HYPERION RESEARCH

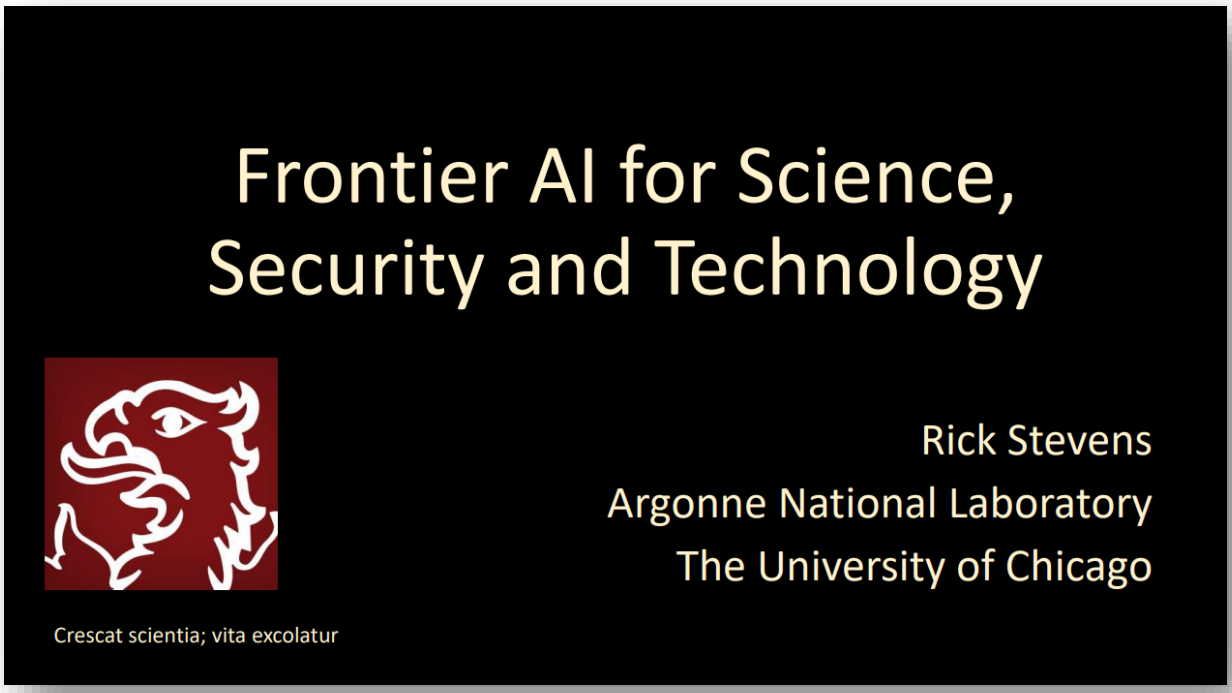
HPC User Forum Update

Growing Momentum for AI in Science with FASST Initiative

Thomas Sorensen
November 2024

IN THIS UPDATE

The HPC User Forum was established in 1999 to promote the health of the global HPC industry and address issues of common concern to users. In September 2024, the 85th HPC User Forum took place at Argonne National Laboratory. This update summarizes a presentation from that conference given by Rick Stevens, Associate Laboratory Director and Argonne Distinguished Fellow. He is Argonne's Associate Laboratory Director for the Computing, Environment, and Life Sciences (CELS) Directorate and a Professor of Computer Science at the University of Chicago as well as playing other roles in national and academic advanced computing development. Stevens held multiple discussions at the 85th HPC User Forum, one of which was an overview of the Frontiers in Artificial Intelligence for Science, Security, and Technology (FASST) Initiative.

