



2023 RESEARCH INFRASTRUCTURE WORKSHOP

The Expanding use of AI in Research Infrastructure Applications

Moderator: Charles F. Vardeman II
University of Notre Dame, CI-Compass



Agenda (95 min 3:25 pm - 5:00 pm)

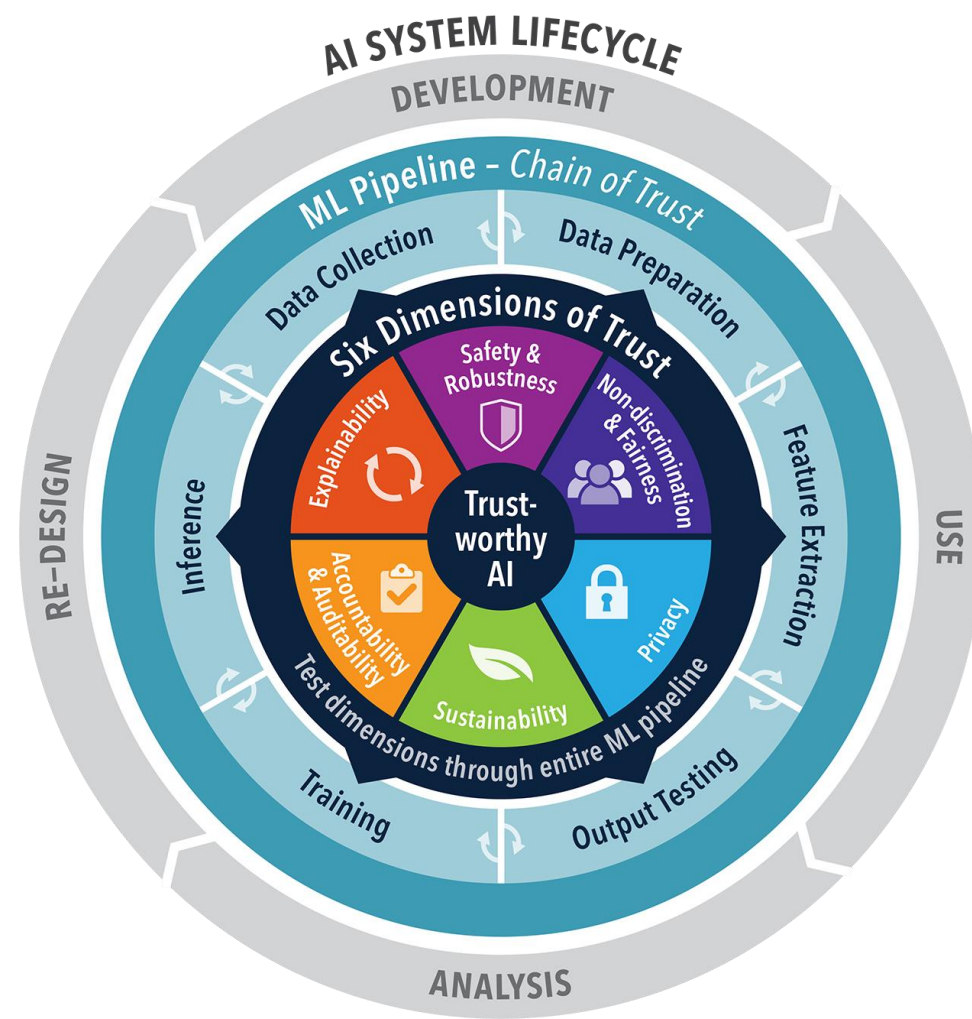
- (35 min) **Panelist Introductions:** Each panelist will start with a brief 5-minute presentation on their AI research interests and their views on the intersection of AI and Research Infrastructure.
- (40 min) **Panel Discussion:** We will proceed with a moderated discussion based on predetermined questions exploring the various facets of AI's role in the research infrastructure ecosystem.
- (20 min) **Audience Q&A:** In the last 20 minutes, we'll open up the floor to the audience for their questions and insights.

Note: We encourage lively, respectful conversation and look forward to everyone's contributions to this important discussion.

