How Can AI Accelerate Science, and How Can Our Government Help?

Tom M. Mitchell
Carnegie Mellon University
July 2024

Recent dramatic advances in AI, including Large Language Models such as GPT, Claude, and Gemini raise the possibility that one very positive impact of AI might be to dramatically accelerate research progress across a wide variety of scientific fields, from cell biology, to materials science, to weather and climate modeling, to neuroscience. Here we briefly summarize this AI-Science opportunity, and what the U.S. government can do to seize this opportunity.

1. The AI-Science Opportunity

The vast majority of today's scientific research, across nearly every field, can be characterized as "lone ranger" science. In other words, scientists and their research teams of a dozen or so researchers, come up with an idea, conduct an experiment to test it, write up and publish the result, perhaps share their experimental data on the internet, and then repeat this process. Other scientists can build on these results by reading the published paper, but this process is error-prone and highly inefficient for several reasons: (1) individual scientists have no hope of reading all of the published articles in their field, and therefore operate in partial blindness to relevant other research, (2) the full details of the experiments described in the journal publications necessarily omit many details, making it difficult or impossible for others to replicate or build on their results, and (3) the analysis of a single experimental dataset is typically done in isolation, failing to incorporate data (and hence valuable information) from other relevant experiments conducted by other scientists.

Over the coming decade, AI can help scientists overcome all three of the above problems by shifting this "lone ranger" paradigm for scientific research toward a paradigm of "community scientific discovery." In particular, AI can be used to create a new kind of computerized research assistant that helps human scientists overcome these problems by

- Discovering regularities in complex data sets, including data sets built up from *many experiments conducted across many laboratories*, in contrast to "lone ranger" analyses of single, much smaller and less representative data sets. This can lead to much more comprehensive and accurate analyses, by basing the analysis on orders of magnitude larger data sets that are beyond the ability of humans to even examine.
- Using AI Large Language Models like GPT to read and digest every relevant publication in the field, and thereby help the scientist form new hypotheses based not only on experimental data from their and other labs, but also based on the hypotheses and arguments found in the published research literature, resulting in hypotheses that are much more informed than is possible without such natural language AI tools.

- Creating "foundational models" that capture the growing knowledge of the field in a single place, and that provide computer-executable models of this knowledge, by training these models on many diverse types of experimental data collected across laboratories and scientists. These executable "foundational models" can serve the same role that equations (e.g, f = ma) serve, in that they make predictions about some quantities based on other observed quantities. However, these foundational models, unlike typical equations, can capture the empirical relationships among hundreds of thousands of different variables, rather than a handful of variables.
- Automating or semi-automating the design and the robotic execution of new experiments, thereby accelerating the rate of new relevant experimentation, and improving the repeatability of scientific experiments.

What are the potential scientific breakthroughs that might result from this paradigm shift in the practice of science? Here are a few examples:

- Shrink by 10x the time and dollar cost of developing new vaccines for new disease outbreaks.
- Accelerate research in materials science, potentially leading to breakthrough products such as room temperature superconductors, better batteries, and thermoelectric materials that convert heat to electrical power without generating emissions.
- Combine sets of experimental data from cell biology at a volume and at a diversity that has never before been attempted, resulting in a "foundational model" of how human cells function, leading to the ability to simulate quickly the results of many potential experiments *in silico* before taking the more costly step of running the experiment *in vivo* in the lab.
- Combine sets of experimental data from neuroscience - from data on single neuron behavior to full-brain fMRI imaging - to build a "foundational model" of the human brain at multiple levels of detail, integrating data at a scale and diversity never before attempted, and resulting in a model that can predict neural activity the brain uses to encode different types of thoughts and emotions, how these can be evoked by different stimuli, the impact of medications on neural activity, and the effectiveness of different therapies for treating mental disorders.
- Improve our ability for forecasting weather, both to customize predictions to highly localized regions (e.g., a single farm), and to extend our ability to forecast weather much further into the future.

2. What Can the U.S. Government Do to Seize this Opportunity?

Transforming this opportunity into reality requires several components:

- *Significant experimental data.* One lesson from text-based foundational models is that the more data they are trained on, the more capable they become. Empirical scientists are also well aware of the value of more, and more diverse experimental data. To achieve multiple order-of-magnitude advances in science, and to train the types of foundational models we desire, will require a very significant advance in our ability to share and jointly analyze diverse data sets contributed across the entire scientific community.
- *Access to scientific publications and the ability to read them by computer.* A key part of the opportunity here is to shift from today's state where a scientist is unlikely to be able to read even 1% of the relevant publications in their field, to a state where a computer assists them by reading

100% of these publications, summarizing them and their relevance to the current scientific question, and providing a conversational interface to discuss their content and implications. This will require not only access to online literature, but also AI research on how to construct such a "literature assistant"

