AI4SCIENCE: ARGONNE NATIONAL LAB AND BEYOND

FRANCIS J ALEXANDER

Argonne National Laboratory
And University of Chicago
Consortium for Advanced Science and Engineering (CASE)

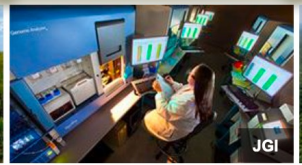
Presented at ai4science: a workshop
Harvard University, Cambridge MA
February 1, 2024

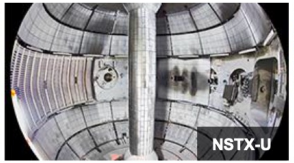
U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

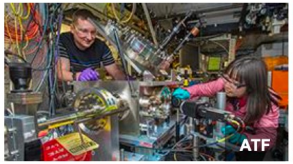
DOE Has Multiple Efforts in AI for Science

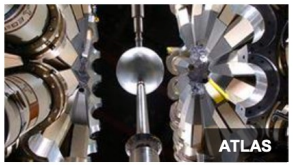
Supports 17 Nat'l Labs and Researchers at Hundreds of Universities

FY 2021
28 scientific user facilities
36,000+ users









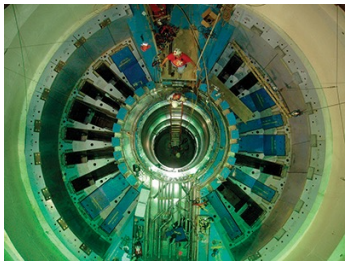
U.S. DEPARTMENT OF ENERGY
Office of Science

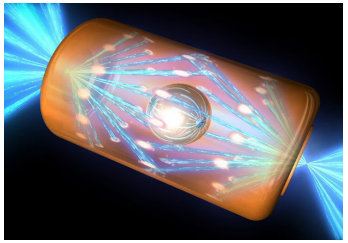
Acronym decoder at https://science.osti.gov/User-Facilities

AI4SCIENCE: DOE'S UNIQUE POSITION

- **Operates the most capable computing systems and the world's largest collection of advanced experimental facilities**
- Responsible for US nuclear security through deep partnerships across government
- **Largest producer of classified and unclassified scientific data in the world**
- Strongest foundation combining physical, biological, environmental, energy, mathematical and computing sciences
- **Largest scientific workforce in the world**
- Strong ties with private sector technology and energy organizations and stakeholders

World's best experimental facilities and supercomputers




DOE HAS BEEN GATHERING WIDE COMMUNITY INPUT (>1300 RESEARCHERS)

What changed in three years?

- Language Models (e.g. ChatGPT) released
- Artificial image generation took off
- AI folded a billion proteins
- AI hints at advancing mathematics
- AI automation of computer programming
- Explosion of new AI hardware
- AI accelerates HPC simulations
- Exascale machines start to arrive