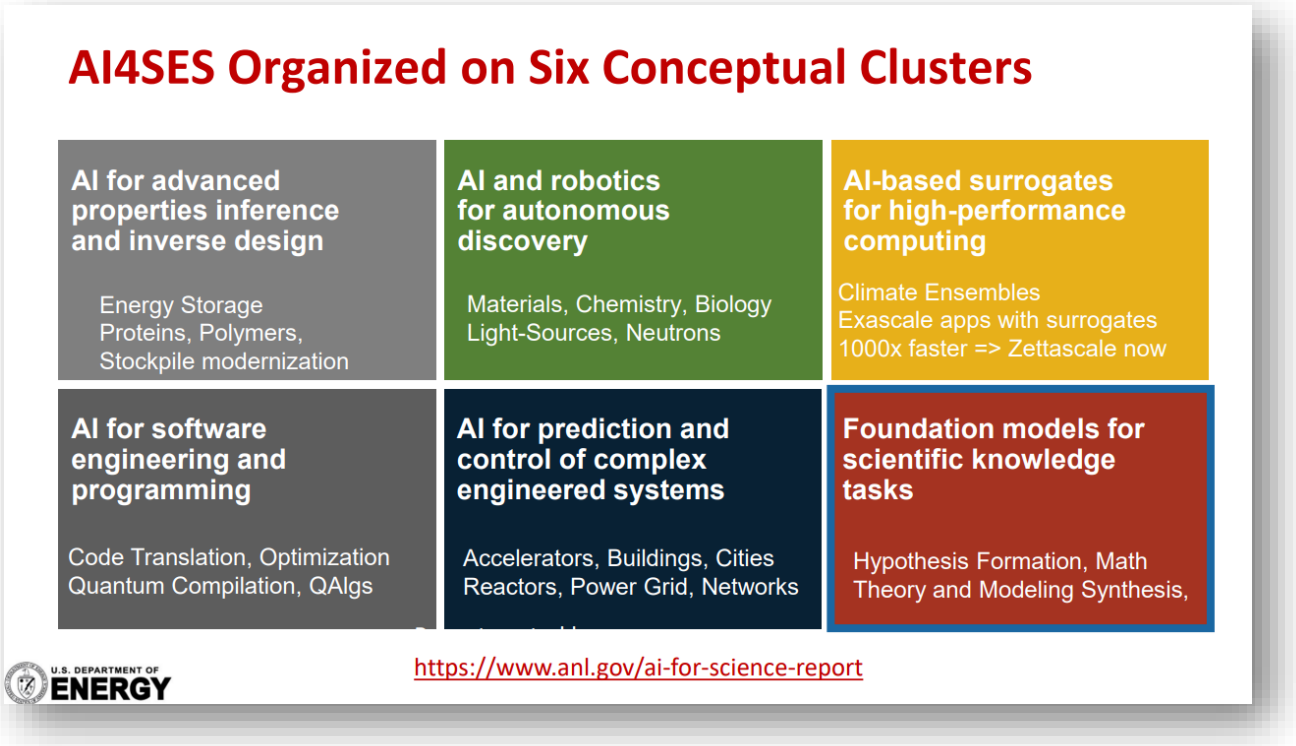


Source: Rick Stevens, 2024

PRESENTATION: FRONTIER AI FOR SCIENCE, SECURITY AND TECHNOLOGY (FASST) BY RICK STEVENS, ASSOCIATE LABORATORY DIRECTOR AT ARGONNE NATIONAL LABORATORY

The genesis of the FASST initiative started in 2019 with a series of town hall meetings involving over 1,300 members from three major DOE computing science labs (Argonne, Lawrence Berkeley, and Oak Ridge), including a meeting in Washington, D.C. These meetings gathered and coalesced data pertaining to the goals and interests among the scientific community on the role AI could play in their research and discovery efforts. Through much of the organizational activity, which had hit a roadblock with the Covid-19 pandemic and subsequent restrictions, the momentum of AI technology, specifically generative AI, continued to mount. As developments in the technology gained speed, efforts were redoubled, now including input from the six major DOE labs (Argonne, Oak Ridge, Lawrence Berkeley, Lawrence Livermore, Los Alamos, and Sandia) in 2022. This group concluded that multimodal and multitask models were taking over, a deviation from previous predictions. These town hall meetings predicted the strength of what would ultimately be called foundation models. Soon after, workshops began in the summer of 2022 divided into six buckets detailed in Figure 1.