- *Computational and networking resources*. Text-based foundational models such as GPT and Gemini are famous for the huge processing resources spent in their development, and significant computational resources will be needed to develop foundational models in different scientific domains as well. However, the computational needs in many AI-Science efforts can be significantly smaller than the computation needed to train LLMs such as GPT, and therefore achievable with investments similar to those ongoing at government research labs. For example, AlphaFold, an AI model that has already revolutionized analysis of proteins for drug design, used much less training computation than text-based foundation models such as GPT and Gemini – the cost of computation to train AlphaFold is a few hundred thousands dollars, compared to a few hundred million dollars to train today's LLMs. Furthermore, once trained, running AlphaFold on a new protein costs less than a dollar. Beyond GPU costs, we will also need significant computer networking to support data sharing, but the current internet already provides an adequate starting point for transmitting large experimental data sets. Thus, the hardware costs for supporting AI-driven science advances may be quite modest in comparison with the potential benefits.
- *New Machine Learning and AI methods*. Current machine learning methods have been found to be extremely valuable for discovering statistical regularities in data sets too large for human inspection (e.g., AlphaFold was trained on a large set of protein sequences and their painstakingly measured 3D structures). A key part of the new opportunity is to extend current machine learning methods, which find *statistical correlations* in data, in two important directions: (1) to move from discovering *correlations* to discovering *causal relationships* in data, and (2) to move from learning from only large structured data sets, to learning from large structured data sets *plus* the vast research literature; that is, to learn as human scientists do from both experimental data and the published hypotheses and arguments of others expressed in natural language. The recent advent of LLMs with advanced capabilities to digest, summarize, and reason about large text collections can form the basis for new machine learning algorithms of this kind.

What should our government do? The key is to support each of the four components noted above, and to rally multiple scientific communities to explore novel AI-based approaches to accelerate their research progress. Accordingly, the government should consider several types of actions:

- *To explore specific opportunities in specific scientific fields*, fund multi-institutional research teams in each of many scientific fields, to produce a vision and preliminary results showing how AI might be used to dramatically accelerate progress in their field, and what is needed to scale the approach. This effort should NOT be funded in grants to individual institutions, because the biggest advances are likely to come from integrating data and studies across many scientists at many institutions. Instead this is likely to be most effective if performed by teams of scientists across many institutions, proposing opportunities and approaches that carry with them the incentives to engage their full scientific community.

- *To accelerate creation of new experimental datasets to train new foundation models, and to make data available to the full community of scientists*:
    - Create data sharing standards to make it easy for one scientist to (re)use the experimental data created by a different scientist, and to form the basis for a national data resource in each relevant science. Note there are earlier successes in setting and using such standards, that can provide starting templates for standards efforts (e.g., the success in sharing data in the human genome project).
    - Create and support data sharing websites for each relevant field. Just as GitHub has become the go-to website for software developers to contribute, share and reuse software code, create a GitHub for scientific data sets that serves as both data repository and search engine for discovering data sets most relevant to a particular topic, hypothesis, or planned experiment.
    - Conduct a study of how to construct incentives to maximize data sharing. Currently, scientific fields vary widely in the degree to which individual scientists share data, and the degree to which for-profit institutions make their data available for basic scientific research. Building a large, sharable national data resource is such an integral component of the AI-science opportunity, that constructing a compelling incentive structure for data sharing will be key to success.
    - Where appropriate, fund development of automated laboratories (e.g., robotic labs for experiments in chemistry, biology, etc., accessible to a wide collection of scientists over the internet) to efficiently run experiments, and to produce data in a standard format. One major side-benefit of creating such laboratories is that they will also drive the development of standards for stating precisely the experimental procedure to be followed, thereby improving reproducibility of experimental results. Just as we can benefit from a GitHub for data sets, we can also benefit from a related GitHub for sharing, modifying and reusing components of experimental protocols.
- *To create the new generation of AI tools needed:*
    - Fund relevant basic AI research specifically targeted to develop approaches applicable to scientific research. This should include developing "foundation models" interpreted broadly, as tools to accelerate research in different fields, and to accelerate the paradigm shift from "lone ranger" science to a more powerful "community scientific discovery" paradigm.
    - Support in particular research on reading the research literature to critique and suggest refinements to stated input hypotheses, and generally to assist scientists in accessing the results from the scientific literature in a way that directly relates to their current problem.
    - Support in particular research on extending ML from discovering *correlations* to discovering *causality*, especially in settings where new experiments can be planned and executed to test hypotheses about causality.
    - Support in particular research on extending ML algorithms from taking only *big data* as input, to taking as input both *big experimental data, and the full research literature in the field*, in order to produce output analyses that are jointly informed by the statistical regularities in the experimental data, and by the stated hypotheses, explanations, and arguments discussed in the research literature.