NSF Meta-Workshop: AI to Accelerate Science and Engineering Discover (AI2ASED)

Monday 2 October 2023 - Tuesday 3 October 2023

Book of Abstracts

Contents

Introduction and Welcome / 1

NSF Leadership

Corresponding Author: mlittman@nsf.gov

NSF directors
Margaret Martonosi (Assistant Director, CISE)
Michael Littman (Division Director, CISE/IIS)
Nina Amla (Senior Science Advisor, CISE)
Christopher C. Yang (Program Director, CISE/IIS)

Introduction and Welcome / 2

Workshop organizers

Corresponding Authors: aidong@virginia.edu, schsu@uw.edu

3

Breakout 2: Science and Engineering for data-driven discovery

Room1 Edgar Lobaton (Moderator) Anuj Karpatne (Scribe) Note1

Room2 Bedrich Benes (Moderator) Wei Ding (Scribe) Note2

Room3 Phil Harris (Moderator) Paul Hanson Scribe) Note3

Lightning talk I / 4

Neural operators:AI + Science

Corresponding Author: anima@caltech.edu

Lightning talk II / 5

Lightning talks (8mins each)

Corresponding Author: ruby.leung@pnnl.gov

Lai-Yung (Ruby) Leung (PNNL)

Lightning talk III / 6

AI + Computation: Use inspired challenges from Manufacturing and Design

Corresponding Author: baskarg@iastate.edu

Baskar Ganapathysubramanian (Iowa State)

Report writing breakout / 7

Writing breakout sessions

Room1 AI-advanced Science & Science-informed AI: Wei Ding (Moderator)
Room2 LLM and Continuous ML: Joshua Agar (Moderator)
Room3 Explainable and Robust AI: Anuj Karpatne (Moderator)
Room4 Education & Outreach, Community and Cyberinfrastructure: Paul Hanson (Moderator)

Panel session / 8

Synergies between Accelerating Computer-Enabled Discovery topics

Corresponding Authors: balajir@colorado.edu, schsu@uw.edu, omar@oden.utexas.edu, mhong@umn.edu, hearst@berkeley.edu, aidong@virginia.edu

Moderator: Jennifer Dy (NEU)
HCI: Marti Hearst (Berkeley)
Data, AI and Machine Learning: Aidong Zhang (UVA), Shih-Chieh Hsu (UW)
Digital Twins: Omar Ghattas (UTexas)
Smart Sensing and Analytics: Mingyi Hong (UMichigan)
Rigorous & Reproducible Reasoning: Rajagopalan Balaji (Colorado)
Programmable/Self-Driving Labs: TBC

Breakout 1 / 9

Breakout1: AI for data-driven discovery

Room1 Eric Toberer (Moderator) Anuj Karpatne (Scribe) note1

Room2 Xinghua Mindy Shi (Moderator) Wei Ding (Scribe) note2

Room3 Jianwu Wang (Moderator) Paul Hanson (Scribe) note3

Breakout III / 10

Breakout 3: Science and Engineering for data-driven discovery

Each room picks 2 or 3 topics to discuss Barrier, Challenge, Opportunities and Recommendations.

Room1 AI-advanced Science & Science-informed AI: Xia Ning (Moderator) Wei Ding (Scribe) note1

Room2 LLM and Continuous ML: Jing Gao (Moderator) Joshua Agar (Scribe) note2

Room3 Explainable and Robust AI: Phil Harris (MIT) Anuj Karpatne (Scribe) note3

Room4 Education & Outreach, Community and Cyberinfrastructure: Nirav Merchant (Moderator) Paul Hanson (Scribe) note4

Closing / 11

Closing

Corresponding Author: ccyang@nsf.gov

Lightning talk I / 12

AI for Science in Quantum, Atomistic, and Continuum Systems

Author: Shuiwang Ji[1]

