Los Alamos NATIONAL LABORATORY

Can AI Answer Every Science Question?

Diane Oyen
Artificial Intelligence Team Leader
Information Sciences Group

April 24, 2024

LA-UR-24-23931

Can AI Answer Every Science Question?

- Probably not, but it has answered some already
- What other science questions can AI answer?
- **Who** is going to answer those questions?

- Foundation models for science (aka LSM = Large Science Model) are looking like a good bet

Why LSM?

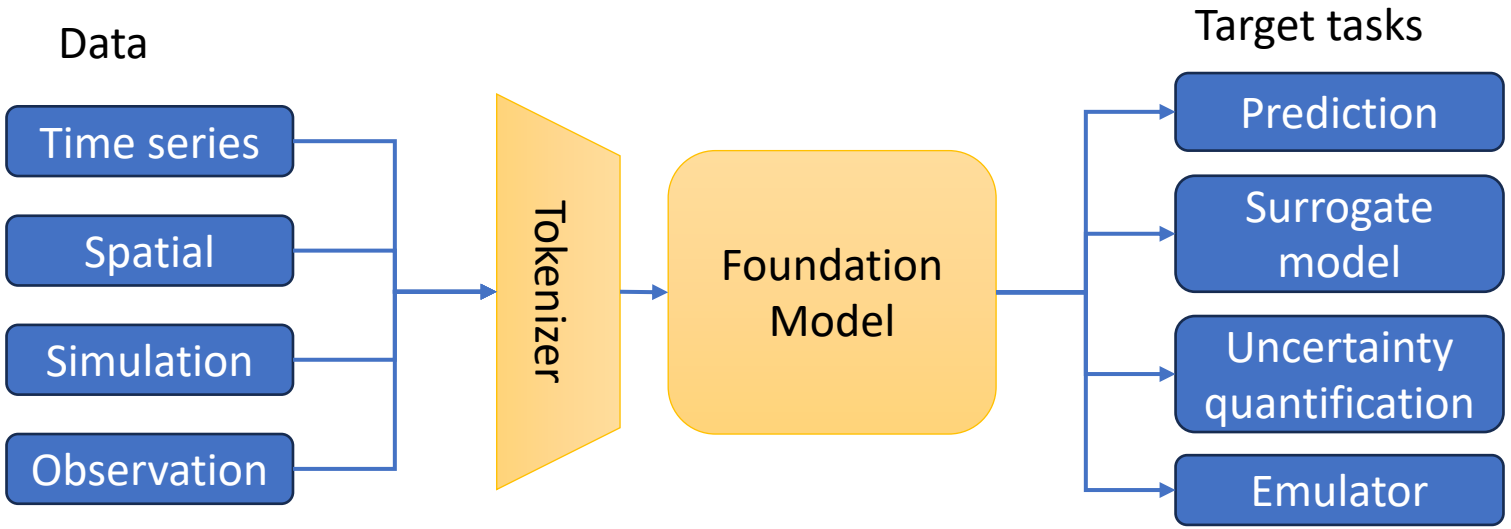
- FASST Initiative: Frontiers in Artificial Intelligence for Science, Security, and Technology https://fas.org/publication/recent-advances-in-ai-statement-kaushik/
  - Anticipated ECP-scale investment in AI
- Trillion-Parameter Consortium https://tpc.dev
  - “bring together groups interested in building, training, and using large-scale models with those who are building and operating large-scale computing systems.”
  - DOE labs, industry, international supercomputing orgs
- U of Michigan https://micde.umich.edu/news-events/annual-symposia/2024-symposium/
  - “SciFM are parameterized physical theories that are usually trained on a broad range of scientific data and capable of being applied to a range of downstream tasks, such as discovering patterns and generating scientific hypotheses, insights, and engineering designs.”

What is a foundation model?

