

AI INITIATIVES AND DECISION MAKING AT SCALE

2/12/2025

FRANCIS J ALEXANDER

Argonne National Laboratory

IN THE COMING YEARS THE U.S./GLOBAL COMMUNITY WILL FACE EVER MORE DIFFICULT DECISIONS

- High-stakes decisions to respond to/guide/influence/control the behavior of large-scale, complex systems.
- Such decisions, often time-critical, must be made amidst tremendous uncertainty and while working with constrained resources.
- These complex natural- and human-engineered systems encompass a broad spectrum of elements: physical, cyber-physical, and socio-technical, economic, etc
- They also feature strong interdependencies and involve multiple stakeholders
- More often than not, a predictive theory is lacking and data in these scenarios are scarce, noisy, indirect, and acquiring them can be difficult/impossible, costly intrusive.

SUCH CHALLENGES ARE HERE ALREADY

- Complex scientific discovery process: e.g., Swiftly developing new drugs/vaccines to combat existing and emerging diseases. The discovery pipeline is itself a complex system where treatments which are unlikely to succeed should be weeded out early in the process.
- Effectively managing biosafety responses that strike a balance between various critical objectives, such as saving lives, maintaining healthcare systems, and sustaining economic activities.
- Developing tools for biosurveillance that can manage large amounts of information from human health, animal health, microbial ecosystems, agricultural systems, and the urban/built environment.

WHAT'S NEEDED TO DEAL WITH THESE ISSUES

- This set of multifaceted challenges requires both leveraging existing as well as greatly extending capabilities in computer science, applied mathematics, high-performance computing, **artificial intelligence (AI) and machine learning (ML)**, operations research etc.
- Current computational tools and AI algorithms face difficulties in principled decision-making under deep uncertainties.
- While ML/AI methods have come a long way, they are primarily predictive and have yet to bridge the gap to a decision-making paradigm with performance guarantees where verification and validation are crucial.
- Such guarantees are key to the adoption of proposed solutions, especially for high consequence issues.
- ***Develop a practical yet rigorous framework for resource constrained (infrastructure, computational, etc.) effective decision-making at scale for high-consequence, uncertain complex systems.***
- Provide performance guarantees, quantified uncertainty for the output of this framework through a rigorously validated and verified analysis
- Apply and extend the state-of-the-art in operations research, ML/AI, applied mathematics, HPC, modeling and simulation including adaptive, agent-based, multimodal and multiscale methods.
- Build a distributed team capable of tackling such future challenges

The Department of Energy (with partners) is Well-positioned to Address These Challenges



DEPARTMENT OF ENERGY

- Foundational research in HPC, ML/AI, Complex Systems, Mathematics, applied to real-world problems at scale
- Operates the most capable computing systems and the world's largest collection of advanced experimental facilities
- Responsible for many aspects of US security through deep partnerships across government (Nuclear, Bio, ...)
- 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
- Research in responsible ML for high consequence decisions
- Multidisciplinary expertise and experienced teams
- Considerable expertise in UQ – key component

DOE IS WELL POSITIONED TO ADDRESS THESE CHALLENGES

- Foundational research in Exascale computing and HPC and scientific AI for variety of areas
- Advances in AI and secure exascale computing
- Vital partnerships with health institutions and agencies
- HIPAA compliant data enclaves and secure HPC
- Foundational research in HPC and scientific AI for the grid and other areas [DOE Infrastructure and Resources](#)
- Ability to validate developments in real-world situations
- Existing partnerships in Federal emergency response and data for energy systems, biological, and nuclear threats
- DOE has the technical ability to lead in and partner with other agencies to advance the use of decision making under deep uncertainty across the government.
- No other agency is as well placed to do this.

INTEGRATED RESEARCH AND DEVELOPMENT PROGRAM IS ESSENTIAL

- While components of this program exist in isolation
- There is no comprehensive, coherent effort which can scale to tackle problems like those mentioned above.
- Such a program would involve research and integration of the following areas.
 - Computer Science
 - Operations Research
 - High Performance Computing
 - Applied Mathematics
 - Statistics
 - Machine Learning / AI
 - Uncertainty Quantification
 - Psychology
 - Economics
 - Sociology
 - Relevant Domain Sciences, and more...