[1] *Texas A&M*

Corresponding Author: sji@tamu.edu

In this talk, I will provide an overview of research on developing AI methods to understand the natural world from the subatomic (wavefunctions and electron density), atomic (molecules, proteins, materials, and interactions), to macro (fluids, climate) scales. My talk will focus on how to capture symmetries in physical systems using equivariant models. I will also touch on a few other technical challenges, including explainability, out-of-distribution generalization, and knowledge transfer with foundation and large language models. My talk will be a summary of our recent review paper on AI for science available at https://arxiv.org/abs/2307.08423

Lightning talk III / 13

The Annual Accelerate Conference

Author: Brandon Sutherland[1]

[1] *Acceleration Consortium*

Corresponding Author: brandon.sutherland@utoronto.ca

In this short lightning talk I will discuss the Acceleration Consortium's annual Accelerate conference, which we ran in 2022 and 2023 in Toronto and are in the early stages of planning 2024 in a different host city. Accelerate spans the entire field of accelerated discovery with AI and automation: computational tools, high-throughput and autonomous experimentation, the ethics of accelerated discovery, and commercialization potential.

Lightning talk II / 14

Cosmology and Astrophysics with MachinE Learning Simulations (CAMELS)

Author: Daniel Angles-Alcazar[1]

[1] *University of Connecticut*

Corresponding Author: angles-alcazar@uconn.edu

The Cosmology and Astrophysics with MachinE Learning Simulations (CAMELS) project aims to overcome major obstacles limiting our understanding of the fundamental properties of the Universe by (1) providing thousands of state-of-the-art hydrodynamic simulations of cosmological structure formation covering a broad range of sub-grid models for the physics of galaxy formation and (2) developing novel machine learning algorithms to maximize the extraction of information from cosmological surveys while marginalizing over uncertainties in galaxy formation physics. In this lightning talk, I will summarize the CAMELS workshop hosted at the Simons Foundation in the Fall 2022, bringing together a growing community of scientists leveraging the CAMELS Public Data Repository to discuss recent progress, challenges, and future directions.

Lightning talk III / 15

NSF-NIH Joint Workshop on Emerging AI in Biology

Author: Carl Kingsford[1]

[1] *Carnegie Mellon University*

Corresponding Author: carlk@cs.cmu.edu

New techniques in AI are rapidly being developed, extended and applied to challenging problems in biology. At the same time, as new assays, new data efforts, and greater understanding is developed in biology, the class and scope of problems that are amendable to AI approaches is growing. In order to survey the current frontier of the interface between AI methodology and biology and to chart future directions and challenges, we held an "NSF-NIH Joint Workshop on Emerging AI in Biology"in June 2023 that gathered approximately 40 experts on the intersection of research in AI and biology. I will present some of the insights and discussion from this workshop. Topics include challenges related to biological applications in the following areas: federated learning; privacy, security and fairness; transfer learning; automated science and active learning; explainability and causality; and scalability.

Lightning talk I / 16

CancerGPT: Few-shot Drug Pair Synergy Prediction using Large Pre-trained Language Models

Author: Ying Ding[1]

[1] *The University of Texas at Austin*

Corresponding Author: yd4956@eid.utexas.edu

Abstract: Large pre-trained language models (LLMs) have been shown to have significant potential in few-shot learning across various fields, even with minimal training data. However, their ability to generalize to unseen tasks in more complex fields, such as biology, has yet to be fully evaluated. LLMs can offer a promising alternative approach for biological inference, particularly in cases where structured data and sample size are limited, by extracting prior knowledge from text corpora. Our proposed few-shot learning approach uses LLMs to predict the synergy of drug pairs in rare tissues that lack structured data and features. Our experiments, which involved seven rare tissues from different cancer types, demonstrated that the LLM-based prediction model achieved significant accuracy with very few or zero samples. This talk highlights several research efforts to tackle drug pair synergy prediction in rare tissues with limited data.