- **Stanford**: “Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model.” https://crfm.stanford.edu/
- **Executive Order**: “AI model that is trained on broad data; generally uses self-supervision; contains at least **tens of billions** of parameters; is applicable across a wide range of contexts…” https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
- **Congressional Act**: “Artificial intelligence model that is trained on broad data; generally uses **self-supervision**; generally contains at least **1 billion** parameters; is applicable across a wide range of contexts; and exhibits, or could be easily modified to exhibit, high levels of performance at tasks that could pose a serious **risk to security**, national economic security, national public health or safety, or any combination of those matters.” https://www.congress.gov/bill/118th-congress/house-bill/6881/text

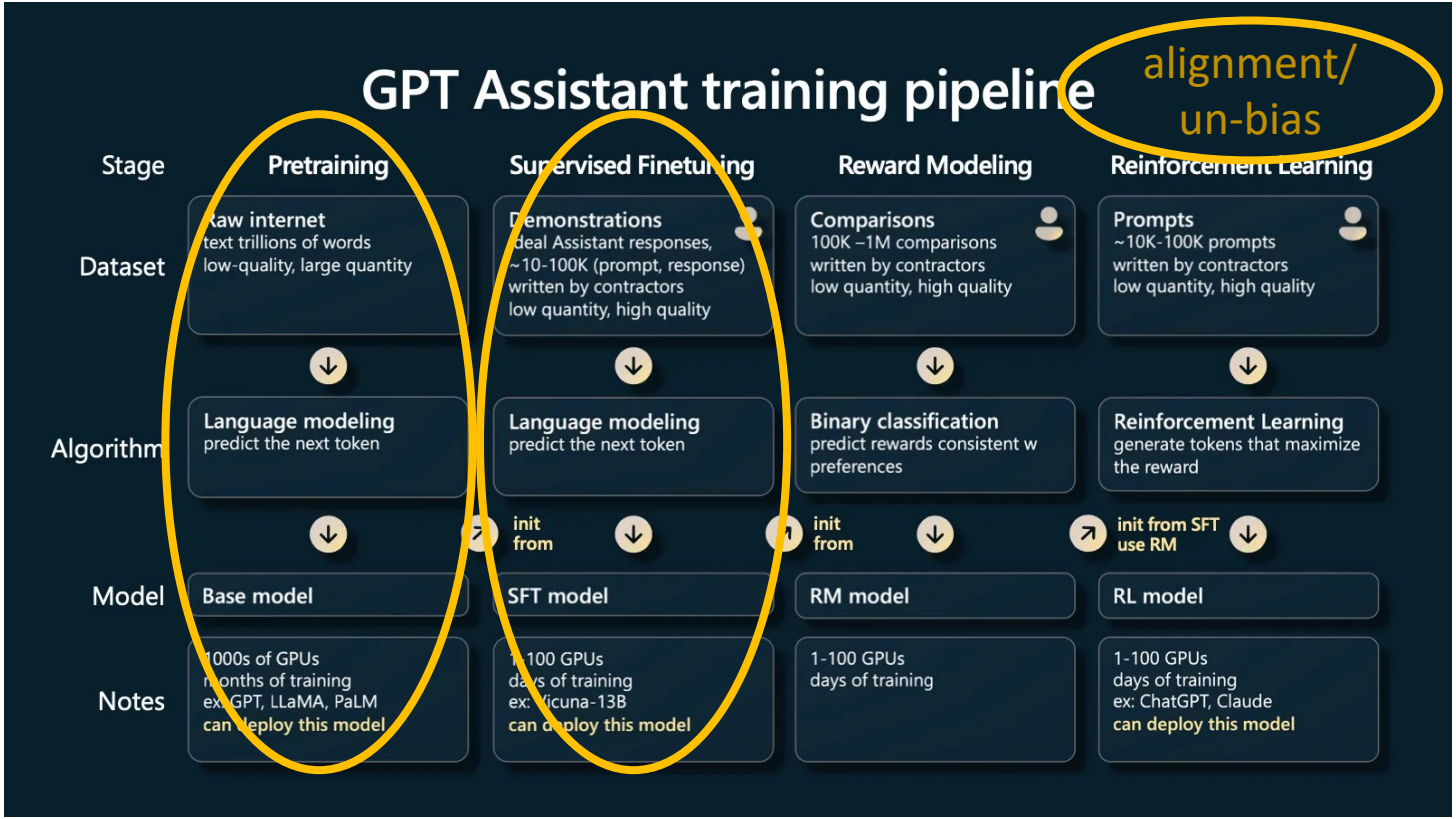
What is a foundation model?

- **Stanford**: “Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model.”

