory AI agents to access the latest technology. As these models are updated as the smart data store contents grow, laboratories will grow more intelligent over time with little effort from bench scientists.