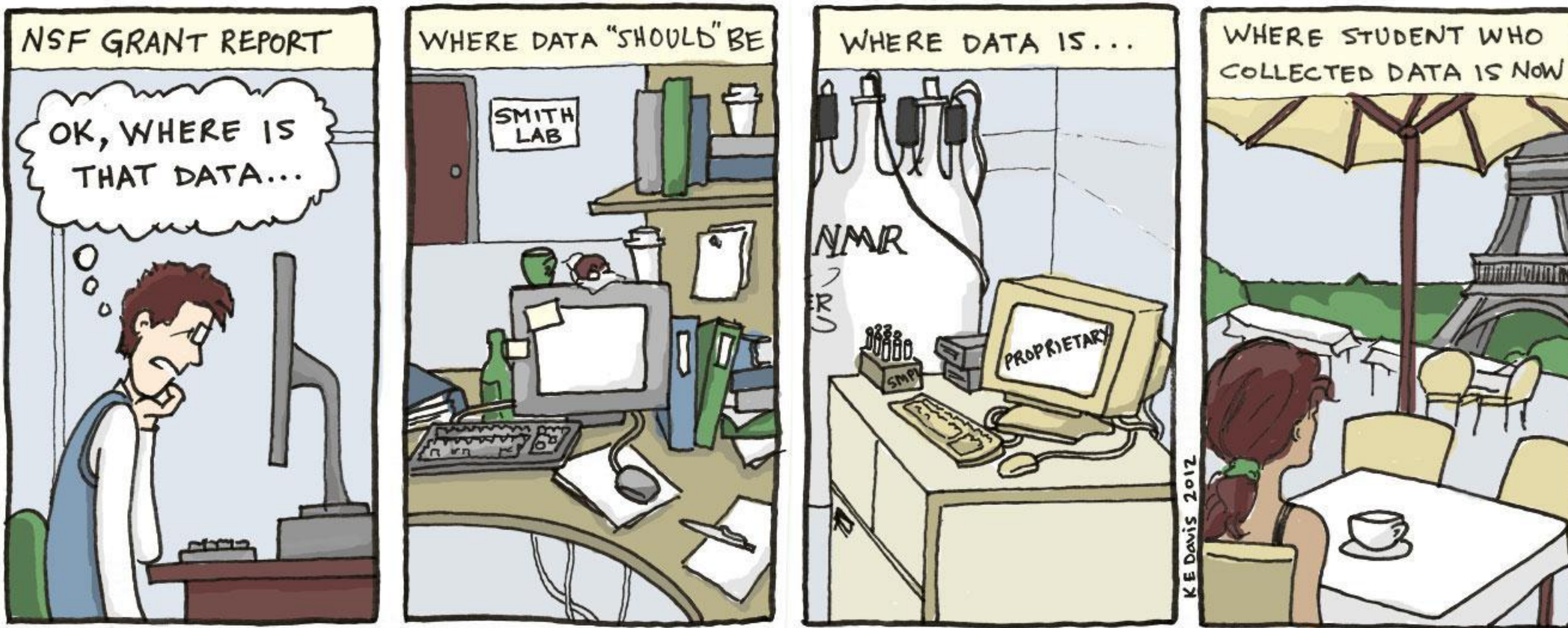
Panel Moderator: Charles F. Vardeman II

TIMELINE:

- PhD: UND Theoretical Chemistry - Co-Creator of OpenMD Molecular Dynamics Simulation Engine
- Computational Scientist, Research Assistant Professor, Notre Dame Center for Research Computing.
- NSF Data and Software Preservation for Open Science (DASPOS)
- Early prototypes based on Ontology Design Pattern, Vocamp methodology using Domain Expert for Modeling. Precursor to “Open Knowledge Network (OKN)” effort.
- CI-CoE Pilot. Use of “structured data” using schema.org markup to enhance dataset discovery. Creation of Earth Science Information Partners (ESIP) Science on schema.org cluster.
- CI-Compass. FAIR Data and AI Enablement
- PI Trusted AI Frameworks: US Navy/DoD effort to establish workforce development and AI Engineering research. IU/IUPUI/Notre Dame/Purdue (SCALE).

