Scientific Data Security and Integrity

T. Munson
Mathematics and Computer Science Division, Argonne National Laboratory
tmunson@mcs.anl.gov

March 2020

In 2050, useful quantum computing will remain 30 years away and we will have harvested the low-hanging fruits of the scientific AI revolution and will have defined the class of AI complete problems, whose complement includes complex problems that either cannot be tackled by AI or are better solved using traditional methods. While not entering another AI winter, AI research will instead be focused on improving the speed of solving AI complete problems, establishing links between various AI complete problems, and pushing the theoretical concept of quantum AI complete problems as a means of moving past the classical AI barriers, while awaiting the dawn of useful quantum computers.

In this world, AI will have relieved scientists of mundane tasks, freeing them to think about problems at a high level. Grant writing and panel reviews will become a thing of the past, replaced by AI assisted grant writers and evaluators. Programming languages will evolve to express the concepts required to push scientific boundaries, while automation and AI methods will be used to translate high-level descriptions into code that runs efficiently on the available machines. However, legacy Fortran code will linger — although inroads will be made with expert AI assistants documenting their internals — and C++ will become even more complicated and template-based codes will still take a long time to compile, even when assisted by AI acceleration. Computing resources will be readily available, with cloud-based computing replaced by on demand supercomputing. Despite these computing resources, scientific data will still be expensive to generate and in short supply. Scientists will become increasingly specialized and technological solutions, such as a google translate assistant between different scientific domains, will help drive integrated projects across domains.

These advancements will fundamentally change the way that science is accomplished, where data will be the key commodity and scientists will be required for their intuition and to turn data into knowledge. AI methods will be able to produce data that looks to be scientifically meaningful, so the ability to detect and reject AI generated data will become essential to prevent contamination of data archives with fake data. Rigorous security of scientific data archives will be required to protect their integrity, along with the associated provenance and curation of scientific data. More time and effort will be allocated to validation and verification of AI generated models in the specialized scientific domains, and reproducibility will become a standard requirement for publishing results. Scientific advances will rely on combining domain-specific knowledge in new and interesting ways to uncover flaws in existing methodologies and suggest experiments to fill in the gaps. Validation and verification activities will become more complex and expand beyond the domain-specific models to consider simultaneous validation of the entire workflow.

While different from today, scientists in 2050 will have many opportunities to advance theory, computational science, applied mathematics and computer science. This confluence of activities and elimination of mundane tasks, will allow scientists to expend more of their efforts on understanding scientific phenomena.