FIGURE 1



Source: Rick Stevens, Argonne National Laboratory, 2024

"Little did we know the bottom one (foundation models for scientific knowledge tasks) would be the hyper-exponential thing, but we knew that all of these were important for the DOE mission." The group followed through with these workshops, produced reports, and by the summer of 2023 they began to see traction. During this time, Stevens and leaders from other laboratories engaged with policymakers and legislators in Congress on the details of their goals and vision for AI development and usage in their scientific computing. In September of that year, Stevens testified to members of Congress on the necessity of timely action and investment into AI as a strategic international imperative for scientific discovery, national security, energy security, and more. This testimony received considerable attention and although the group continued to promulgate these messages, there still was not yet a DOE initiative to go along with this momentum.

That fall of 2023, the White House released an executive order on AI. This voluminous document contained detailed orders for multiple agencies that had already been in development throughout the year with some input from members of national labs. In the document there were two major elements that Stevens had been working towards: first, the implementation of the plan that had been penned in 2022 – AI for Science, Energy, and Security; second, the recognition that powerful AI tools could be dangerous in the wrong hands and that NNSA labs should own the analysis of advanced AI for chemical, biological, radiological, and nuclear threats. These developments were considered critical by members of national laboratories like Stevens', who were having their predictions and notions recognized in these new policies.

FIGURE 2

Frontiers in Artificial Intelligence for Science, Security and Technology (FASST)

The proposed **Frontiers in AI for Science, Security, and Technology (FASST)** initiative leverages DOE's enabling infrastructure to deliver key assets for the national interest:

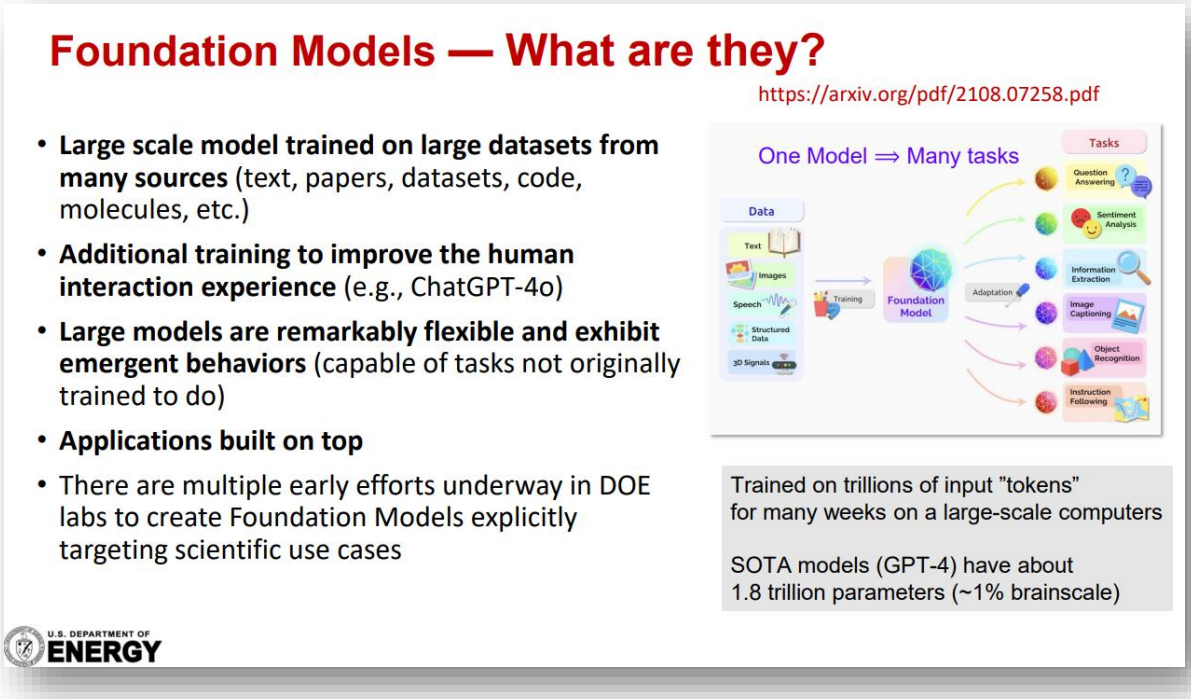
- **Advance National Security**. The development of AI models for national security applications, such as threat detection and strategic deterrence is crucial to maintaining America's defensive posture.
- **Attract and build a talented workforce**. FASST is the most ambitious AI initiative of its kind. This mission will attract, train, and retain top scientific talent for a leading capability deployed in the public interest.
- **Harness AI for Scientific Discovery**. FASST will develop AI tools that will dramatically reduce the time to discovery and extend the nation's competitive edge in technological innovation.
- **Address Energy Challenges**. FASST will unlock new clean energy sources, optimize energy production, and improve grid resilience, and build tomorrow's advanced energy economy. America needs low-cost energy to support economic growth and FASST can help us meet this challenge.
- **Develop technical expertise necessary for AI governance**. FASST will provide insight and independent expertise to quickly inform and validate standards and regulations for a responsible and safe AI industry.