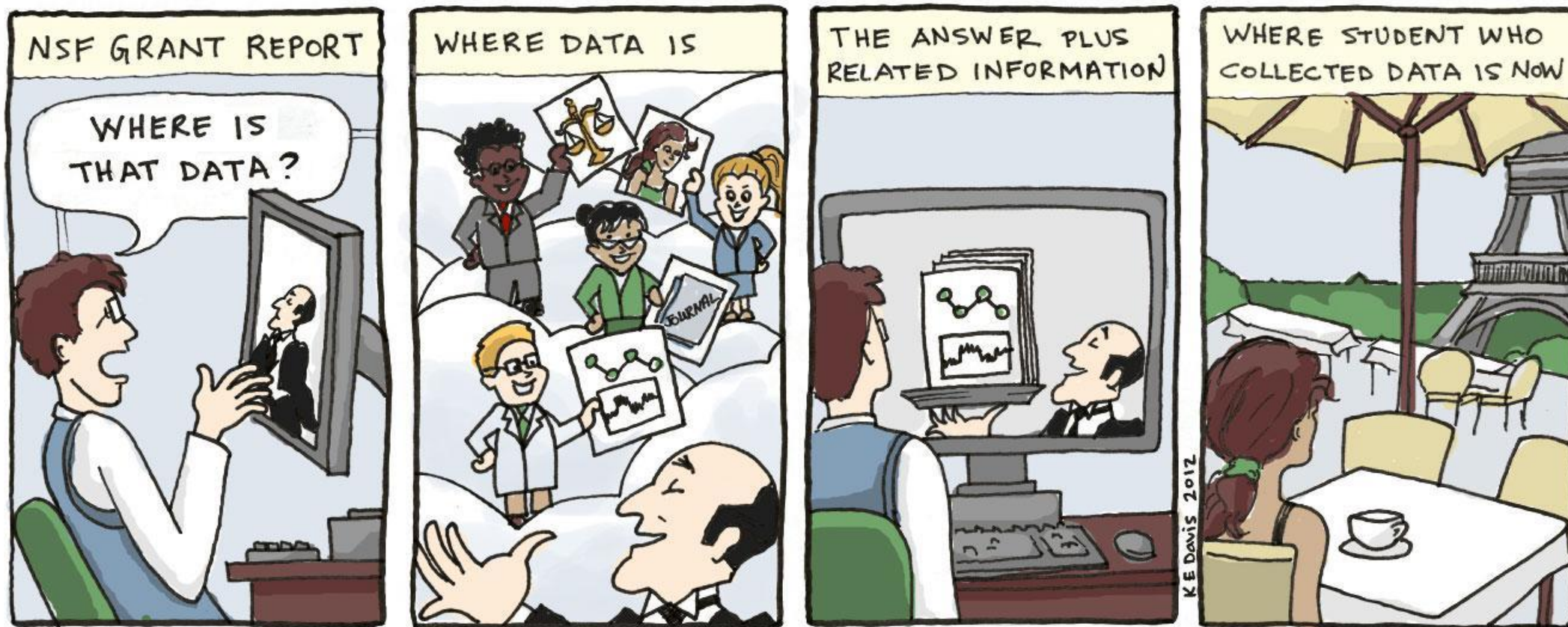


A decade ago...



CREDIT: Kristina Davis (CRC, UND)

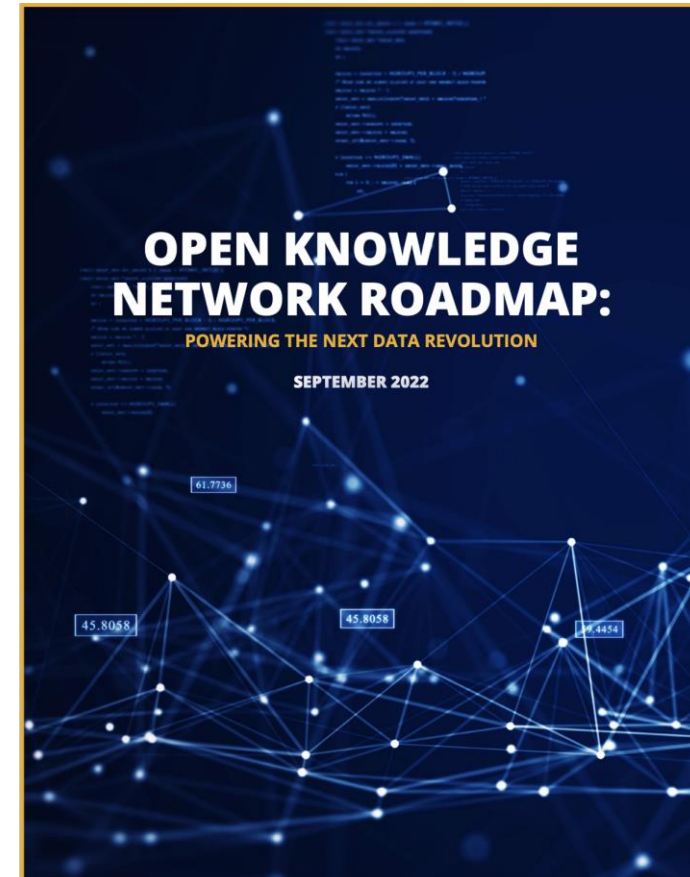
A vision for the future from a decade ago...