17

NSF leadership

Corresponding Author: dmanders@nsf.gov

Lightning talk III / 18

Pandemic Research for Preparedness and Resilience

Author: Madhav V Marathe[1]

[1] *University of Virginia*

Corresponding Author: mvm7hz@virginia.edu

A Research Roadmap for the Next Pandemic PREPARE (Pandemic Research for Preparedness and Resilience) is an NSF CISE-sponsored virtual organization tasked with fostering research collaborations and synthesizing critical pandemic-related computing research into a roadmap to help inform NSF funding opportunities that will aid our nation's effective response to the next pandemic. Since we started this project in October 2020, we have hosted eight virtual workshops featuring 72 subject-matter experts as speakers, panelists, and committee members. Collectively, these sessions were attended by over 2000 researchers and viewed on YouTube more than 3800 times. Please see prepare-vo.org1 for more details.

Through the aforementioned workshops, plus conversations with community members, podcast interviews, and literature review, we have gathered a good deal of information which we have synthesized as input into a set of recommendations meant to advise NSF leadership as they determine funding for programs that will help our world prepare to take on the next pandemic. This work represents input from a multidisciplinary assemblage of international researchers, and recommendations are offered in the following areas: Importance of Multidisciplinary Collaborations and Industry-Academia-Government (IAG) Partnerships; Cyberinfrastructure, Data, Data Analysis, and Responsible AI and Tools; and Societal Impacts.

I will briefly summarize the recommendations from the report with a specific focus on role of AI and Data Science in Pandemic response.

Lightning talk II / 19

Foundation Models for Science: What happens when you train large (language) models for Science?

Author: Shirley Ho[1]

Co-author: Mike McCabe [1]

[1] *Flatiron Institute*

Corresponding Authors: sho@flatironinstitute.org, mmccabe@flatironinstitute.org

In recent years, the fields of natural language processing and computer vision have been revolutionized by the success of large models pretrained with task-agnostic objectives on massive, diverse datasets. This has, in part, been driven by the use of self-supervised pretraining methods which allow models to utilize far more training data than would be accessible with supervised training. These so-called "foundation models" have enabled transfer learning on entirely new scales. Despite their task-agnostic pretraining, the features they extract have been leveraged as a basis for task-specific finetuning, outperforming supervised training alone across numerous problems especially for transfer to settings that are insufficiently data-rich to train large models from scratch. In this talk, I will show our preliminary results on applying this approach to a variety of scientific problems and speculate what are possible future directions.

Lightning talk I / 20

Fokker-Planck Inverse Reinforcement Learning: A physics-constrained approach to Markov Decision Process models of cell dynamics

Authors: Chengyang Huang[1]; Krishna Garikipati[1]; Siddhartha Srivastava[1]; Xun Huan[1]

[1] *University of Michigan*

Corresponding Author: krishna@umich.edu

In this short talk I will discuss our recent work on an approach to introducing connections between the Fokker-Planck equation and learning algorithms for dynamical systems that follow Markov Decision Processes

Lightning talk II / 21

Using domain-aware metrics for deploying AI/ML in weather/climate applications

Author: Peetak Mitra[1]

[1] *Excarta*

Corresponding Author: peetak@excarta.io

Conventional AI/ML metrics (such as RMSE) for optimization often do not translate well for weather/climate-specific applications including for energy grid management, or modeling key physical prognostics that are driven by an underlying dynamical process. In this short talk, we will explore the importance of using domain-aware metrics for model training, post-training evaluation and eventual deployment-in-the-wild for climate-specific AI/ML software. Some of these learnings were aggregated from organizing the *Tackling Climate Change with Machine Learning* workshop at NeurIPS 2022 and from leading various technical projects in the industry across the TRL landscape in the AI/ML application to weather/climate.

