Networking and Information Technology
Research and Development

***The government seeks individual input; attendees/participants may provide individual advice only*.**

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes
June 7, 2023

Virtual Meeting

Participants

Alejandro Suarez (NSF)	Mallory Hinks (NCO)
Daniel Bullock (NSF)	Manish Parashar
Ewa Deelman (University of Southern California)	Miron Livny (Wisconsin)
Florence Hudson	Shantenu Jha (Rutgers University)
Glenn Lockwood (Microsoft)	Svitlana Volkova (Aptima)
H Birali Runesha (University of Chicago)	Tom Gibbs (NVIDIA)
Hal Finkel (DOE)	Val Anantharaj (ORNL)
Jim Basney (Illinois)	William Miller (NSF)
Kevin Thompson (NSF)	

Introductions: This meeting was chaired by Jay Park (NSF) and Hal Finkel (DOE SC)

Opportunities of Large Pretrained Generative Models and Embodied AI for Scientific Workflows

Svitlana Volkova (Aptima)

In this presentation, Dr. Svitlana Volkova, Chief AI Scientist at Aptima, explored the opportunities and limitations of large pre-trained generative models and embodied AI for scientific workflows. She discussed the advancements in AI, the architecture of large pre-trained generative models, and the emerging capabilities of these models. Dr. Volkova also talked about the fundamentals of in-context learning and recent science advances that can be leveraged to improve scientific workflows. She highlighted the potential of large pre-trained generative models in scientific reasoning, embodied AI, and scientific data workflows. She also emphasized the limitations of these models, particularly in grounding and planning.

During the Q&A session, Dr. Daniel Bullock asked about the differences between scientific and non-scientific reasoning for synthetic systems, and possible different epistemologies that may be used. Dr. Volkova discussed the complexity of scientific reasoning tasks and how startups like Euler AI are developing benchmarks for scientific question answering but noted that this is not the same as scientific reasoning. Dr. Miron Livny asked about what cyber infrastructure people should do differently to support these workflows. Dr. Volkova suggested investing in AI approaches specifically for science and using these models to boost productivity. She also discussed the potential of using large pre-trained models for boosting productivity and answering science-related questions. She emphasized the need to be transparent about the risks associated with developing new foundation models and investigating their limitations. In terms of infrastructure, she suggested leveraging the cloud and hybrid approaches. Dr. Livny highlighted the challenge of training and suggests having a large number of models. There was also discussion around the challenges of proprietary models and the need for more open models and infrastructure. Dr. Volkova emphasized the importance of having both HPC and cloud infrastructure for scientific tasks. She noted that existing HPC infrastructure is critical for security reasons, but alone is not sufficient for current scientific tasks. She suggested partnering with cloud providers in a way that is cost-effective while also maintaining security and leveraging a hybrid infrastructure.

<u>New scientific workloads and Why We Need More Automation</u>

Ewa Deelman, University of Southern California

Dr. Ewa Deelman gave a presentation discussing the importance of workflow automation in managing complex scientific workflows. She presented examples of workflows, such as generating a seismic hazard map of Southern California and the collaborative and adaptive sensing of the atmosphere, and how workflow automation can help manage the large number of tasks and data involved. Dr. Deelman also discussed the challenges of managing increasingly complex cyber infrastructure and the changing expectations of users. She highlighted the capabilities of the Pegasus workflow management system and the NSF-funded Pegasus Access Pilot, which allows users to create and submit workflows within Jupyter Notebook and provision resources using the HD Condo Annex tool. She also discussed the growing use of AI-based methods within workflows and the need for automation to ensure validation and replication of results.

During the Q&A session, Daniel asked about the feasibility of developing a pool of open science front end resources that can be interchangeable with backend resources. Dr. Deelman suggested striving for a standardized approach to cyber infrastructure to provide users with a better experience. Miron asked about protecting against cyber risks in large pre-trained models, and Svitlana asked for Dr. Deelman's thoughts on being proactive against such risks. The group discussed the challenge of proactively protecting against cybersecurity risks and the idea of using adversarial models or auditing to detect and prevent them. Miron emphasized the importance of reliable auditing for establishing trust and protecting restricted data. Hal asked about integrating edge-based data sources into these systems. Dr. Deelman discussed the traditional approach of using lightweight condor containers to execute jobs at the edge and relying on robust cyber infrastructure tools like HT Condor, which her group has been using since its inception in 2001. She emphasized the importance of building on existing tools and the benefits of the principles they have developed, such as separating work for description and execution, allowing the same workflow to run on different target resources without changing the description. She acknowledged the need for more work on dealing with network issues and resource constraints when dealing with edge.

The group discussed the challenge of preparing for next-generation cybersecurity threats in open science, particularly related to AI and edge computing. Dr. Deelman acknowledged the need for more work in dealing with edge flavors, and Svitlana asked for her thoughts on preparing for future threats. Ewa admitted the difficulty in predicting what the next threats will be and suggested potentially going back technologically to better understand the implications of new and bigger models.

Val Anantharaj from Oak Ridge National Laboratory discussed the exploration of foundation models for science, particularly in climate and computational sciences. He suggested that, at this exploratory stage, it is important to focus on innovation and not be held back by concerns about cyber issues. Val also noted the rapid pace of innovation in this field, with an innovation cycle of around three months. He discussed the challenges of keeping up with the rapid pace of innovation in foundation models for science. He noted that the innovation cycle is about three months and suggested that multiple sprints may be necessary to refine explorations and come to intermediate conclusions. Tom Gibbs from Nvidia offered to help with running models faster on a larger scale. Val emphasized the critical size of the number of parameters and network required for reliable performance and the need for continuous training. Svitlana Volkova agreed on the importance of keeping up with innovation and noted that labs have an advantage in terms of data and expertise.

Svitlana emphasized the importance of domain expertise and data in driving innovation in foundation models for science, noting that startups cannot solve the climate crisis alone. Val discussed the challenges of scaling up modalities and the differences in approach compared to classical machine learning. Tom Gibbs shared his experience working on foundation models for genomic and proteomic prediction and notes the challenge of dealing with latency on the edge.

Next Meeting July 12, 2023