2020 DOE Office of Science ASCR Advisory Committee report recommending major DOE AI4S program

https://www.anl.gov/ai-for-science-report

U.S. DEPARTMENT OF ENERGY | Office of Science

National Nuclear Security Administration

WORKSHOPS ORGANIZED ON SIX CROSSCUTTING THEMES

AI for advanced properties inference and inverse design

Energy Storage
Proteins, Polymers,
Stockpile modernization

AI and robotics for autonomous discovery

Materials, Chemistry, Biology
Light-Sources, Neutrons

AI-based surrogates for high-performance computing

Climate Ensembles
Exascale apps with surrogates
1000x faster => Zettascale now

AI for software engineering and programming

Code Translation, Optimization
Quantum Compilation, QAlgs

AI for prediction and control of complex engineered systems

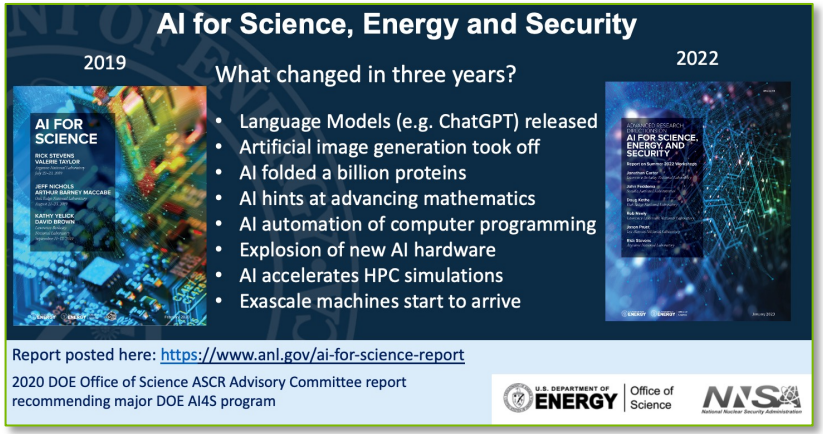
Accelerators, Buildings, Cities
Reactors, Power Grid, Networks

Foundation, Assured AI for scientific knowledge

Hypothesis Formation, Math
Theory and Modeling Synthesis,

WE WANT TO BUILD THE WORLD'S MOST POWERFUL FOUNDATION MODELS FOR SCIENCE

- New exascale platforms provide ideal platforms for training and evaluating O(10^{12} parameters) language models for science and engineering.
- Building state-of-the-art LLMs will require large allocations of machine time (e.g., O(30-100) exaflop-days) for training and downstream tuning, alignment, and evaluation.
- The scale of effort required to prepare training data, and the scale of computing resources needed to build and train models, suggests that it would be optimal to collaborate on very large models rather than working independently on many smaller models.



Many from this global community—over 1,200 from academia, federal laboratories, and industry—began to work together in 2019 through seven US Department of Energy-sponsored workshops and two public reports (in 2020 and 2023).

Download the 2023 Report:

https://tpc.dev

TRILLION PARAMETER CONSORTIUM: THREE GOALS

Goal 1. Build an open community of researchers that are interested in creating state-of-the-art large-scale generative AI models (FMs/LLMs) aimed broadly at advancing progress on scientific and engineering problems, by sharing methods, approaches, tools, insights, and workflows.

Goal 2. Incubate, launch, and loosely (voluntarily) coordinate specific projects to build specific models at specific sites and attempt to avoid unnecessary duplication of effort and to maximize the impact of the projects in the broader AI and scientific community. Where possible we will work out what we can do together for maximum leverage vs. what needs to be done in smaller groups.

Goal 3. Create a global network of resources and expertise that can help facilitate teaming and training the next generation of AI and related researchers interested in the development and use of large-scale AI in advancing science and engineering.

Initial TPC Partners Come from Around the Globe

AI Singapore
Allen Institute For AI
AMD
Argonne National Laboratory
Barcelona Supercomputing Center
Brookhaven National Laboratory
CalTech
CEA
Cerebras Systems
CINECA
CSC – IT Center for Science
CSIRO
ETH Zürich / CSCS
Fermilab National Accelerator Laboratory
Flinders University
Fujitsu Limited
HPE
Indiana University
Intel
Juelich Supercomputing Center
Kotoba Technologies, Inc.
LAION
Lawrence Berkeley National Laboratory
Lawrence Livermore National Laboratory
Leibniz Supercomputing Centre
Los Alamos National Laboratory
Microsoft
National Center for Supercomputing Applications
National Energy Technology Laboratory
National Institute of Advanced Industrial Science & Technology (AIST)
National Renewable Energy Laboratory
National Supercomputing Centre, Singapore
NCI Australia
New Zealand eScience Infrastructure
Northwestern University
NVIDIA
Oak Ridge National Laboratory
Pacific Northwest National Laboratory
Pawsey Institute
Princeton Plasma Physics Laboratory
RIKEN
Rutgers University
SambaNova
Sandia National Laboratories
Seoul National University
SLAC National Accelerator Laboratory
Stanford University
STFC Rutherford Appleton Laboratory, UKRI
Texas Advanced Computing Center
Thomas Jefferson National Accelerator Facility
Together AI
Tokyo Institute of Technology
Université de Montréal
University of Chicago
University of Delaware
University of Illinois Chicago
University of Illinois Urbana-Champaign
University of Michigan
University of New South Wales
University of Tokyo
University of Utah
University of Virginia