Lightning talk II / 22

From Harnessing the Data Revolution to Harvesting the Data Revolution

Author: Philip Coleman Harris[1]

[1] *Massachusetts Inst. of Technology (US)*

Corresponding Author: philip.coleman.harris@cern.ch

Developments in modern computation and instrumentation have led to the possibility of recording enormous amounts of data, the data revolution. Along with this incredible data flow, a new demand has emerged for algorithms that can run on all this data to "Harness the Data Revolution." Large datasets are rapidly encompassing many scientific domains, including high-energy physics, Astronomy, Neuroscience, Genomics, Materials Science, Biology, Climate science, Materials science, among others. The use of parallel processing strategies, coupled with deep learning, placed within modern cyberinfrastructure has emerged as a solution to handle the data revolution. However, new developments in AI algorithms and an educated workforce are needed to achieve state-of-the-art algorithms. This talk presents a list of emerging solutions and strategies towards algorithms and approaches that allow us to handle this data. Ultimately we can go from harnessing the data revolution to harvesting the data revolution.

Lightning talk II / **23**

Accelerating AI Applications in Environmental Sciences

Authors: Eric Khin[1]; Rob Redmon[2]; Yuhan "Douglas" Rao[3]

[1] *NOAA National Centers for Environmental Information*

[2] *NOAA Center for AI*

[3] *Cooperative Institute for Satellite Earth System Studies/NOAA National Centers for Environmental Information*

Corresponding Authors: rob.redmon@noaa.gov, eric.a.khin@noaa.gov, yrao5@ncsu.edu

In this lightning talk, we will provide an overview of NOAA Center for AI's approach to foster an open community discussion that gather members from academic researchers, industry leaders, and government researchers and managers around the topics of AI development in environmental sciences. Since 2022, the annual NOAA AI workshop transitioned into an open community forum where all interested members in the community will come together to set the research agenda for AI in environmental sciences. This year's NOAA AI Workshop centered around two themes - building benchmarking frameworks for AI R&D and facilitating research-to-applications transition for AI in environmental sciences. The community identified the key challenges in AI-ready data, cyber-/social-infrastructures, and workforce development that need to be addressed to fully embrace the potential of AI in environmental sciences.

Lightning talk I / **24**

AI Enabled Scientific Revolution

Author: Vipin Kumar[1]

[1] *University of Minnesota*

Corresponding Author: kumar001@umn.edu

There is an increasing consensus in the wider scientific community that AI is poised to disrupt science by unlocking entirely new approaches, driving new scientific inquiry, and enabling greater

scientific leaps with far-reaching societal consequences. In addition, challenges unique to scientific problems offer an opportunity to dramatically advance AI. However, there are substantial barriers that are faced by AI in the context of science, and addressing these barriers will require support for advances in AI that are driven by the unique needs of scientific problems. Workshop on AI Enabled Scientific Revolution was held at NSF in February 2023 to discuss a new frontier in AI that could revolutionize the traditional discovery process across multiple scientific disciplines. This in-person workshop was attended by 28 researchers spanning all aspects of AI (including ML, robotics, computer vision, and NLP) as well as researchers who had extensive experience at the intersection of AI and one or more scientific applications, including environmental sciences (e.g., climate, hydrology), materials science, high energy physics, astrophysics, chemistry, and biomedical sciences. Attendees were from academia, industry, and philanthropic organizations, as well as NSF and other government agencies. My talk will provide a summary of the wide ranging discussions at the workshop as well as concrete recommendations to incentivize the development of next-generation AI and its adoption in scientific practice that will dramatically accelerate scientific discovery across a range of domains.

Reference: workshop report

Lightning talk III / 25

The Frontiers of Artificial Intelligence-Empowered Methods and Solutions to Urban Transportation Challenges

Authors: Lili Du[1]; Yinhai Wang[2]

