Artificial intelligence for science and engineering: A priority for public investment in research and development

T. Hey, UK Research and Innovation, United Kingdom

Introduction

The rapid growth of scientific data generated both by scientific experiments at large national and international facilities and by model simulations on supercomputers epitomises Jim Gray's "Fourth Paradigm" of data-intensive science. The use of artificial intelligence (AI) technologies to help automate the generation and analysis of such datasets is increasingly necessary. This essay describes the great potential for the use of AI and deep learning technologies to transform many fields of science. It draws particular attention to the conclusions of Town Hall meetings organised by the US Department of Energy (DOE). These meetings explored the potential for AI to accelerate science and the need for major public research and development (R&D) funding. Such funding could enable multidisciplinary teams of academic researchers to generate comparable breakthroughs to those of commercial companies such as Google DeepMind.

Deep learning (DL) neural networks – a sub-discipline of AI – came to prominence in 2012 when a team led by Geoffrey Hinton won the ImageNet Image Recognition Challenge (Krizhevsky et al., 2012). Their entry in the competition, AlexNet, was a DL network consisting of eight layers. The learning phase was computed on graphics processing units (GPUs), a specialised form of electronic circuit frequently used in software-based games. By 2015, building on this initial research, a Microsoft Research team used a DL network with more than 150 layers trained using clusters of GPUs to achieve object recognition error rates comparable to human rates (He et al., 2016).

DL networks are now a key technology for the IT industry and used for a wide variety of commercially important applications. These include facial recognition, handwriting transcription, machine translation, speech recognition, autonomous driving and targeted advertising. More recently, Google's UK subsidiary, DeepMind, used DL neural networks to develop the world's best Go playing systems with their AlphaGo variants.

Of particular interest for science is DeepMind's AlphaFold protein-folding prediction system (Senior et al., 2020). Their latest version of AlphaFold convincingly won the most recent Critical Assessment of Protein Structure Prediction (Jumper et al., 2021). Nobel Prize winner Venki Ramakrishnan said, "This computational work represents a stunning advance on the protein folding problem, a 50-year-old grand challenge in biology. It has occurred decades before many people in the field would have predicted. It will be exciting to see the many ways in which it will fundamentally change biological research" (DeepMind, 30 November 2020).

This essay reviews the changing face of much data-driven scientific research, and the impact of DL and other AI technologies.

The four paradigms of scientific discovery

Turing Award winner Jim Gray was the first to use the term "Fourth Paradigm" to describe the next phase of data-intensive scientific discovery (Gray, 2009). In the first paradigm, which lasted over 1 000 years, science was empirical, based solely on observation. Then, in 1687, after the discoveries of Kepler and Galileo, Isaac Newton published the *Mathematical Principles of Natural Philosophy*. This established his three laws of motion that defined classical mechanics and provided the foundation for his theory of gravity. The mathematical laws of nature provided the basis for theoretical explorations of scientific phenomena, a second paradigm for scientific discovery. Nearly 200 years later, Maxwell formulated equations for his unified theory of electromagnetism, and then, in the early 20th century, Schrödinger's equation described quantum mechanics. The use of these two paradigms – experimental observation and theoretical calculation – has been the basis for scientific understanding and discovery for the last few centuries.

In 2007, working on a study of computing futures, Jim Gray and the Computer Science and Telecommunications Board realised that computational science was a third paradigm for scientific exploration. It involved a shift towards simulation based on, and generating large volumes of, scientific data created in the first instance by digital instruments.

In his talk to the Board, Gray (2009) concluded the world of science had changed:

> *The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm for scientific exploration (Gray, 2009).*

Each paradigm has limitations. Experiments can be slow and difficult to do at scale or even at all. Moreover, large-scale instruments, such as the Large Hadron Collider (LHC) or the Square Kilometre Array radio telescope, are expensive to build and maintain. In addition, the output of each experiment is usually analysed separately, within its own silo. This limits the potential for new knowledge to analysis of the output and input parameters of individual experiments.[1]

Mathematical models can also have limitations, including the need to simplify assumptions to create the models in the first place. In addition, scientists are often unable to solve the resulting set of complex equations to produce easily explored analytical solutions. Computer simulations can be used to address both of these limitations to a certain extent. Simulating mathematical models for a wide range of different research areas – such as climate science, molecular dynamics, materials science and astrophysics – has proved successful, and supercomputers are now used routinely for such simulations. Exascale supercomputers can perform 10^{18} floating point calculations per second. However, even supercomputer simulations are limited by the mathematical models and data representations being simulated. Moreover, one simulation represents only one instance of the problem, based on a particular set of initial conditions and constraints. In addition, some simulations can take many days or weeks to complete, thus limiting the exploration of large parameter spaces.