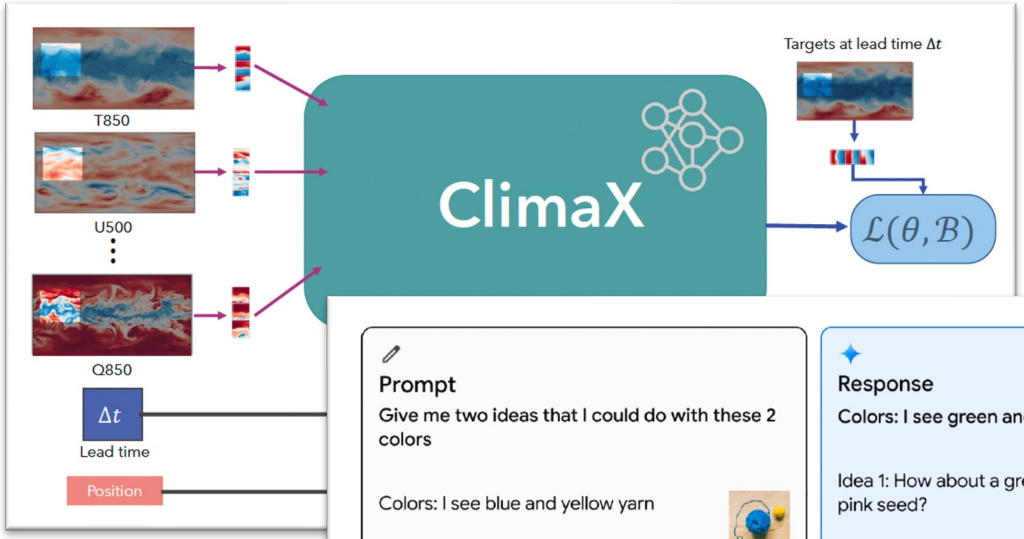


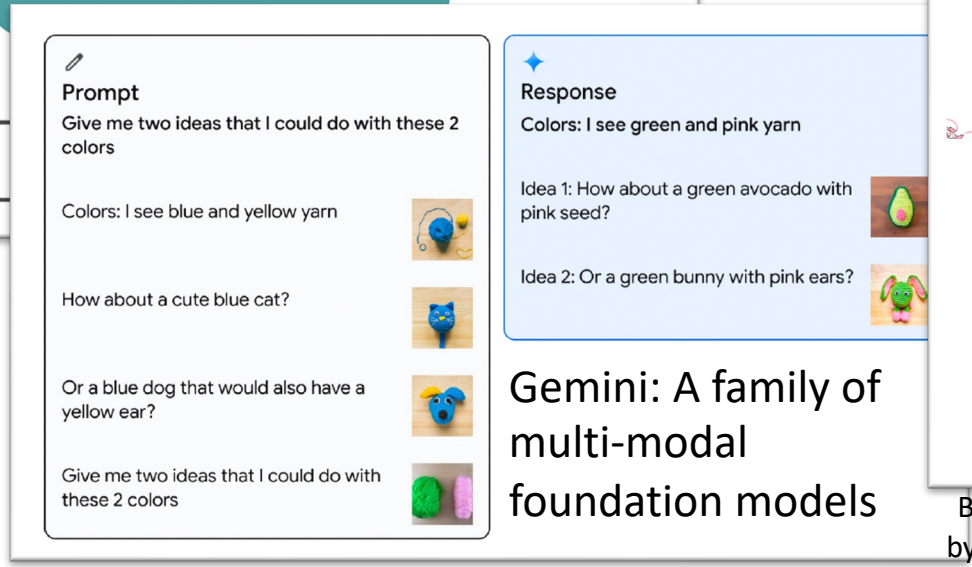
Building a Foundation Model

Karpathy's "state of GPT" https://medium.com/@chassweeting/the-state-of-gpt-by-andrew-kaparthy-fad2f007c1b9