[1] *University of Florida*

[2] *University of Washington*

Corresponding Authors: yinhai@uw.edu, lilidu@ufl.edu

With the quickly growing quantity and variety of transportation data, Artificial intelligence (AI) technologies are revolutionizing transportation research from system management to automated vehicle and infrastructure control. Emerging AI technologies combined with other analytical methods will lead to improved scientific understandings, transformative methods, and innovative, proactive management solutions for urban transportation infrastructure systems (UTIS). To explore the frontiers of AI-empowered methods, solutions, best practices, and workforce development for addressing urban transportation challenges, we held a two-phase workshop on June 4-5, 2022, in Seattle, WA, and on December 15, 2022, in Gainesville, FL, respectively. The workshop gathered researchers from relevant disciplines, industry experts, policymakers, educators, and workforce developers, fostering a collaborative environment for comprehensive discussions and exchanges. This presentation will share key findings of the workshop, including research opportunities, application-ready technologies, limitations, emerging implementation, workforce development, and education needs, to further stimulate transformative research in pertinent communities.

Lightning talk I / 26

Unified Knowledge Representation for Science

Author: Wei Wang[1]

[1] *UCLA*

Corresponding Author: weiwang@cs.ucla.edu

The vast amount of knowledge accumulated in various science disciplines has been traditionally maintained in a way that is difficult for AI systems to use, due to differences in formats, standards,

and types. This makes it challenging to integrate and share knowledge across different domains and to use it to build intelligent systems. To address these challenges, there is a pressing need to develop AI/ML models that can automatically train foundational models for knowledge representation. These models should be able to extract and integrate knowledge from multiple sources, in different formats and types, and be able to update themselves incrementally as new knowledge becomes available. This will require developing advanced algorithms that can handle uncertainty, ambiguity, and variability, and that can learn to generalize from specific examples to more abstract concepts and categories.

In addition, research is needed on how to utilize these pre-trained knowledge models in building AI systems for science adventure. This requires developing new methods for reasoning, inference, and decision-making that can leverage the knowledge in the models to solve complex problems and make informed decisions. It also requires developing user interfaces and visualization tools that can enable scientists and engineers to interact with the knowledge models in a natural and intuitive way, and to explore and analyze the knowledge in different ways.

In summary, developing AI/ML models for knowledge representation and utilizing them in building intelligent systems for science adventure is a challenging but important research direction that has the potential to transform the way we discover, understand, and apply knowledge across different domains.

Lightning talk II / 27

Workshop on Machine Learning and Artificial Intelligence to Advance Earth System Science: Opportunities and Challenges

Author: L. Ruby Leung[1]

[1] *Pacific Northwest National Laboratory*

Corresponding Author: ruby.leung@pnnl.gov

This presentation briefly summarizes a workshop convened by the National Academies of Sciences, Engineering, and Medicine on February 7, 10, and 11, 2022, on the opportunities and challenges of using ML/AI to advance Earth system science, including their ethical development and use. The workshop explored how ML/AI approaches can contribute to improving understanding, analysis, modeling, prediction, and decision making. The 3 days of the workshop were organized around 3 broad themes: (1) Emerging approaches for using, interpreting, and integrating ML/AI for Earth system science; (2) Challenges and risks of using ML/AI for Earth system science; and (3) Future opportunities to accelerate progress.

Breakout report I / 28

Room1

Corresponding Author: etoberer@mines.edu

Breakout report I / 29

Room2

Corresponding Author: mindyshi@temple.edu

Breakout report I / 30

Room3

Corresponding Author: jianwu@umbc.edu

Breakout report II / 31

Room1

Corresponding Author: edgar.lobaton@ncsu.edu

Breakout report II / 32

Room2

Corresponding Author: bbenes@purdue.edu

Breakout report II / 33

Room3

Corresponding Author: philip.coleman.harris@cern.ch

Breakout report III / 34

Room2

Corresponding Author: jinggao@udel.edu

Breakout report III / 35

Room1

Corresponding Author: ning.104@osu.edu

Breakout report III / 36

Room3

Corresponding Author: nirav@arizona.edu

37

ACED workshop

Corresponding Author: ccyang@nsf.gov

Breakout report III / 38

Room3

Corresponding Author: philip.coleman.harris@cern.ch

Breakout III / 39

Breakout3 topics introduction

Corresponding Author: aidong@virginia.edu