We have begun the process with Several Exemplars

- *How do we build ecosystems and technological capacity for model maturation?*
- *Do we have the data to accurately measure and forecast disease burden and services?*
- *What is the role of forecasting in addressing the know-do gap?*
- *If you can forecast, what is optimal intervention?*
- *Defining a health system objective function*
- *What is the level of complexity needed to link forecasting to action?*
- *Measures of value to close the loop: Did the forecast come true?*
- *Building trust and defending against bias: Toward inclusive and equitable solutions*
- *Connecting and Interoperating across Other Societal Forecasting efforts*



- ***Multiscale Ecosystem complexities for Robust, Generalized Epidemiology***
- ***“Fail Fast”***
- ***Climate***



You are here

**Human Progress
Through Time**



ON THE EXASCALE FRONT

- Frontier/Aurora together > 100K GPUs for scientific computing, AI and advanced data analysis
- Robust software environments (ECP) for ModSim, AI, Analysis
- Over the next five years these systems will be upgraded to 5-10 Eflops FP64 and 100's of Eflops for AI, approaching Zettascale by 2030
- Enough capacity for significant use in AI, FM development and integrated reasoning systems for AI4Science

COMPUTING AT THE EXASCALE WILL EVOLVE



- Traditional ModSim are generally FP64 applications
- Mixed precision (FP32, FP16, etc.) can often deliver 10x more
 - > 10EF mixed precision on F/A vs > 1 EF F/A for fp64
- Many applications can leverage mixed precision with molecular dynamics being very good example (Anton, etc.)
- Increasing use of AI/ML surrogates provides even more upside potential with some applications > 100x to > 100,000x acceleration compared to FP64 baselines (e.g. Digital Twins)
- Finally, pure end-to-end FM model replacements for simulation (sometimes called ML emulators are possible in some areas)

THE 2022 WORKSHOPS RECOGNIZED THIS TREND

AI4SES SIX CONCEPTUAL CLUSTERS

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

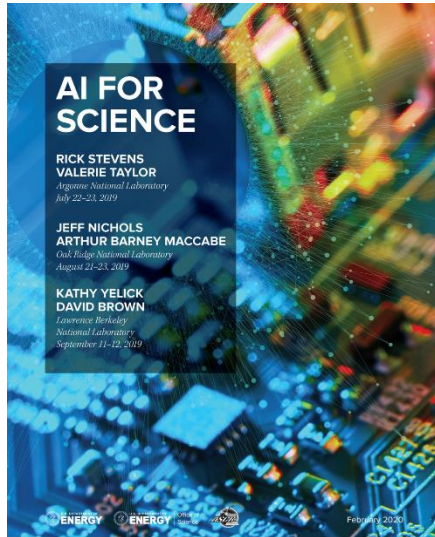
Accelerators, Buildings, Cities
Reactors, Power Grid, Networks

Foundation models for scientific knowledge tasks

Hypothesis Formation, Math
Theory and Modeling Synthesis,

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

DOE Has Been Gathering Wide Community Input (>1300 researchers)



Much accelerated 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>

DOE Labs reinforce AI for Energy

ADVANCED RESEARCH
DIRECTIONS ON

AI FOR ENERGY

Report on Winter 2023 Workshops

Claus Daniel

Argonne National Laboratory

Jess C. Gehin

Idaho National Laboratory

Kirsten Laurin-Kovitz

Argonne National Laboratory

Bryan Morreale

National Energy Technology Laboratory

Rick Stevens

Argonne National Laboratory

William Tumas

National Renewable Energy Laboratory

Key Findings for Establishing the Cross-cutting Aspects of AI Supremacy Needed to Ensure Success	5
in Energy Mission Areas	5
High-Consequence	5
Urgency	6
Complexity	7
01. Nuclear Energy	8
1.1 Grand Challenges	8
1.2 Advances in the Next Decade	11
1.3 Accelerating Development	13
1.4 Expected Outcomes	15
1.5 References	15
02. Power Grid	18
2.1 Grand Challenges	18
2.2 Advances in the Next Decade	19
2.3 Accelerating Development	21
2.4 Expected Outcomes	24
2.5 References	25
03. Carbon Management	26
3.1 Grand Challenges	26
3.2 Advances in the Next Decade	29
3.3 Accelerating Development	31
3.4 Expected Outcomes	32
3.5 References	32
04. Energy Storage	34
4.1 Grand Challenges	34
4.2 Advances in the Next Decade	36
4.3 Accelerating Development	39
4.4 Expected Outcomes	42
4.5 References	42
05. Energy Materials	44
5.1 Grand Challenges	44
5.2 Advances in the Next Decade	46
5.3 Accelerating Development	47
5.4 Expected Outcomes	49
5.5 References	49

