THE UNIVERSITY OF ARIZONA

RESEARCH, INNOVATION & IMPACT
Data Science Institute

The Data Drip

Special edition
July 2024

A Digest of Data Science News -
University of Arizona

DATA SCIENCE INSTITUTE (DSI)

We're presenting at the IT Summit!

The 12th annual **University of Arizona 2024 IT Summit** is scheduled for **August 6th in-person** at the Student Union Memorial Center and **virtually on August 7th** which includes workshops and presentations with remote live streams. Join the **Data Science Institute**, **CyVerse**, **ICDI**, and others as we network with peers and learn more about advances in technology. We'll see you there!

Register for August 6th and **August 7th**. Pop in to hear us at these sessions on the AI Empowerment track.

Tuesday, August 6th

- **Seeing Beyond the AI Hype Cycle**
- **Collaborate in the Cloud with CyVerse**
- **So You Want to Build an AI Application?**
- **AI Working Group Reports**

Wednesday, August 7

- **Day in the Life of a Data Scientist**
- **Building your own Chatbots with a Purpose**
- **UArizona & NSF Infrastructure for AI**
- **Making Cloud Infrastructure Accessible with CACAO**
- **Sensor Fabric: AI from Wearable Data**
- **Do More with your Data Using RAG: Building Chatbots**

- **Teaching infrastructure for AI/ML**
- **Learning & Maker Space for ML/AI: UArizona Data Lab**
- **Campus AI Practitioners Community Building**

Aug. 6th details for selected sessions on the AI Empowerment track

10:30 - 10:50 AM Seeing Beyond the AI Hype Cycle

- Nirav Merchant, Director, Data Science Institute

Looking to keep up with the dizzying pace of AI innovations when they're often closed source and proprietary? Fortunately, there is a vibrant ecosystem of open source tools and platforms with permissive licenses, coupled with no cost and highly scalable AI infrastructure available through NSF ACCESS and the recent National AI Research Resource (NAIRR) initiative. With these NSF resources alongside campus computational infrastructure, IT and AI expertise, U of A researchers and students can build AI powered solutions that meet the Open Science and reproducibility requirements from federal agencies. We all have an important role to play in shaping AI innovation as creators, consumers and educators. This requires community engagement, hands-on training for pragmatic and informed adoption of AI capabilities as they mature over time.

1:30 - 1:50 PM Collaborate in the Cloud with CyVerse

- Edwin Skidmore, Dir. Infrastructure, Cloud Native Services, CyVerse
- Tony Edgin, Senior Research Engineer, Core Software, CyVerse
- Sarah Roberts, Team Lead & Senior Research Software Engineer, Core Software, CyVerse

Get an overview and demonstration for how CyVerse Data Store and cloud automation platforms are democratizing access to scalable commercial and private cloud resources. The CyVerse Data Store integrates with any S3 compatible storage service—like AWS Simple Storage Service and the Open Storage Network (OSN)—to allow users to share their data securely with collaborators at any institution while keeping their S3 credentials private. Access is easy through web and desktop tools, and container (docker) based analysis tools. If you are launching cloud-based virtual machines in services like AWS Elastic Compute Cloud (EC2) or NSF's CloudBank and Jetstream2, or NIH STRIDES, save time and make your work reproducible by automating deployment. CyVerse Cloud Automation and Continuous Analysis Orchestration (CACAO) combines the Infrastructure as Code (IaC) capabilities of Terraform and Ansible to create recipes that are readily deployed through an easy to use and simple web interface for end users.

2:00 - 2:20 PM So You Want to Build an AI Application?

- Nirav Merchant, Director, Data Science Institute
- Shravan Aras, Assistant Director for Sensor Analysis & Smart Platforms, Center for Biomedical Informatics & Biostatistics

Learn where to start! AI based applications typically require a significant amount of computational and data resources, and higher rigor in software engineering and data management. Find out how to make effective use of campus and NSF resources, alongside platforms like Hugging Face, to explore AI models and build and deploy solutions that span campus and commercial resources.

3:00 - 3:45 PM AI Working Group Reports

- Lanita Collette, Deputy Chief Information Officer
- Barney Maccabe, Executive Director, Institute for Computation and Data-Enabled Insight
- Justin Starren, Chief, Division of Health and Biomedical Informatics

Join to hear summaries of the results from work the campus AI working groups have been doing over the last year.