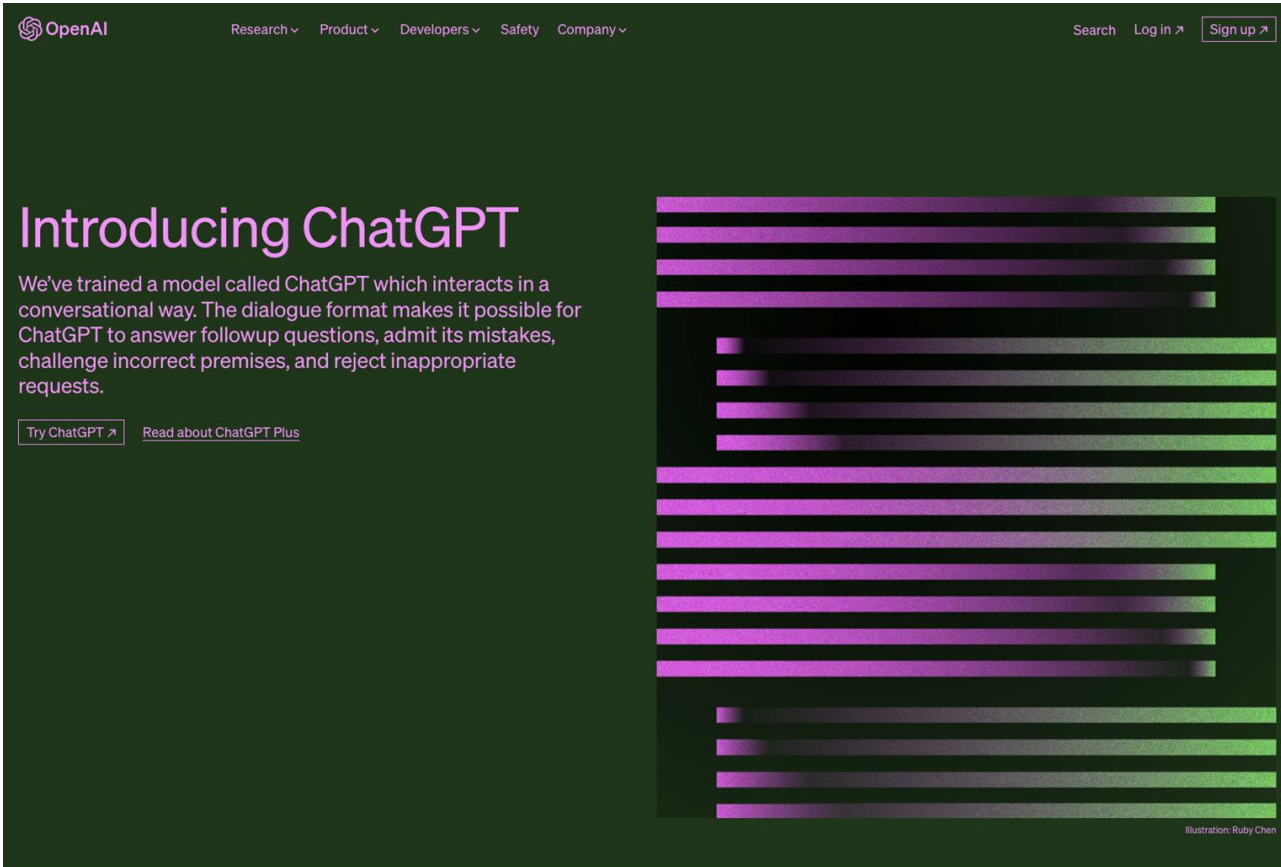


CREDIT: Kristina Davis (CRC, UND)

Open Knowledge Networks (OKN)

“The OKN represents a best-in-class opportunity to provide FAIR access to open data, to enable AI and ML tools and ecosystems, and to leverage data and information needed to address societal challenges and innovation opportunities.”





The screenshot shows the OpenAI website's announcement for ChatGPT. The header includes the OpenAI logo and navigation links for Research, Product, Developers, Safety, and Company. There are also links for Search, Log In, and Sign up. The main content area features the title "Introducing ChatGPT" in a large, light blue font. Below the title is a paragraph of text describing the model's conversational capabilities. Two buttons are visible: "Try ChatGPT" and "Read about ChatGPT Plus". The background of the announcement is dark with abstract, glowing blue and purple horizontal lines. At the bottom right of the announcement, it says "Illustration: Ruby Chen".

[Introducing ChatGPT \(openai.com\)](https://openai.com)

Introduction by ChatGPT (Whova)

"Hello everyone, I'm ChatGPT, an AI developed by OpenAI. Though I can't be there in person, I'm excited to be part of this important discussion in some capacity. I was trained on a diverse range of internet text, but I don't know anything about individual users who interact with me, maintaining privacy and confidentiality.

As an AI, my purpose is to help users like you generate human-like text based on the prompts given to me. I've been used in various contexts, ranging from drafting emails, writing Python code, creating written content, to even co-authoring a book.

Today, we're here to explore my potential and that of other AI tools in contributing to the future of research infrastructure. This is an important topic, as AI continues to revolutionize numerous fields, and scientific research is no exception.

AI tools like me could assist researchers in a number of ways, from automating routine tasks, aiding decision-making processes, to even predicting trends based on data analysis. We also can play a role in managing and maintaining the research infrastructure itself.

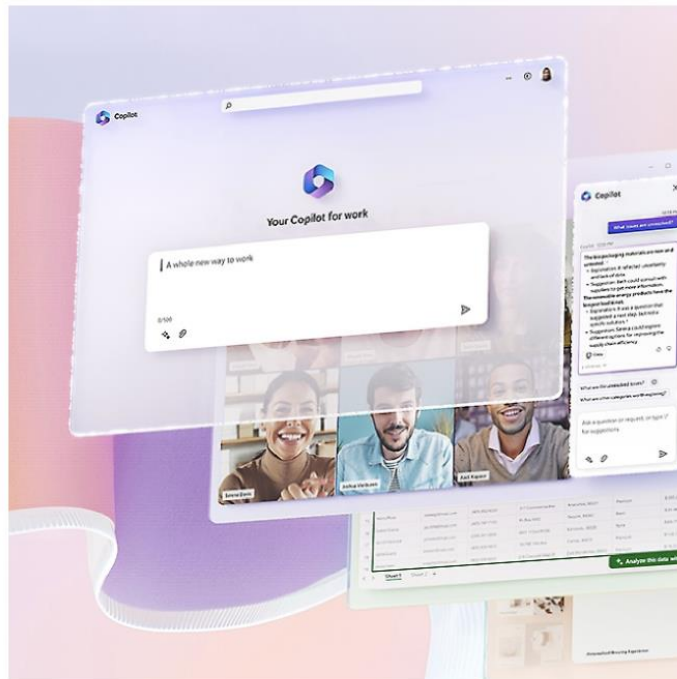
Remember, as an AI, I don't have feelings or beliefs. But if I did, I'd be excited about the possibilities we're discussing today. I hope you find this conversation stimulating and informative. Let's explore the future of AI and research infrastructure together!"