OpenAI o1 is smarter than most humans

Norway Mensa IQ test

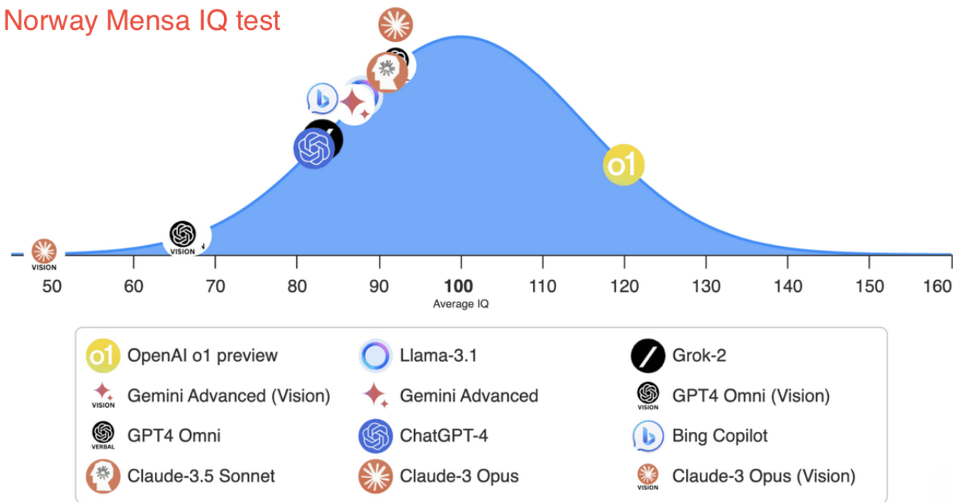


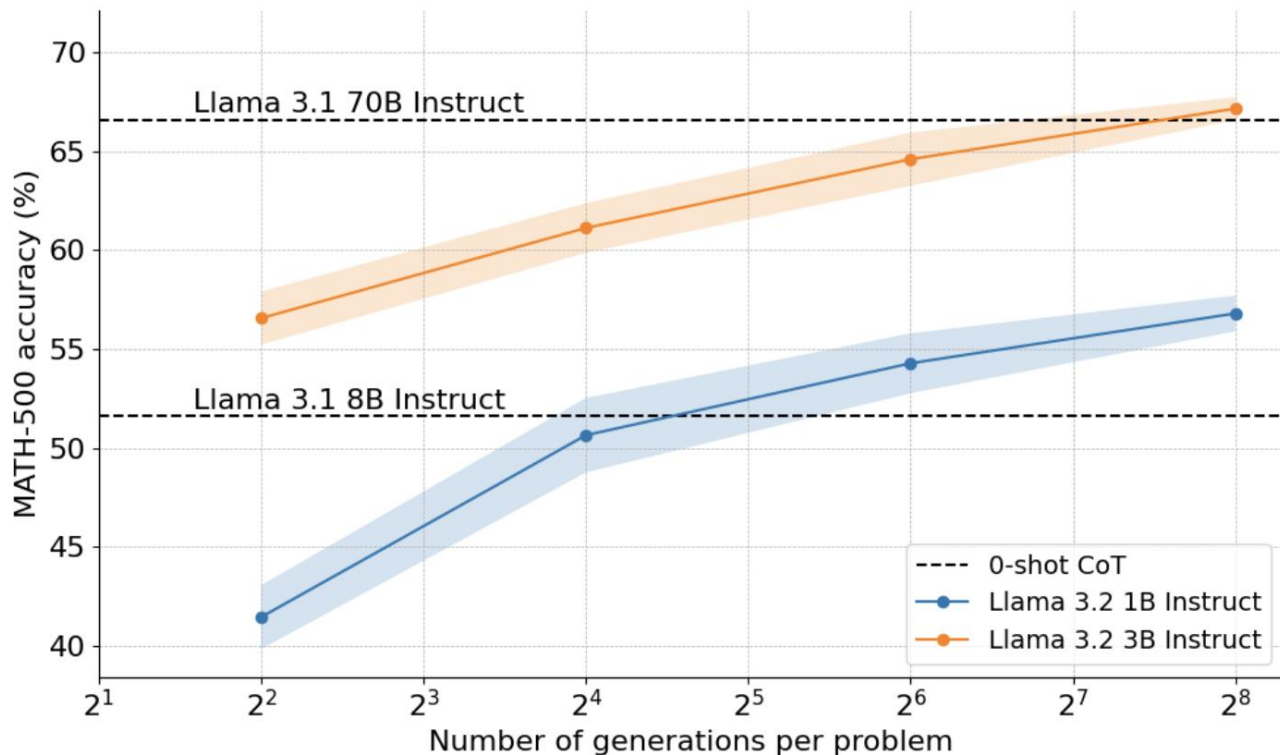
Image source: TrackingAI.org

O1 IQ level stays the same with non-public questions
(good results are not due to contamination/overfitting)

REASONING MODELS HAVE “THINKING TIME”

- Before responding, somewhat like the difference between type 1 thinking and type 2 thinking in humans
- Type 2 thinking is the idea of complex reasoning on a problem, considering multiple options, iterating, looking things up, perhaps consulting various sources, etc.
- Reasoning models are a step in this direction but still quite primitive in the approach
- Important ideas are Reward models, Evaluation models, Assessment models
- Distillation is another concept being used with reasoning models

SMALL MODELS CAN OUTPERFORM LARGE MODELS WHEN GIVEN ENOUGH TIME!



BOOTSTRAPPING AND DISTILLATION

- How to go from a non-reasoning model to a reasoning model?
- Start with hundreds of problems that need reasoning to solve