Aug. 7th details for selected sessions on the AI Empowerment track

9:00 - 9:50 AM Day in the Life of a Data Scientist

- Enrique Noriega-Atala, Computing Sciences Researcher, Computer Science Department
- Andrew Bennett, Assistant Professor Hydrology & Atmospheric Sciences

Take a journey in building AI-powered applications at the University of Arizona. Data scientists are tasked with extracting value and perform lead discovery from complex data sets. Every discipline has their specialized tools, analysis methods and workflows to process their disciplinary data modalities (image, text, audio, etc.). Hear from two U of A researchers on how they have built their tool boxes and their journey in building ML/AI solutions for their projects.

10:00 - 10:50 AM Building your own Chatbots with a Purpose

- Jay Timsina, Enterprise Architect, UITS
- Mithun Paul, Research Scientist & Educator, Data Science Institute
- Drennen Brown, Director, Information Technologies and Spatial Services
- Jon Kevan, Instructional Design & Support Manager, UA Health Sciences Global & Online
- Alan Hamilton, Executive Director, Arizona Simulation Technology & Education Center, UA

Health Sciences

- Mario Vasquez, User Experience Designer, UITS

Large Language Models (LLM) and Generative Pre-trained Transformers (GPT) with chat interfaces like ChatGPT have ushered a new era for how users are interacting with complex knowledge platforms. There are many tangible benefits in integrating these LLM and GPT based solutions with data from sources that are not readily accessible or available on the public internet. Multiple groups at the University of Arizona are prototyping solutions where they have successfully developed applications that combine the power of these technologies. Get highlights from some of the use cases and demonstrations, followed by a panel discussion.

11:00 - 11:50 AM UArizona & NSF Infrastructure for AI

- Tyson Swetnam, Co-PI and Science Lead, Cyverse
- Chris Reidy, Manager, Research & Discovery Technologies, UITS

Discover the physical and virtual resources available for AI teaching, training, and research from the Research Computing group in UITS and from University of Arizona's owned and operated CyVerse, an NSF project. Learn what current resources are freely available to all University of Arizona staff, faculty, and students and how to gain access to them. Find out also about national AI efforts, including the National Artificial Intelligence Research Resource Pilot (NAIRR) and Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS).

1:00 - 1:20 PM Sensor Fabric: AI from Wearable Data

- Shravan Aras, Assistant Director for Sensor Analysis & Smart Platforms, Center for Biomedical Informatics & Biostatistics

Wearables are an essential part of multiple research projects and studies, but researchers struggle in managing data from sensors and deriving value from it. Learn about Sensor Fabric and its architectural design for making data AI ready. Find out how the presenters developed the Sensor Fabric to address many of the common challenges faced for ingesting wearable sensor data from Fitbit, Apple Watch, Sensor Kit, Empatica, etc. into a scalable and cost-effective cloud-based solution. This includes use of object storage such as S3 and schema catalogs such as AWS Athena. See how data from Sensor Fabric can be efficiently and quickly used to create AI models written in PyTorch and also using AutoML techniques offered by both GCP and AWS, along with connections to open-source visualization tools running in CyVerse.

1:00 - 2:50 PM Making Cloud Infrastructure Accessible with CACAO

- Edwin Skidmore, Director of Infrastructure, CyVerse
- Jeremy Frady, DevOps Engineer, Cloud Native Services, CyVerse, Data Science Institute

- John Xu, Research Software Engineer, CyVerse, Data Science Institute
- Michele Cosi, Data Science Educator, CyVerse, Data Science Institute

Learn how to leverage CACAO (Cloud Automation and Continuous Analysis Orchestration), an open-source, cloud-native platform, to use infrastructure-as-code (IaC) or domain-specific language (DSL) templates for your deployments. Simplify the challenges with deploying complex software installations and configurations on multiple virtual machines or cloud services, including those with specialized hardware like GPUs. Make manual installation less time-consuming and error prone. With templates, your deployments are also reproducible. Hear about best practices from multiple NSF projects that are utilizing CACAO for a range of activities, including cloud access for science and engineering, ML-based simulations, data science workshops in Ecology and Geospatial analysis, and LLM based workflows and inference.

1:00 - 2:50 PM Do More with your Data Using RAG: Building Chatbots

- Iqbal Hossain, Director, Research Data Science, Research Innovation & Impact
- Enrique Noriega-Atala, Computing Sciences Researcher, Computer Science Department
- Mithun Paul, Research Scientist & Educator, Data Science Institute
- Jay Timsina, Enterprise Architect, UITS

Get an introduction to Retrieval-Augmented Generation (RAG) systems within our campus environment, revolutionizing how we manage and utilize information in the realm of generative AI. Explore their transformative potential in streamlining workflows, facilitating knowledge retrieval, and fostering innovation. See how they work in successful deployments, such as at KMap, and find out the strategies employed and lessons learned in the process. Examine RAG system architecture, including the mechanics of vector databases, semantic search capabilities, and effective prompting methodologies.