(this list continues to expand)

https:// tpc.dev

AURORA

Leadership Computing Facility Exascale Supercomputer Overview

Peak Performance

≧ 2 Exaflops DP

Intel GPU

Intel® Data Center GPU Max Series 1550

Code named "PVC"

Intel Xeon PROCESSOR

Intel® Xeon® CPU Max Series with HBM

Code named "SPR+HBM"

Platform

HPE Cray-Ex

System Size

166 Compute Racks
10,624 nodes
21,248 CPUs
63,744 GPUs

Compute Node

2 CPU, 6 GPU
1 TB DDR5
1 TB HBM
8 Fabric NICs
Node Unified Memory Architecture

Aggregate System Memory

DDR5 10.9 PB, 5.95 PB/s
HBM CPU 1.36 PB, 30.5 PB/s
HBM GPU 8.16 PB, 208.9 PB/s

System Interconnect

HPE Slingshot 11
Dragonfly topology with adaptive routing
2.12 PB/s Peak Injection BW
0.69 PB/s Peak Bisection BW

High-Performance Storage

220 PB
31 TB/s DAOS bandwidth
1024 DAOS nodes

Programming Environment

oneAPI
- C/C++
- Fortran
- SYCL/DPC++
- Python

Aurora MPICH and oneCCL
OpenMP offload
Kokkos, RAJA
Intel PerformanceTools, Intel gdb
Tensorflow, PyTorch
DDP, Horovod, DeepSpeed
oneDAL and ScikitLearn
Python Libraries
JupyterHub
Julia, Numba
Spark
MLDE, SmartSim

WHAT IS AURORAGPT?

- AuroraGPT is a series of LLMs (7B, 70B, 200B, 1000B, etc)
- Trained on a mixture of general text, code and scientific domain knowledge (Biology, Physics, Materials/Chemistry, Climate, Computer Science, Nanoscience, Cancer,
- Domain knowledge extends beyond information in Common Crawl (RP2, Dolma, Pile), ArXiv, PMC, etc. to include text encoded forms of structured scientific data from variety of domain data resources
- AuroraGPT is expected to have multiple phases of development
  - Phase 1 – Text only models – raw and instruct models (2023/2024)
  - Phase 2 – Basic multimodal models (2024/2025)
  - Phase 3 – Advanced scientific multi-model models (2025/2026)
- **Explore pathways** towards a “Scientific Assistant” model
- **Build with international partners** (RIKEN, BSC)
- **Multilingual** – English, French, German, Japanese, Spanish, Italian
- **Multimodal** – images, tables, equations, proofs, time-series, graphs, fields, sequences, etc.

AUTONOMOUS DISCOVERY @ARGONNE

- **The vision**
  - A system that starts with a high-level description of a hypothesis and autonomously carries out computational and experimental workflows to confirm or reject that hypothesis
  - **Use of AI in robotics and simulations to close the loop on planning, execution, and analysis of experiments**
- Builds on
  - **AI approaches to planning** (multiple steps), and integration of results, causality, etc.
  - **Machine learning/simulation** to design and predict exp properties and outcomes
  - **Automation of experimental protocols** (robotic steps and workflows)
  - **Active Learning or RL** for selection of next experimental targets, etc.