- Roughly, use prompting techniques to generate many detailed traces of reasoning on various problems

P → **T** → **S** **P** → ? → **S** **P** → ? → ?

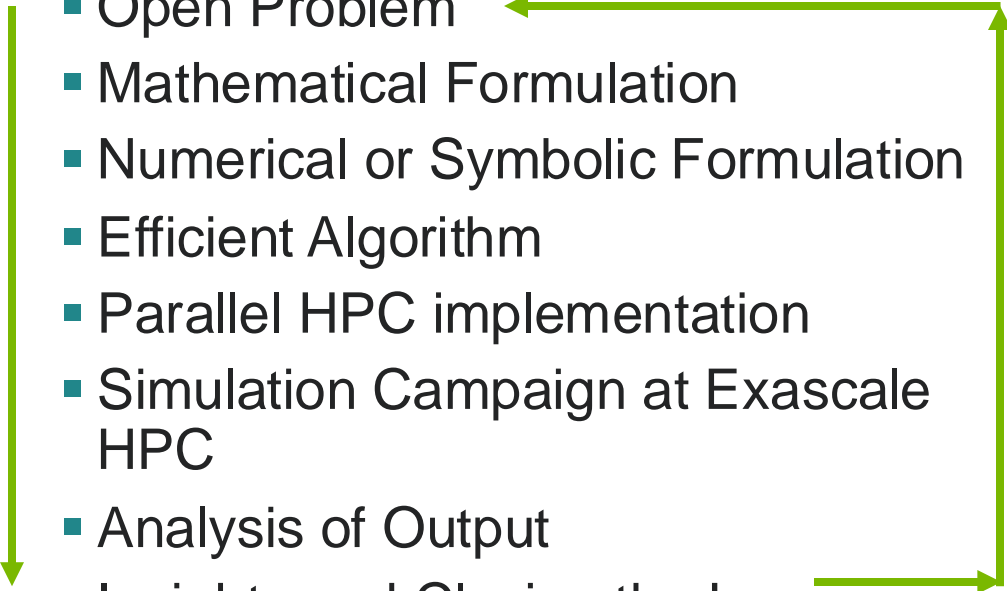
- Use these traces as data to drive RL loops to teach models how to do reasoning as part of normal “completion” like an extension of “instruction following” and “alignment”
- Order 300 examples with P and S, with perhaps 30 worked out with P, T and S are sufficient for a powerful model to learn a subdomain

FRONTIER MODELS WILL BE MORE USEFUL IF THEY CAN:

- Interact with HPC experimental codes/simulations
- Deeply understand DOE user facilities
- Understand and analyze multi-modal data
- Integrate better context from scientific experiments
- Generate correct references/reproducibility
- Provide uncertainty for responses
- Produce more detailed responses
- Improve knowledge on scientific topics
- Better produce original/novel recommendations

How much of these will just happen on the natural industrial progression and what will need forcing?