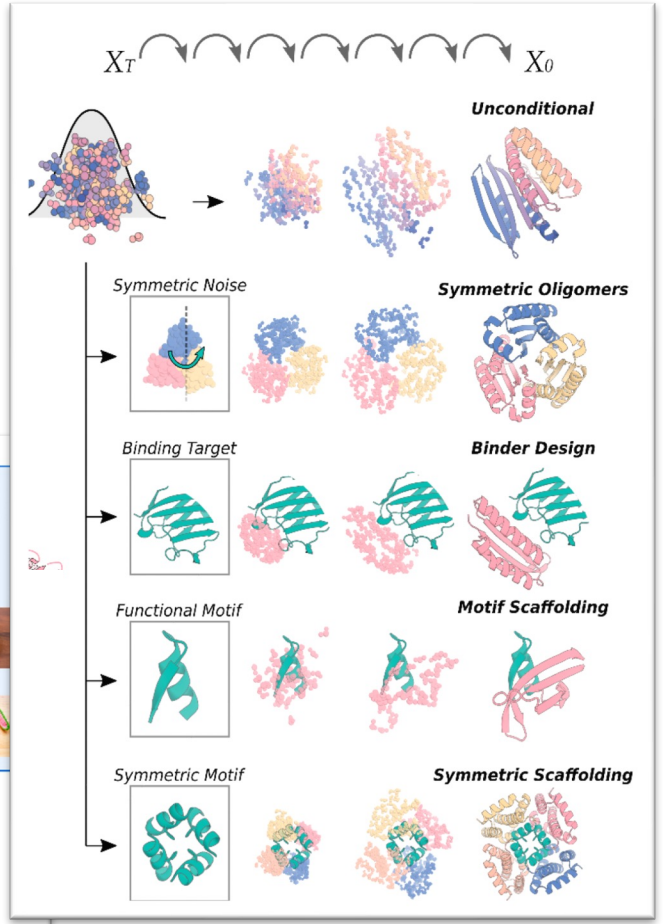


Success of foundation models





Gemini: A family of multi-modal foundation models



Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models

FM for Science (aka LSM)

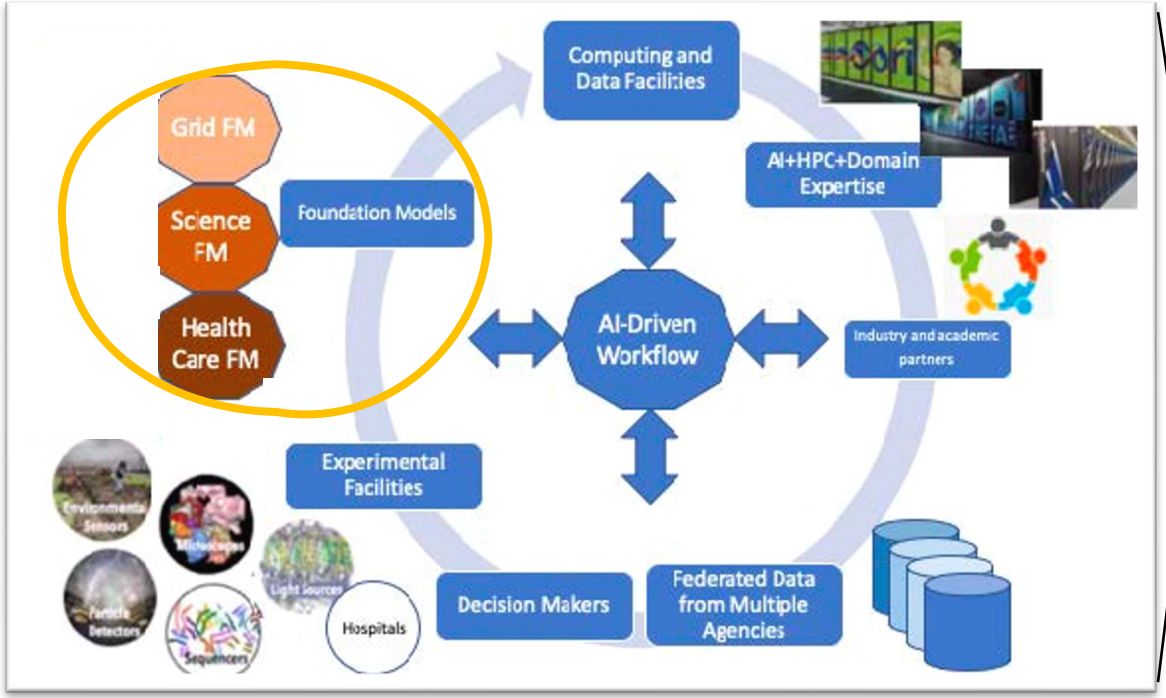
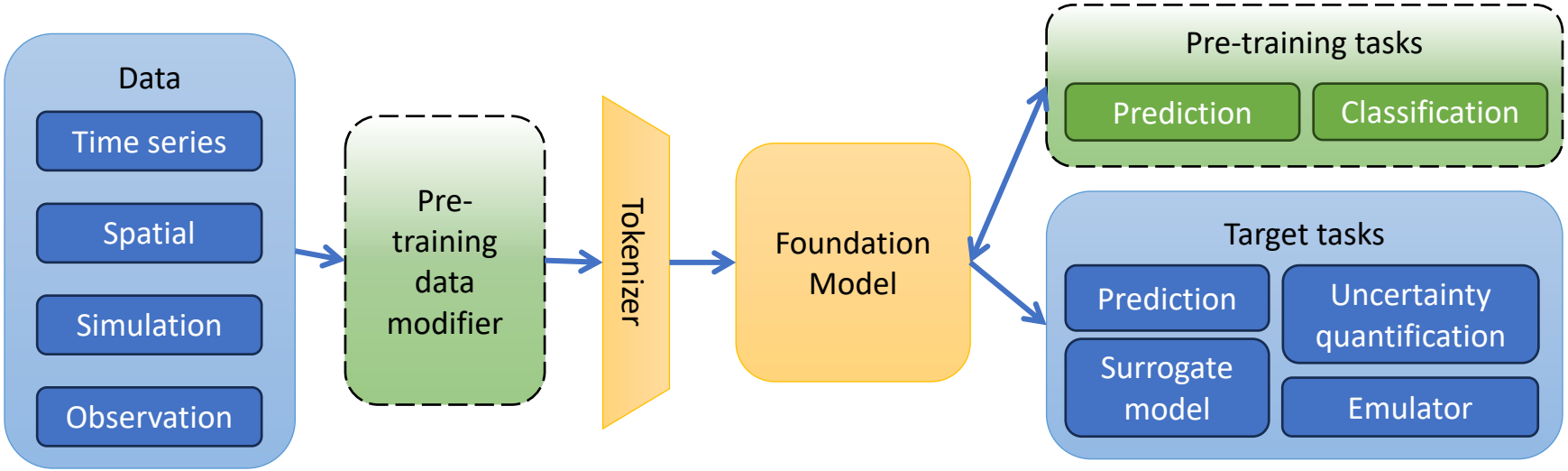