1:30 - 1:50 PM Teaching infrastructure for AI/ML

- Tyson Swetnam, Co-PI and Science Lead, CyVerse
- Chris Reidy, Manager, Research & Discovery Technologies, UITS

Do you need to include hands-on AI/ML learning material in courses and workshops? ML/AI based assignments need specialized software stacks (Jupyter, MLFlow, Pytorch etc.) with dependencies that make them hard to manage and maintain. Additionally, they often require specialized hardware like GPUs and significant data handling capacity. Find out about the resources available to faculty wanting to develop new creative courses and assignments. Learn common practices that ease the burden of managing the ML/AI software resources, and how to utilize no cost infrastructure available through NSF ACCESS and UA HPC.

2:00 - 2:50 PM Learning & Maker Space for ML/AI: UArizona Data Lab

- Carlos Lizarraga, Computation and Data Science Educator, Data Science Institute
- Brenda Huppenthal, GRA, Computer Science, UArizona Data Lab
- Megh Krishnaswamy, GRA, Linguistics, UArizona Data Lab
- Chan "CK" Chi-Kwan, Associate Astronomer, Steward Observatory

Do you need to stay on top of the evolving landscape of machine learning (ML) and artificial intelligence (AI)? Learn and experiment with new developments and advancements at the UArizona DataLab. Participate in codefests/hackathons on cross-cutting topics of interest. This hub and clearinghouse for the U of A community provides:

• extracurricular and non-formal training through short format workshops in ML/AI tools and technologies

• consultations on research projects for incorporating ML/AI methods into research lab workflows

• assistance in onboarding and training students and staff who are involved in implementing ML/AI solutions

Gain an understanding of the progress and roadmap for non-formal learners to utilize DataLab, and for departments and other units to participate in DataLab to showcase their activities and highlight the vibrancy of ML/AI learning opportunities offered across campus.

3:00 - 3:50 PM Campus AI Practitioners Community Building

- Michael Dorland, IT Support Analyst, Campus IT Partnerships
- Barney Maccabe, Executive Director, Institute for Computation and Data-Enabled Insight
- Ash Black, Director AI & Industry, Institute for Computation and Data-Enabled Insight/AI Core
- Shravan Aras, Assistant Director for Sensor Analysis & Smart Platforms, Center for Biomedical Informatics & Biostatistics
- Darcy van Patten, Chief Technology Officer
- Chan "CK" Chi-Kwan, Associate Astronomer, Steward Observatory

Hear from the panel on steps taken to bring together the University community using AI. Be part of the conversation for next steps.

NEWS AND ACTIVITIES

BIO5 hosts the GSC24 conference in August

The **Genomic Standards Consortium conference** with a focus on "Challenges of Reproducibility in Genomics" will

be held at the University of Arizona at BIO5 Institute on August 5-9, 2024. The GSC conference is organized by Genomic Standards Consortium, an open, community organization dedicated to making genomic data discoverable. Note: In-person registration is closed, but **virtual registration** is available until August 5.

Supporting new graduate & professional students

Stop by the Data Science Institute, CyVerse, and ICDI tables during the **New Graduate and Professional Student Orientation event** on August 21, 2024. The event showcases a comprehensive introduction to some of the most important resources that the University of Arizona has to offer.

SAVE THE DATE

NEW GRADUATE AND PROFESSIONAL STUDENT ORIENTATION

WEDNESDAY, AUGUST 21, 2024

THE UNIVERSITY OF ARIZONA
Graduate College

CALENDAR

Be sure to check often the **Data Science Institute calendar** for useful data science workshops, trainings, and events offered by the Data Science Institute, University of Arizona DataLab, CyVerse and other departments across the University of Arizona campus.

Share **The Data Drip** with your friends and colleagues! **Subscribe** by completing a brief form. Let us know if you have data science news that you want us to share in The Data Drip or on the **Data Science Institute** website.

Share!
The Data Drip
A digest of Data Science news at the University of Arizona

Email the Data Science Institute

Email the Data Science Institute and Subscribe to The Data Drip.

SUBSCRIBE TO THE DATA DRIP

Land Acknowledgment

We respectfully acknowledge the University of Arizona is on the land and territories of Indigenous peoples. Today, Arizona is home to 22 federally recognized tribes, with Tucson being home to the O'odham and the Yaqui. Committed to diversity and inclusion, the University strives to build sustainable relationships with sovereign Native Nations and Indigenous communities through education offerings, partnerships, and community service.