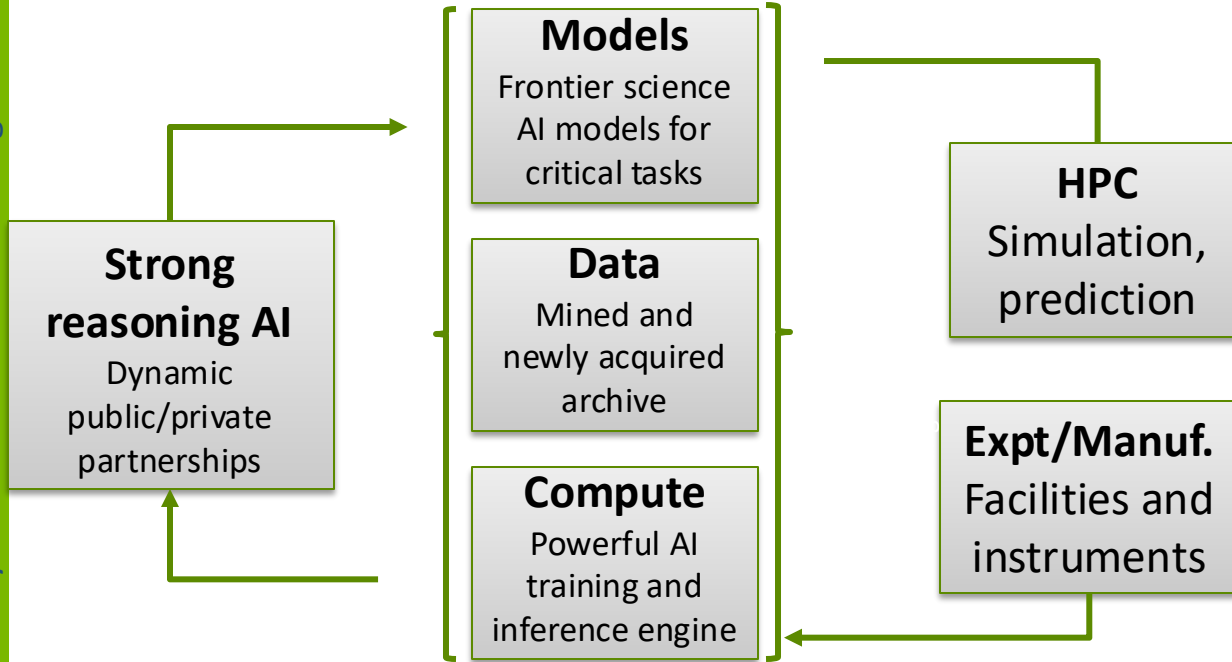
AI ASSISTANT FOR “COMPUTATIONAL” SCIENCE

- Open Problem
 - Mathematical Formulation
 - Numerical or Symbolic Formulation
 - Efficient Algorithm
 - Parallel HPC implementation
 - Simulation Campaign at Exascale HPC
 - Analysis of Output
 - Insights and Closing the Loop
- 



ELEMENTS OF A LARGE-SCALE SCIENTIFIC REASONING PLATFORM

Industry contributions and co-design



Platform capability

1. AI-driven HPC use
2. AI-driven facility use
3. Closed loop for acting on edge data
4. Automated reasoning and hypothesis exploration
5. Accelerated discovery of novel science and solutions

Increasing integrated technical maturity

A closed-loop accelerated scientific reasoning platform (fundamental AI research)