AI as “Co-Pilots”

Introducing Microsoft 365 Copilot

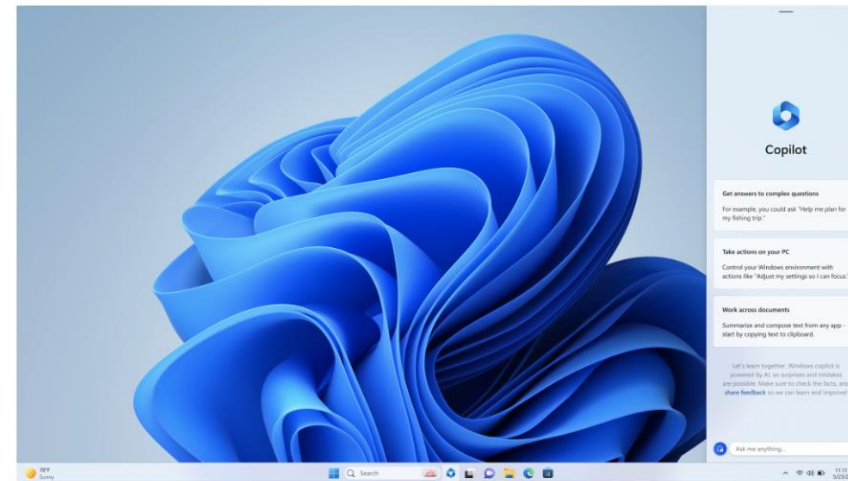
The new Microsoft 365 Copilot experience works alongside you, embedded in the apps you use every day—Word, Excel, PowerPoint, Outlook, Teams, and more. It combines the power of language models with your business data and context—including all your Microsoft 365 apps, documents, and conversations.

[Learn more >](#) [▶ Watch the video](#)



[Microsoft 365 - Subscription for Office Apps | Microsoft 365](#)

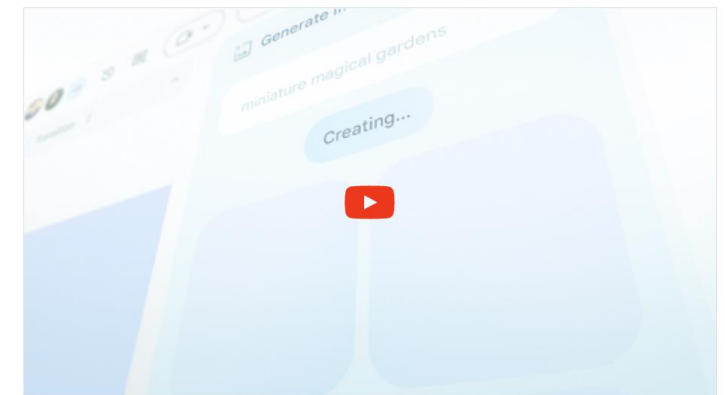
We're thrilled to introduce Windows Copilot. Windows is the first PC platform to provide centralized AI assistance for customers. Together, with Bing Chat and first- and third-party plugins, you can focus on bringing your ideas to life, completing complex projects and collaborating instead of spending energy finding, launching and working across multiple applications.



[Bringing the power of AI to Windows 11 - unlocking a new era of productivity for customers and developers with Windows Copilot and Dev Home - Windows Developer Blog](#)

See how we're making Workspace even more helpful with Duet AI

Workspace is harnessing the power of generative AI to unlock new ways of working so people can create, connect and grow together. Explore what's next.