ARTIFICIAL INTELLIGENCE
GUIDED, ROBOTICALLY EXECUTED EXPERIMENTS

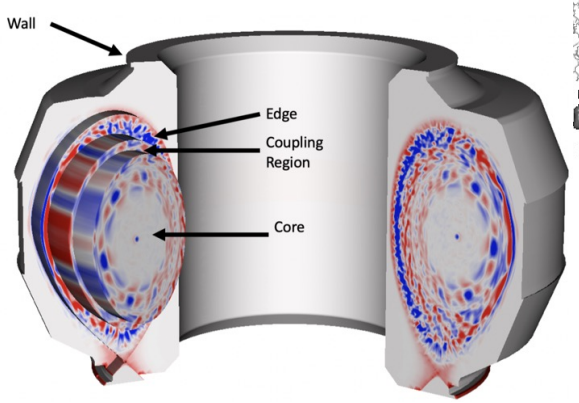
https://github.com/anl-sdl/
https://www.cs.uchicago.edu/~rorymb/

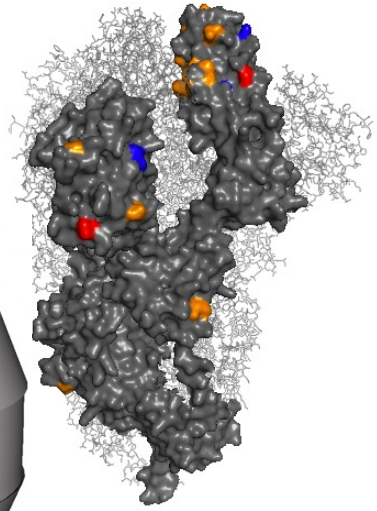
DECISION SUPPORT FOR COMPLEX SYSTEMS AND 'WICKED' PROBLEMS

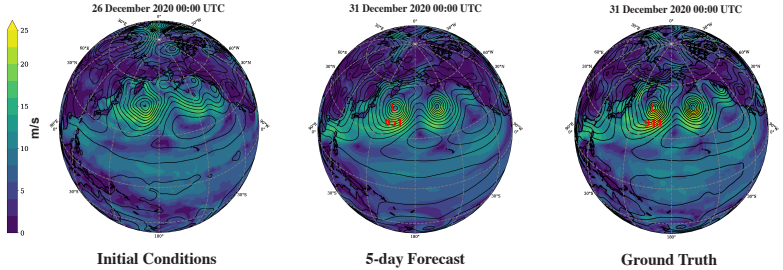
- Complex natural and engineered systems involve a diverse set of physical, cyber-physical, and social-technical aspects and the strong interactions among them.
- These systems typically have complex life cycles and involve multiple stakeholders.
- With currently available computational tools and algorithms, principled decision-making remains a challenge for these types of systems with deep uncertainties, often coined as "wicked."
- Wicked problems are not well-bounded, are framed differently by various groups/individuals, encompass large to existential scientific uncertainties, and may be poorly understood until the point when a solution has been achieved.
- Meeting the challenge of Wicked Problems requires an integrated research and development program that provides the underlying theory and fundamentals, algorithms (mathematics, computer science, HPC, economics, psychology, domain sciences…)
- Building a capability across the labs to address these types of challenges