Simulation of climate models, which needs to be urgently improved, illustrates such limitations. The US National Center for Atmospheric Research (NCAR) collaborates in a new National Science Foundation (NSF) multidisciplinary Center for Learning the Earth with Artificial Intelligence and Physics (LEAP). LEAP will use machine learning (ML) technologies to improve NCAR's Community Earth Systems Model (CESM). The CESM comprises a complex collection of component models that can simulate the interaction

of atmosphere, ocean, land, sea ice and ice sheet processes. However, CESM is limited in its ability to incorporate an accurate mathematical representation of some important physical processes that are difficult to simulate. These include the formation and evolution of clouds at such a fine scale that the model cannot resolve them, and processes to represent land ecology that are too complicated to capture in a simulation. Climate scientists have created simplified subcomponents – known as parameterisations – to approximate these physical processes into CESM. As one of its major goals, LEAP aims to improve these approximations by using ML technologies to incorporate learning from large amounts of Earth system observational data and high-resolution model simulation data.

AI for science and engineering

AI for Science and Engineering applies AI and DL technologies to the huge scientific datasets generated by both supercomputer simulations and modern experimental facilities. Huge quantities of experimental data now come from many sources – from satellites, gene sequencers, powerful telescopes, X-ray synchrotrons, neutron sources and electron microscopes. They are also generated from major international facilities such as the LHC at the European Organization for Nuclear Research (CERN) in Geneva and the European X-ray Free-Electron Laser facility in Hamburg. These facilities already generate many petabytes of data per year and their planned upgrades will create at least an order of magnitude more data. Extracting meaningful scientific insights from these ever-increasing volumes of data will be a major challenge for scientists.

Many initiatives around the globe are now applying AI technologies to manage and analyse the ever-larger and more complex scientific datasets. Commercial tools and technologies for ML provide scientists with a good starting point. However, their application to the wide range of scientific problems requires multidisciplinary collaborative teams, including both computer scientists and physical scientists. In the United States, for example, the National Science Foundation recently established 18 National AI Research Institutes with research partnerships covering 40 states (NSF, 2022). The US DOE funds both the associated large-scale experimental facilities and the supercomputers at the National Laboratories. In 2019, the DOE Laboratories organised a series of Town Hall meetings to examine opportunities and practical next steps for AI to accelerate research in fields under the domain of the DOE's Office of Science (DOE, 2020). These meetings were attended by hundreds of scientists, computer scientists, along with participants from industry, academia and government.

The DOE Town Hall meetings used the term "AI for Science" to broadly represent the next generation of methods and scientific opportunities in computing and data analysis. This included the development and application of AI methods – a combination of ML, DL, statistical methods, data analytics and automated control – to build models from data and to use these models alone or with simulation data to advance scientific research. In line with ideas expressed in many of the contributions to the current publication, the meetings concluded that AI could transform many areas of scientific research over the next decade. It envisioned that AI technologies can:

- accelerate the design, discovery and evaluation of new materials
- advance development of new hardware and software systems, instruments and simulation data streams
- identify new science and theories revealed in high-bandwidth instrument data streams
- improve experiments by inserting inference capabilities in control and analysis loops
- enable the design, evaluation, autonomous operation and optimisation of complex systems from light sources and accelerators to instrumented detectors and high-performance computing (HPC) data centres
- advance development of autonomous laboratories and scientific workflows

- dramatically increase the capabilities of exascale and future supercomputers by capitalising on AI surrogate models (i.e. models that that mimic the behaviour of the simulation models as closely as possible while being computationally much cheaper to evaluate)
- automate the large-scale creation of findable, accessible, interoperable and re-usable (FAIR) data.

The Alan Turing Institute, the national institute for data science and AI in the United Kingdom, reached a similar conclusion. Its "AI for Science and Government" initiative includes a major research effort on AI for science in collaboration with the Scientific Machine Learning Group at the Rutherford Appleton Laboratory, the UK's National Laboratory at Harwell, near Oxford (STFC, 2022).

Thoughts on directions for research (and research policy)

Google DeepMind has applied DL techniques to make significant progress in three different fields of science – protein folding, materials modelling (Kirkpatrick et al., 2021) and fusion plasma control (Degrave et al., 2022). Researchers at DeepMind assembled multidisciplinary teams of experts and used the power of Google's Cloud computing resources for training their DL solutions to make these breakthroughs.

Can academic researchers compete with such efforts? Two actions are needed to address this question:

- A broad multidisciplinary programme is needed to allow scientists, engineers and industry to collaborate with computer scientists, applied mathematicians and statisticians to solve their challenges using a range of AI and ML technologies. This needs coherent and dedicated government funding with processes that encourage such collaboration rather than continuing with stove-piped funding allocated to individual disciplines.
- Such a programme should create a shared cloud infrastructure that allows researchers to access the competitive computing resources and tools that fuel AI R&D. In the United States, the NSF and the White House Office of Science and Technology Policy are creating a roadmap to establish a National AI Research Resource (NSF and OSTP, 2022). This is intended to be a shared research infrastructure that will provide AI researchers with significantly expanded access to computational resources and high-quality data.