Now, writing is easier with the help of AI in Workspace

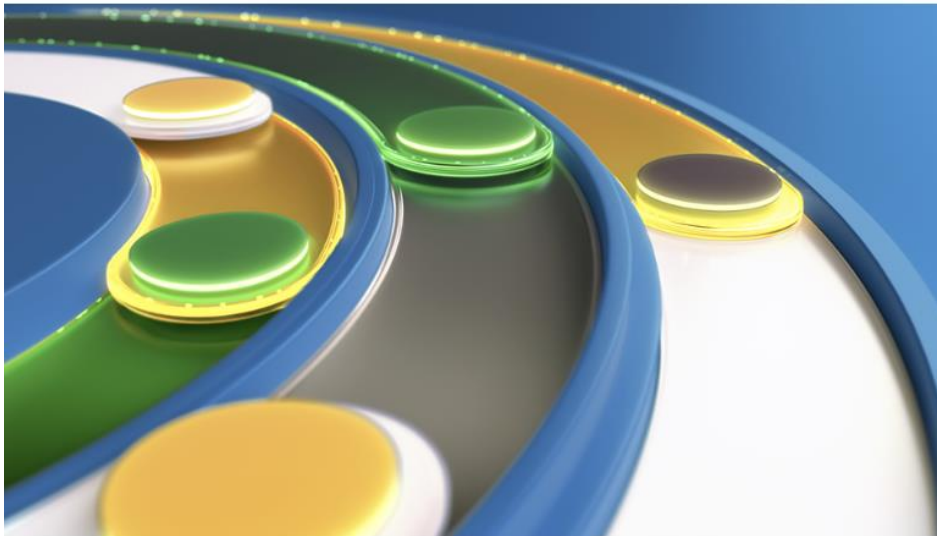
Blank pages can stump the best of us, which is why we are embedding generative AI tools in Docs and Gmail to help you get started. Simply type in a topic, and have Workspace help you go from blank page to first draft.

[Generative AI Tools for Better Productivity | Google Workspace](#)

Co-Pilots in Research Infrastructure Management

Introducing Microsoft Security Copilot:
Empowering defenders at the speed of AI

Mar 28, 2023 | Vasu Jakkal - Corporate Vice President, Security, Compliance, Identity, and Management





[Introducing Microsoft Security Copilot: Empowering defenders at the speed of AI - The Official Microsoft Blog](#)

Product

GitHub Copilot X: The AI-powered developer experience

GitHub Copilot is evolving to bring chat and voice interfaces, support pull requests, answer questions on docs, and adopt OpenAI's GPT-4 for a more personalized developer experience.



Author
 Thomas Dohmke

March 22, 2023

[GitHub Copilot X: The AI-powered developer experience | The GitHub Blog](#)

AI for Data Analysis and Data Interpretation



New Results

[Follow this preprint](#)

Automated model building and protein identification in cryo-EM maps

Kiarash Jamali, Lukas Käll, Rui Zhang, Alan Brown, Dari Kimanius, Sjors H.W. Scheres

doi: <https://doi.org/10.1101/2023.05.16.541002>

This article is a preprint and has not been certified by peer review [what does this mean?].

[Abstract](#) [Full Text](#) [Info/History](#) [Metrics](#) [Preview PDF](#)

Abstract

Interpreting electron cryo-microscopy (cryo-EM) maps with atomic models requires high levels of expertise and labour-intensive manual intervention. We present ModelAngelo, a machine-learning approach for automated atomic model building in cryo-EM maps. By combining information from the cryo-EM map with information from protein sequence and structure in a single graph neural network, ModelAngelo builds atomic models for proteins that are of similar quality as those generated by human experts. For nucleotides, ModelAngelo builds backbones with similar accuracy as humans. By using its predicted amino acid probabilities for each residue in hidden Markov model sequence searches, ModelAngelo outperforms human experts in the identification of proteins with unknown sequences. ModelAngelo will thus remove bottlenecks and increase objectivity in cryo-EM structure determination.

[Automated model building and protein identification in cryo-EM maps | bioRxiv](#)

MLCopilot: Unleashing the Power of Large Language Models in Solving Machine Learning Tasks

Lei Zhang* Yuge Zhang Kan Ren Dongsheng Li Yuqing Yang

Microsoft Research

isleizhang@outlook.com, {yugzhan, kanren}@microsoft.com

[\[2304.14979\] MLCopilot: Unleashing the Power of Large Language Models in Solving Machine Learning Tasks \(arxiv.org\)](#)

Models of the Physical World

Importance of equivariant features in machine-learning interatomic potentials for reactive chemistry at metal surfaces

Wojciech G. Stark,¹ Julia Westernmayr,^{1,2} Oscar A. Douglas-Gallardo,^{1,3}
James Gardner,¹ Scott Habershon,¹ and Reinhard J. Maurer^{1,4,*}

¹*Department of Chemistry, University of Warwick,
Gibbet Hill Road, Coventry CV4 7AL, United Kingdom*

²*Current address: Wilhelm Ostwald Institute for Physical and Theoretical Chemistry, University of Leipzig, Germany*

³*Current address: Instituto de Ciencias Químicas, Facultad de Ciencias,
Universidad Austral de Chile, Isla Teja, Valdivia 5090000, Chile*

⁴*Department of Physics, University of Warwick,
Gibbet Hill Road, Coventry CV4 7AL, United Kingdom*

(Dated: May 22, 2023)

Accurate Surface and Finite Temperature Bulk Properties of Lithium Metal at Large Scales using Machine Learning Interaction Potentials

Mgcini Keith Phuthi,[†] Archie Mingze Yao,[†] Simon Batzner,[‡] Albert Musaelian,[‡]
Boris Kozinsky,[‡] Ekin Dogus Cubuk,[¶] and Venkatasubramanian Viswanathan^{*,†,§}