Source: Rick Stevens, Argonne National Laboratory, 2024

Not long after the executive order, NNSA released a report on AI for Nuclear Deterrence in "as much of an open statement as the NNSA could make about the role AI could play in the deterrence role of the department." This report outlined the goal of leveraging AI technology in their mission of ensuring the usability and success of the nuclear stockpile and other related items. These developments were followed by another workshop at Argonne in the winter of 2022 on AI for Energy. While this topic was already included in earlier workshops, the energy labs had more to contribute, which resulted in these further workshops addressing AI R&D for nuclear, power grid, carbon management, energy storage, and energy materials.

The next spring, after some deliberation, the FASST Initiative debuted at the SCSP AI Expo during a DOE announcement. Notably, during this announcement, Divyansh Kaushik, a VP at Global Beacon Strategies, a Washington, D.C.-based strategic advisory firm, remarked, "It is arguably the most important AI initiative yet from the Biden administration." The next week, Senator Schumer, along with a bipartisan group of other legislators who had held an AI focused study group over the last year, released a report stating that the country should be spending $32B a year on non-defense AI. At that point, the US was only spending about $3B a year on non-defense AI. While this report presented solid rationale, details of funding programs were not provided. The report echoed sentiments of an earlier piece published by the Special Competitive Studies project, signaling that the public and private sectors had begun aligning on this topic.