Fig 2-2, AI for SES Report

AI for Science, Energy, and Security Report. (2023). https://www.anl.gov/ai-for-science-report

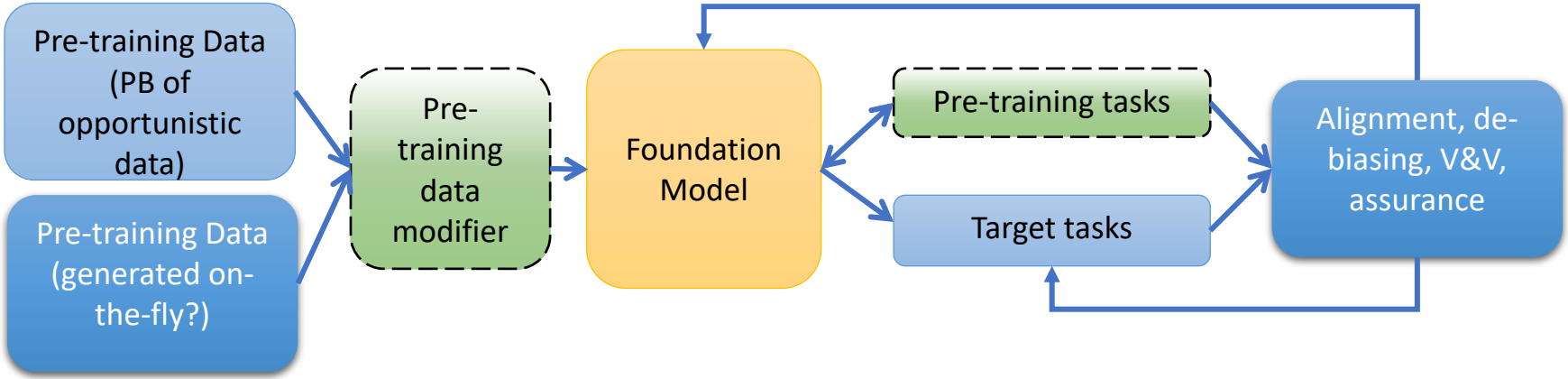
Technical Focus

- Transformer architecture can flexibly incorporate multiple modalities of data
- Tokenization of data to fit into transformer while retaining information
- Trained with self-supervision on massive, unlabeled, uncurated data



What is a Workflow in AI for Science?

- AI pre-training assume “data exists”, but science data could be produced synchronously
- Alignment post-training should be easier to quantify for LSM than for LLM – solving this issue for science should inform other uses



What Science Questions can AI Answer?

- The Scientific AI community needs the Scientific Computing community if we hope to answer this