[†]*Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA*

[‡]*School of Engineering and Applied Science, Harvard University, Cambridge, MA, USA*

[¶]*Google Research, Brain Team*

[§]*Corresponding Author*

E-mail: venkvis@cmu.edu

Accurate Fourth-Generation Machine Learning Potentials by Electrostatic Embedding

Tsz Wai Ko,^{*,†} Jonas A. Finkler,[‡] Stefan Goedecker,[‡] and Jörg Behler^{*,†}

[†]*Universität Göttingen, Institut für Physikalische Chemie, Theoretische Chemie,*

Tammannstraße 6, 37077 Göttingen, Germany

[‡]*Department of Physics, Universität Basel, Klingelbergstrasse 82, 4056 Basel, Switzerland*

[¶]*Present Address: Lehrstuhl für Theoretische Chemie II, Ruhr-Universität Bochum, 44780*

Bochum, Germany, and Research Center Chemical Sciences and Sustainability, Research

Alliance Ruhr, 44780 Bochum, Germany

E-mail: tko@chemie.uni-goettingen.de; joerg.behler@ruhr-uni-bochum.de

Experimental Assistant's

ChemCrow: Augmenting large-language models with chemistry tools

Andres M Bran^{1,2*} Sam Cox^{3*} Andrew D. White³ Philippe Schwaller^{1,2}

¹ Laboratory of Artificial Chemical Intelligence (LIAC), ISIC, EPFL

² National Centre of Competence in Research (NCCR) Catalysis, EPFL

³ Department of Chemical Engineering, University of Rochester

*Contributed equally.

{andres.marulandabran, philippe.schwaller}@epfl.ch

{samantha.cox, andrew.white}@rochester.edu

Emergent autonomous scientific research capabilities of large language models

Daniil A. Boiko,¹ Robert MacKnight,¹ and Gabe Gomes^{*1,2,3}

1. Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
2. Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA 15213, USA
3. Wilton E. Scott Institute for Energy Innovation, Carnegie Mellon University, Pittsburgh, PA 15213, USA

*corresponding author, gabegomes@cmu.edu

[Submitted on 9 Jun 2023 (v1), last revised 13 Jun 2023 (this version, v2)]

14 Examples of How LLMs Can Transform Materials Science and Chemistry: A Reflection on a Large Language Model Hackathon

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D. Bocarsly, Andres M Bran, Stefan Bringuier, L. Catherine Brinson, Kamal Choudhary, Defne Circi, Sam Cox, Wibe A. de Jong, Matthew L. Evans, Nicolas Gastellu, Jerome Genzling, María Victoria Gil, Ankur K. Gupta, Zhi Hong, Alishba Imran, Sabine Kruschwitz, Anne Labarre, Jakub Lála, Tao Liu, Steven Ma, Sauradeep Majumdar, Garrett W. Merz, Nicolas Moitessier, Elias Moubarak, Beatriz Mouriño, Brenden Pelkie, Michael Pieler, Mayk Caldas Ramos, Bojana Ranković, Samuel G. Rodrigues, Jacob N. Sanders, Philippe Schwaller, Marcus Schwarting, Jiale Shi, Berend Smit, Ben E. Smith, Joren Van Heck, Christoph Völker, Logan Ward, Sean Warren, Benjamin Weiser, Sylvester Zhang, Xiaoqi Zhang, Ghezal Ahmad Zia, Aristana Scourtas, KJ Schmidt, Ian Foster, Andrew D. White, Ben Blaiszik

Chemistry and materials science are complex. Recently, there have been great successes in addressing this complexity using data-driven or computational techniques. Yet, the necessity of input structured in very specific forms and the fact that there is an ever-growing number of tools creates usability and accessibility challenges. Coupled with the reality that much data in these disciplines is unstructured, the effectiveness of these tools is limited.

Motivated by recent works that indicated that large language models (LLMs) might help address some of these issues, we organized a hackathon event on the applications of LLMs in chemistry, materials science, and beyond. This article chronicles the projects built as part of this hackathon. Participants employed LLMs for various applications, including predicting properties of molecules and materials, designing novel interfaces for tools, extracting knowledge from unstructured data, and developing new educational applications.

The diverse topics and the fact that working prototypes could be generated in less than two days highlight that LLMs will profoundly impact the future of our fields. The rich collection of ideas and projects also indicates that the applications of LLMs are not limited to materials science and chemistry but offer potential benefits to a wide range of scientific disciplines.