FIGURE 3

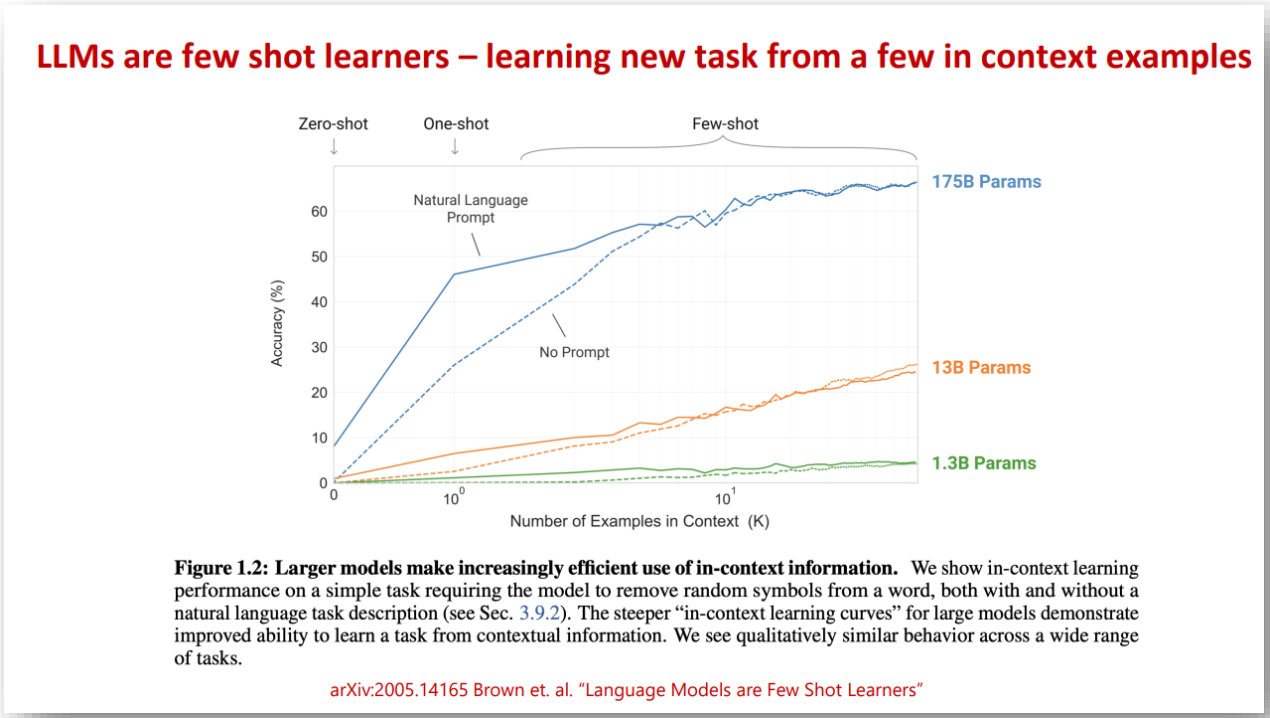


Source: Rick Stevens, Argonne National Laboratory, 2024

Education began across labs and legislators to understand what these projects could look like in the future. These efforts resulted in a piece of legislation called the DOE AI Act. It asked for $12B over 5 years. This was an authorization bill, meaning that policymakers design projects and their goals with budgetary considerations all accounted and numerated. None of the funding is created, rather, it "gives permission to ask for the money." The bill was endorsed by bipartisan legislators, attached to the National Defense Authorization Act (NDAA) and was publicized in a press release.

In the summer of 2023, an office was created that reports to the Deputy Secretary for Energy and Innovation called the Office of Critical and Emerging technologies. It owns coordination responsibilities for AI, quantum, bio-economy, and other emerging technologies. This office has been a communication nexus for the initiative.

FIGURE 4



Source: Rick Stevens, 2024

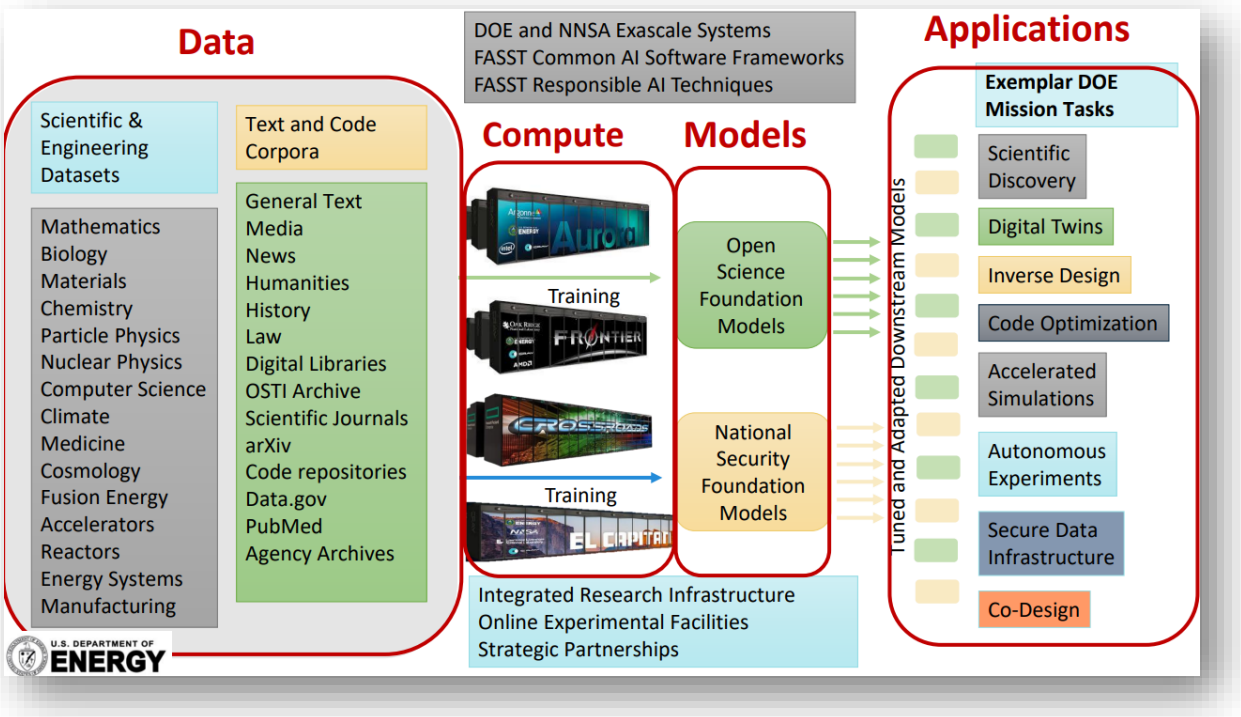
"This is not your grandparents' AI," says Rick Stevens, referring to modern foundational models. "The key concept is building on the notion of foundational models. Large Language Models are an example of foundational models built on language, but foundational models can be applied to other things like protein sequences or materials." This effort is a formalization of Stevens' vision of moving towards a