The DOE work described here also detailed a rich set of topics on which research breakthroughs are needed to broaden and deepen AI's uses in science and engineering. These topics could become targets of public R&D support. In particular, as DOE (2020) describes, participants highlighted the need to:

- Incorporate domain knowledge into AI methods to improve the quality and interpretability of the models. There is a need to go beyond current models driven only by data or simple algorithms, laws and constraints. Especially key would be ML techniques driven by theory and data that could better represent the underlying dynamics specific to particular phenomena.
- Automate the large-scale creation of FAIR data. AI in science requires large datasets from a diverse range of sources – from experimental facilities and computational models to environmental sensors and satellite data streams. Adding some semantic information in the form of machine-actionable metadata could allow AI technologies to automate the creation of FAIR data. This would provide the basis for new data infrastructures to allow more interoperability and re-use.
- Advance foundational topics in the science of AI itself, with a view to developing:
  - frameworks to establish that a given problem is solvable by AI/ML methods
  - frameworks and tools to establish the validity and robustness of AI techniques, indicating the limits of AI techniques, the quantification of uncertainties and the conditions (assumptions and circumstances) that give assurance of AI predictions and decisions
  - frameworks and tools to establish which AI techniques best address different sampling scenarios and enable efficient AI on different computing and sensing environments

  - techniques that help explain the behaviour of the AI model methods
  - AI models that identify causal variables and distinguish between cause and effect
  - methods for AI models to be used to identify causal variables and distinguish between cause and effect.
- Develop new hardware and software environments. Much new AI hardware is being developed in industry for data centres, autonomous driving systems and gaming, among others. Opportunities exist for the research community to work with industry to co-design heterogeneous compute systems that use the new architectures and tools. Software is also needed to enable AI capabilities to seamlessly integrate with large-scale HPC models (see the essay in this volume by Georgia Tourassi, Mallikarjun Shankar and Feiyi Wang) and to generate and operate new scientific workflows. Such AI-enabled workflows will incorporate expert knowledge to accomplish tasks, adapt to new data and results, and refine models on the basis of cost (e.g. in energy use or run-time).

Conclusion

Greatly increased data volumes are expected for the next generation of scientific experiments. This is true for the National Laboratories in the United States with their large-scale experimental facilities, for projects such as the Square Kilometre Array Radio Telescope observatory (SKA, 2022), and for the CERN LHC upgrade, among many others. AI will be needed to automate the data collection pipelines and advance the analysis phase of such experiments. For all of these reasons, major multidisciplinary programmes on AI for science and engineering should be a high priority for public R&D investment. They can greatly increase the rate of scientific discovery and catalyse new commercial developments.

References

DeepMind (30 November 2020), “AlphaFold: A solution to a 50-year-old grand challenge in biology”, DeepMind blog, www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

Degrave, J. et al. (2022), “Magnetic control of tokamak plasmas through deep reinforcement learning”, *Nature*, Vol. 602, pp. 414-419, https://doi.org/10.1038/s41586-021-04301-9.

DOE (2020), *AI for Science, Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science*, US Department of Energy, Office of Science, Argonne National Laboratory, Lemont, https://publications.anl.gov/anlpubs/2020/03/158802.pdf.

Gray, J. (2009), “Presentation at the NRC-CSTB in Mountain View, CA, 11 January 2007”, in *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Hey, T., S. Tansley and K. Tolle (eds.), Microsoft Research, Redmond.

He, K. et al. (2016), “Deep residual learning for image recognition”, in *2016 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, https://doi.org/10.1109/CVPR.2016.90.

Jumper, J. et al. (2021), “Highly accurate protein structure prediction with AlphaFold”, *Nature*, Vol. 596, pp. 583-589, https://doi.org/10.1038/s41586-021-03819-2.

Kirkpatrick, J. et al. (2021), “Pushing the frontiers of density functionals by solving the fractional electron problem”, *Science*, Vol. 374, pp. 1385-1389, https://doi.org/10.1126/science.abj6511.

Krizhevsky, A. et al. (2012), “ImageNet classification with deep convolutional neural networks”, *Advances in Neural Information Processing Systems*, pp. 1097-1105, https://doi.org/10.1145/3065386.

NSF (2022), “NSF-Led National AI Research Institutes”, webpage, www.nsf.gov/news/ai/AI_map_interactive.pdf (accessed 25 November 2022).

NSF and OSTP (2022), National AI Research Resource Task Force website, www.ai.gov/nairrtf/ (accessed 25 November 2022).

Senior, A.W. et al. (2020), “Improved protein structure prediction using potentials from deep learning”, *Nature*, Vol. 577, pp. 706-710, https://doi.org/10.1038/s41586-019-1923-7.

SKA (2022), "The Ska Project”, webpage, www.skatelescope.org/the-ska-project/ (accessed 25 November 2022).

STFC (2022), “Scientific Machine Learning”, webpage, www.scd.stfc.ac.uk/Pages/Scientific-Machine-Learning.aspx (accessed 25 November 2022).

Note

[1] However, AI and deep learning technologies can be applied not just to a single instance of an experiment but also to the analysis of the combined total of information from many such experiments. This can help generate new scientific discoveries and insights.

From:

Artificial Intelligence in Science

Challenges, Opportunities and the Future of Research

Access the complete publication at:
https://doi.org/10.1787/a8d820bd-en

Please cite this chapter as:

Hey, Tony (2023), "Artificial intelligence for science and engineering: A priority for public investment in research and development", in OECD, *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/7b7b1bce-en