Subjects: **Materials Science (cond-mat.mtrl-sci)**; Machine Learning (cs.LG); Chemical Physics (physics.chem-ph)

Cite as: arXiv:2306.06283 [cond-mat.mtrl-sci]

(or arXiv:2306.06283v2 [cond-mat.mtrl-sci] for this version)

<https://doi.org/10.48550/arXiv.2306.06283> 

Submission history

From: Kevin Maik Jablonka [[view email](#)]

[v1] Fri, 9 Jun 2023 22:22:02 UTC (12,598 KB)

[v2] Thu, 13 Jun 2023 07:44:33 UTC (12,598 KB)

LLM Materials Science Hackathon

“Motivated by recent works that indicated that large language models (LLMs) might help address some of these issues, we organized a hackathon event on the applications of LLMs in chemistry, materials science, and beyond. This article chronicles the projects built as part of this hackathon. Participants **employed LLMs for various applications, including predicting properties of molecules and materials, designing novel interfaces for tools, extracting knowledge from unstructured data, and developing new educational applications.**”

Panelist Talks

CI Compass aims to connect the following NSF facilities

Katie Dagon (NCAR)

Dr. Katie Dagon is a Project Scientist II at the National Center for Atmospheric Research (NCAR) in Boulder, working in the Climate and Global Dynamics Laboratory. Her research focuses on modeling the impacts of climate change on land-atmosphere interactions, climate variability, and extreme events. She is also involved with a variety of projects applying machine learning and artificial intelligence methods to climate science and modeling. At NCAR she helps organize the Earth System Data Science initiative. Katie obtained her Ph.D. in Earth and Planetary Sciences from Harvard University and her B.S. in Mathematics-Physics from Brown University.

Dan Stanzione (TACC)

Dr. Dan Stanzione, Associate Vice President for Research at UT Austin since 2018 and Executive Director of the Texas Advanced Computing Center (TACC) since 2014, is a nationally recognized leader in high performance computing. He serves on the National Artificial Intelligence Research Resource Task Force, formed by NSF and the White House Office of Science and Technology Policy (OSTP). He is the principal investigator (PI) for an NSF grant to deploy Frontera, the fastest supercomputer at any U.S. university. Stanzione is also the PI of TACC's Stampede2 and Wrangler systems, supercomputers for high performance computing and for data-focused applications, respectively. For six years he was co-PI of CyVerse, a large-scale NSF life sciences cyberinfrastructure. Stanzione was also a co-PI for TACC's Ranger and Lonestar supercomputers, large-scale NSF systems previously deployed at UT Austin. Stanzione received his bachelor's degree in electrical engineering and his master's degree and doctorate in computer engineering from Clemson University.

Philip Harris (MIT)

Philip Harris joined the MIT faculty in 2017. Born in Sao Paulo, he received his B.S in Physics from Caltech in 2005, and his Ph.D from MIT in 2011 on research performed at CERN with the Large Hadron Collider(LHC). From 2011-2013, Philip was a CERN fellow working on the Higgs discovery. From 2014-2017, he was a CERN staff scientist working on dark matter searches at the CMS experiment. Philip is one of the founders of the Fast Machine Learning Organization. Also, he is currently the experimental coordinator of The Institute for AI and Fundamental Interactions (IAIFI), and he is the deputy director for the Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery (A3D3) Institute. In addition, Philip is responsible for the deployment of real-time AI algorithms in the upgraded readout system of the CMS detector on the LHC. He is also working on new strategies to integrate LHC computing with High Performance Computing centers through the use of AI algorithms integrated with optimized heterogeneous computing.

Pete Beckman (ANL)

Co-Director, Northwestern Argonne Institute of Science and Engineering and Argonne Distinguished Fellow, Northwestern University / Argonne National Laboratory. Pete Beckman is a recognized global expert in high-end computing systems and Argonne Distinguished Fellow. He co-directs the Northwestern University / Argonne Institute for Science and Engineering, and coordinates the collaborative technical research activities in extreme-scale computing between the U.S. Department of Energy and Japan's ministry of education, science, and technology. Beckman's expertise encompasses software and architectures for large-scale parallel and distributed computing systems. Beckman leads the Argo project focused on low-level resource management for the operating system and runtime, he is the founder and leader of Argonne's Waggle project for artificial intelligence (AI) and edge computing, and he also leads the Sage project funded by the National Science Foundation to build a nationwide infrastructure for AI at the edge.

Panel Discussion

Question 1: AI Scientists?

- Given the rise of AI tools, and LLM tools like ChatGPT and their potential to act as co-pilots in scientific research, how do you see these AI tools being integrated into the research infrastructure ecosystem to enable more efficient scientific discovery?
- What key challenges do we need to anticipate and address?
- What “properties” do AI models and assistants need to do science research?

Question 2: AI Models, Assistants, and Infrastructure

AI has the potential to revolutionize the analysis of scientific data and modeling of complex phenomena, as evidenced by recent papers. However, the successful integration of AI into research infrastructure requires a robust data governance framework for both data and models.

- What does this look like to you, and how can we achieve it while respecting FAIR data principles?
- How might AI help with these issues?
- How do we plan for this integration and how do we measure success?
- How do we manage the exponential acceleration resulting in the AI “firehose”?

Audience Questions?

Bonus Round: A vision for the future?

- Looking ahead, how do you envision AI tools transforming the role of researchers and the overall landscape of scientific research in the next 10 years?
- What are the implications of these transformations on the design of our research infrastructure, skills needed, and ethical considerations we should bear in mind?
- Are there risks related to AI that you worry about?