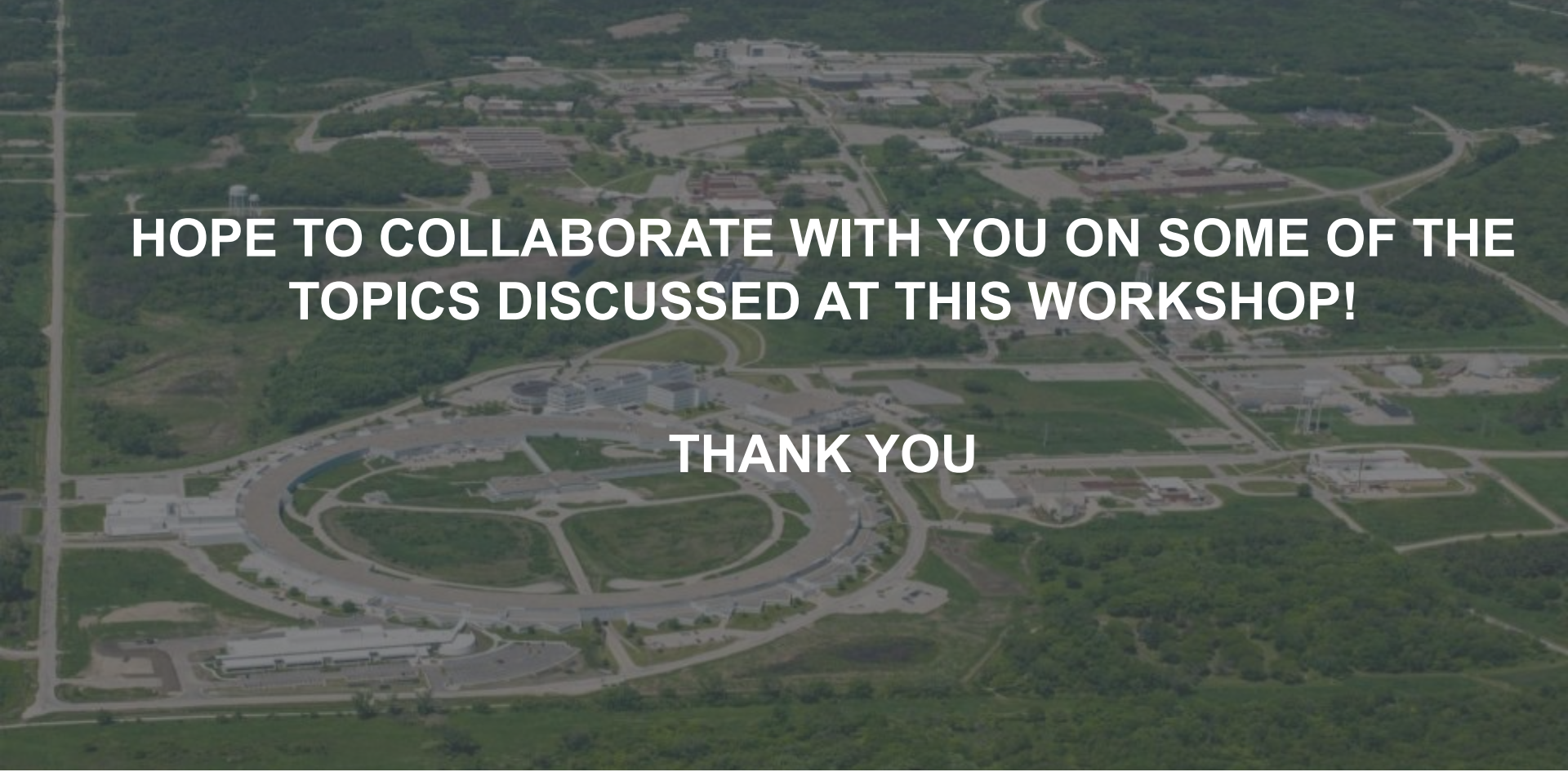
RELATED PROJECTS AT ANL

- **Decision Support for Complex Systems and 'Wicked' Problems**
- **"Fail Fast" Design**
- **Optimal Operator Design (including Optimal Experimental Design)**
- **Computational CoDesign (Goal/Theory/Operator/Algorithm/Hardware)**
- **Number of large domain specific efforts at Argonne National Laboratory:**
  - Predicting Disruptions in Tokamaks
  - Climate/Weather Using LLM's
  - Genomics predicting likely viral evolution
  - Pandemic Response: EMERGE Modeling and optimal control









HOPE TO COLLABORATE WITH YOU ON SOME OF THE TOPICS DISCUSSED AT THIS WORKSHOP!

THANK YOU