CompBioMed Conference

September 25-27 2019

Title:

AI for Big Science

Speaker:

Professor Tony Hey

Chief Data Scientist

Rutherford Appleton Laboratory

Science and Technology Facilities Council

Didcot OX11 0QX

Abstract

This talk will review some of the challenges posed by the huge growth of experimental data generated by the new generation of large-scale experiments at UK national facilities at the Rutherford Appleton Laboratory site at Harwell near Oxford. Such 'Big Scientific Data' comes from the Diamond Light Source and Electron Microscopy Facilities, the ISIS Neutron and Muon Facility, and the UK's Central Laser Facility. Increasingly, scientists are now needing to use advanced machine learning and other AI technologies both to automate parts of the data pipeline and also to help find new scientific discoveries in the analysis of their data. For commercially important applications, such as object recognition, natural language processing and automatic translation, Deep Learning has made dramatic breakthroughs. Google's DeepMind has now also used Deep Learning technology to develop their AlphaFold tool to make predictions for protein folding. Remarkably, they have been able to achieve some spectacular results for this specific scientific problem. Could Deep Learning be similarly transformative for other scientific problems? After a brief review of some initial applications of machine learning at the Rutherford Appleton Laboratory, we focus on challenges and opportunities for AI in advancing pharmaceutical and materials

science. Finally, we discuss the importance of developing some realistic machine learning benchmarks using Big Scientific Data coming from a number of different scientific domains. For the computer vision community, it was the ImageNet database that provided researchers with the capability to evaluate algorithms for object detection and image classification at large scale. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) allowed researchers to compare progress in detection across a wider variety of objects and led directly to the present Deep Learning and GPU revolution. We believe that the creation of a credible 'Scientific Machine Learning' (SciML) collection of benchmarks could prove useful and significant for the scientific research community. The talk concludes with some initial examples of our 'SciML' benchmark suite and a discussion of the research challenges these benchmarks will enable.