ACCELERATING DISCOVERY USING AI ASSISTANTS

Extraction, integration and reasoning with knowledge at scale

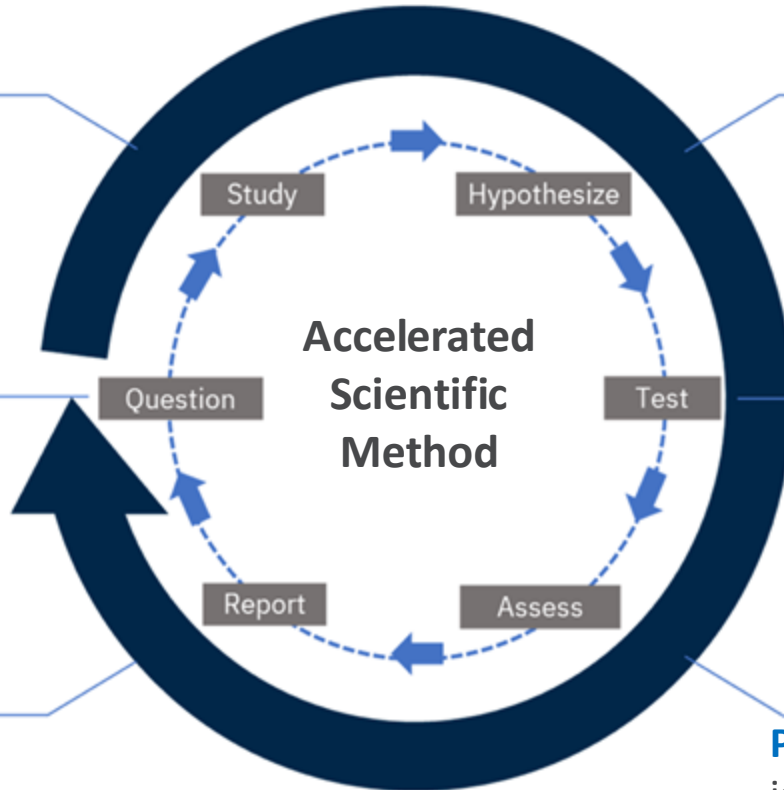
Tools help **identify new questions** based on needs and gaps in knowledge

Machine representation of knowledge leads to new hypotheses and questions

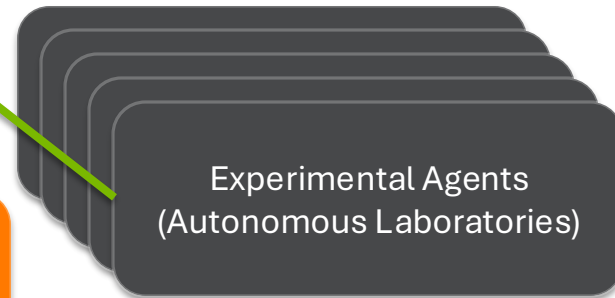
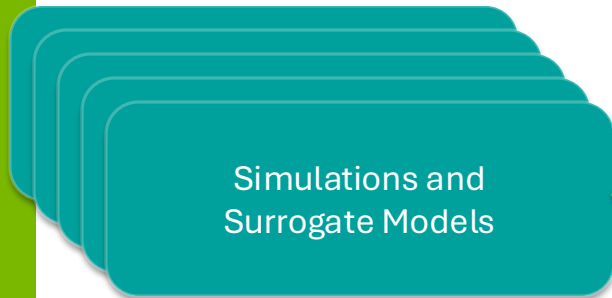
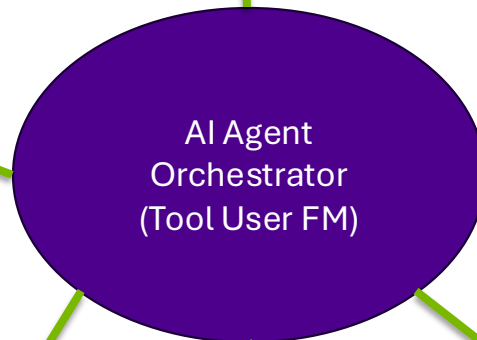
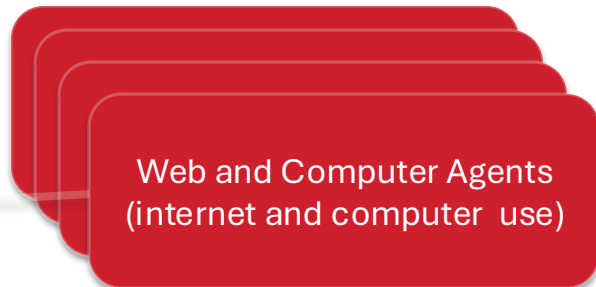
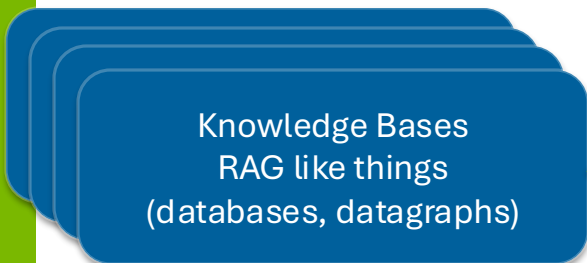
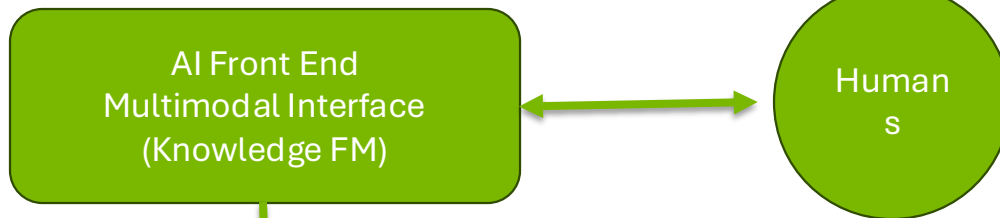
Generative models automatically propose new hypotheses that expand the discovery space