small number of general purpose models was also reflected in a paper on key transformers released in 2017 by Google researchers with other academic contributors. The idea is that all of these large general models at-scale are learning the same representation of the world: the platonic model hypothesis. This paper argues that scale matters and the detailed structural differences of models (how many layers, detailed attention mechanisms, etc.) don't matter as much. They are learning the same representation whether they're coming from OpenAI, Anthropic, or other sources. To Stevens, this means the foundation models, scale, and training on multi-modal data is convergent in some sense.

The goal is models built like large language models, trained by huge amounts of data, then specialized downstream for specific tasks. They're trained on text, images, sound, and so on with a single model in the middle, with it all tokenized and integrated, and applications are built on top of this. Google Gemini is built in this way and was perhaps one of the first major public examples of this design. It has discreet back end decoders based on the data modality. Things that are audio get interpreted as audio, things that are text get decoded as text and so on. These models are trained on a large corpus of data in which science is just a small fraction. This is a problem for laboratories like those in the DOE when they are trying to conduct their advanced research.

By this time the FASST Initiative had a direction, budget planning numbers, and a proposed activities scope; but what does it actually look like? FASST as a program is designed to have four pillars of activity: Data, Compute, Models, and Applications.

FIGURE 5



Source: Rick Stevens, Argonne National Laboratory, 2024

These pillars of activity include trying to organize the scientific and engineering data that exists inside the DOE and their sister agencies for training and building out compute both for large scale training and inference. This includes building models but only a handful of key foundation models as well as building many applications on top of them to serve the scientific and engineering research communities. These foundation models are to be mapped to specific DOE mission areas including clean energy systems, computational intelligence, nuclear security, advanced manufacturing, and more. Two models that Stevens noted as especially important are designed for knowledge integration and computation management. These scientific models function in a different language space than widely used large language models, being prompted by structural language as opposed to a spoken or written language.

There are efforts to broaden these narrowly talented models to include entire fields. For instance, there are DOE discussions in progress with the National Cancer Institute to build a single model that is trained on all cancer information available. A model like this would be trained on a diverse data set and have many downstream applications but would require a tremendous amount of data and compute power to train. The FASST initiative and DOE are also aiding the development of future models by creating formalisms for calculating training time and demands on different machines based on parameter and token counts.

In this “FASST in a nutshell” presentation by Rick Stevens, the fast-paced and critical development of scientific and engineering AI integration is fully illuminated. Openness and hastiness in reacting to this quickly developing technology space is crucial for maintaining and growing leadership in the established mission areas of national science research organizations.

For more information or to view this and other presentations given at HPC User Forums dating back to 2008, visit www.hpcuserforum.com.

About Hyperion Research, LLC

Hyperion Research provides data driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