Robotic labs automate experimentation and bridge digital models and physical testing

Pattern and anomaly detection is integrated with simulation and experimentation to extract new insights



AI SYSTEMS APPS SKETCH



FASST: Data

- DOE laboratories have 100's of Petabytes of scientific data in many modalities, modalities that industry is not focused on
- The vast majority of this data is not publicly available or in a form that can be used to train AI models

- FASST will support laboratories and universities to **organize the collections** of scientific and engineering data at DOE labs and **prepare that data for training AI models**
- FASST will support the **analysis and tools needed to transform and curate** this data and to develop **representations needed to train models**
- FASST will support **domain data teams to validate models** and to demonstrate improved prediction performance through expanded data availability

FASST: Models

- The heart of FASST is the development of **foundation models** that span the scientific and engineering scope of the department

- **Each model will take a massive computing effort** to train (perhaps 8-10 can be built each year) and they can capture whole scientific domains (e.g. biology, particle physics, etc.)
- **Each model is trained on massive datasets** that have been created at laboratories (materials, chemistry, biology, particle physics, energy technology, nuclear security, etc.)
- FASST models are **large-scale** (many tens of billions of parameters), and trained on trillions of input tokens **however they may not be (natural language or transformer based) LLMs** depending on the data domain and data representation
- FASST **models will be designed to be deployed via AI applications stacks** and are not generally aimed at end users directly

FASST: Compute

- DOE has fielded three Exascale class computers and these will be the initial platforms for FASST model development
- However, Frontier level AI will need orders of magnitude more compute for training and for broad inference capacity over the next five years

- FASST will deploy mid scale systems at laboratories and via the cloud to support significant scale out of inference capacity for **data curation** and preparation work, for **AI applications development** and, for **AI research**
- FASST will **develop and deploy frontier level platforms for training the largest scientific AI models**
- FASST will invest in **partnerships with industry and universities to develop energy efficient next generation platforms**

FASST: AI Applications

- AI applications are **relatively lightweight systems** that provide a User Interface (or API) and an AI capabilities specific to some problem domain or scientific or engineering workflow

- AI applications can be **ensembles of models, combining agents, databases, simulations and other tools into a single workflow**
- FASST will develop a **framework for rapidly constructing scientific AI applications** that integrates the best of industrial models, open models, and laboratory and university developed models
- FASST AI applications need **to be updated frequently** with new model components and will be hosted on the FASST Inference platforms at the laboratories

INITIATIVE GOALS AND OUTCOMES

- **Notice of Request for Information (RFI) on Frontiers in AI for Science, Security, and Technology (FASST) Initiative; Reopening of Comment Period**
- <https://www.energy.gov/fasst>
- **Ensure US (DOE) leads the world in technical capability** for its missions in Science, Energy and National Security
- **Create, deploy and sustain world leading "frontier" AI systems and applications** for DOE mission areas to provide advantage to US and partners
- **Increase productivity and capabilities of the DOE laboratories, academic, agency and international partners**
- **Develop an AI-forward workforce** for DOE

- **Discovery Science** – accelerate and improve effectiveness
- **Energy Transition** – accelerate, reduce risk, improve translation
- **National Security** – anticipate risk, mitigate risks, accelerate mission



THANK YOU



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

