# RII Track-1: Data Analytics that are Robust and Trusted (DART): From Smart Curation to Socially Aware Decision Making

## Year 3 Annual Report

Award Number: 1946391
Jurisdiction: Arkansas
PI: Jennifer Fowler
Co-PI: Jackson Cothren
Awardee Institution: Arkansas Economic Development Commission
Awardee DUNS: 619312569
Award Start Date: July 1, 2020
Award End Date: June 30, 2025 (Estimated)
Report Submission Date: April 1, 2023
Reporting Period: April 1, 2022 – March 31, 2023

# Table of Contents

# Overview

## Vision

The Arkansas research community - academic, government, and industry - collaborating often and easily on a shared computing platform with access to high performance computing nodes, peta-byte scale storage, fast and reliable big data transfer, and shared software environments which facilitates replicable, reproducible, and cutting-edge data science research. Reliable, scalable, explainable, and theoretically grounded data science approaches to data life cycles and modeling allow the public to better understand how machine learning and artificial intelligence effects their lives. When they engage with data science products on their smart devices, on social media platforms, and on the web, the improved and robust privacy and safety protections and fair results increase their trust of data collection and the resulting information, allowing for broader use of data science to benefit society. In Arkansas, the educational ecosystem provides learners with a well-designed, consistent, sequenced, scaffolded, and modular education in data science with further educational and/or job opportunities available at appropriate points in their careers.

## Mission

The mission of DART is to improve research capability and competitiveness in Arkansas by creating an integrated statewide consortium of researchers and educators working to establish a synergistic, statewide focus on excellence in data analytics research and training.

## Goals

The growing array of tools - powerful high-level programming languages, distributed data storage and computation, visualization tools, statistical modeling, and machine learning - along with a staggering array of big data sources, has the potential to empower people to make better and more timely decisions in science, business, and society. However, there remain fundamental barriers to practical application and acceptance of data analytics in these areas, any one of which could derail or impede its full development and contributions. These four topics form the integrative research and education activities on which DART will focus. Goals defined by each project team– cyberinfrastructure (CI), data curation and life cycle (DC), social media and networks (SM), social awareness (SA), and learning and prediction (LP), and education (ED)- as well as strategies in other project elements contribute to one or more of these topics.

1. **Big data management:** Before data streams and datasets can be used in learning models, they must be manually curated, or at the least, curated for a specific problem. We still rely on human analysts to assess the content and quality of source data, engineer features, define and transform data models, annotate training data, and track data processes and movement.

2. **Security and privacy:** Government agencies and private entities collect (often with only an individual's implicit consent), process it often in near-real-time, and deliver products or services based on these data to consumers and constituents. There are increasing worries that both the acquisition and subsequent application of big data analytics are not secure or well-managed. This can create a risk of privacy breaches, enable discrimination, inject biases, and negatively impact diversity in our society.
3. **Model interpretability:** Machine learning models often sacrifice interpretability for predictive power and are difficult to generalize beyond their training and test data. But interpretability and generalizability of trained models is critical in many decision-making systems and/or processes, especially when learning from multi-modal and heterogeneous big data sources. There is a continuing to need to better balance the predictive power of complex machine learning models with the strengths of statistical models to better configure deep learning models to allow humans to see the reasoning behind the predictions.
4. **Data-Skilled Workforce:** As data-driven science and decision making become commonplace, our state and nation will need to rely on a well-educated workforce at almost all levels of responsibility to be aware of the power and pitfalls of using data in decision making. This topic represents a significant addition in year 2. It is a natural and effective way to think about how education and workforce development efforts integrate with research efforts.

## Key Accomplishments During Year 3

Year 3 was a pivotal year for the project. This year, we hosted the first in-person graduate student and faculty retreat, which was the first time some of the participants met in person. DART participants made keynote presentations at international conferences, published articles and served as editors in prestigious academic journals, submitted numerous proposals that were funded by Federal, private, and institutional sources, and received numerous honors and awards, which will be discussed in further detail throughout this report.

During Year 3, we recruited new members to both our external advisory board (EAB) and industry advisory board (IAB). Our new EAB member is Dr. Scotty Strachan, who serves as the Co-PI for Cyberinfrastructure for Nevada EPSCoR's current Track-1 project. He is the principal research engineer at the Nevada System of Higher Education and was co-PI on the EPSCoR Cyberinfrastructure workshop that took place during the 2022 National EPSCoR conference in Portland Maine. New IAB members include Dr. Yanbin Ye, who graduated from UALR and is working as a Data Science Director at Walmart; Diane Schmidt, Head of Enterprise Data Governance at the London Stock Exchange Group; Ruinan Wang, Principal Product Manager at Amazon Web Services; and Amy Elrod, Director of Talent Retention at Acxiom. Some existing IAM members have new professional roles, including Adewale Obadimu who left LinkedIn and is now at Diligent Robotics. Heather Snell moved from Arvest

to JB Hunt, and Kash Mehdi left Informatica for a European firm called Data Galaxy. Finally, Ty Keller left Hytrol and a meeting is scheduled to identify a replacement from Hytrol for the IAB.

A major highlight is the establishment of a working prototype that solved a critical technical challenge for the Arkansas Research Platform (ARP) in collaboration with the University of Arkansas systems office and IT staff from several institutions. Details are provided in the CI portion of the report below.

**Key Abbreviations of Project Components:**

Cyberinfrastructure component (CI)

Data Curation & Life Cycle component (DC)

Social Media component (SM)

Social Awareness & Privacy component (SA)

Education component (ED)

Learning and Prediction component (LP)

## Significant Problems

One challenge the project has faced is nearly unprecedented turnover of personnel. We have experienced numerous faculty leaving the state for positions in industry and academic institutions in non-EPSCoR states. One possible explanation is that the fields of research in the DART project are in such high demand nationwide, and we are experiencing a relatively normal 'brain drain' because Arkansas can't offer high salaries and societal amenities available in larger metro areas and non-EPSCoR states. If this is the case, it is worrisome to think of what will happen when other EPSCoR states pursue cutting edge topics for Track-1 projects. Another factor that may contribute to the seemingly high turnover is simply that the project has the largest number of faculty participants of any previous Track-1 project in Arkansas, resulting in proportionate rates of turnover.

Another concern is the current political climate in Arkansas, though it is not unique to our state. The tension between the state legislature and the Federal government appears to grow, and over the past several years our state legislature has passed a number of laws to reduce the power of the Governor and executive branch (including the loss of the Governor's ability to veto laws). A number of bills were introduced (but not ultimately passed) during our current legislative session that would potentially impact some of DART's work. One proposed to not only end any affirmative action in the state, but also make it punishable by a class A misdemeanor. Another proposed to prohibit all Russian, Chinese, or Iranian nationals from purchasing property in Arkansas. Many of our project participants are from these countries, as

well as numerous other Arkansas residents working in science and engineering. These measures did not ultimately pass, but made enough progress to cause concern.

## Changes in Strategy

The project intends to submit a revised strategic plan to NSF in Spring 2023, due to personnel turnover and feedback from advisory boards and evaluation checkpoints. The January 2023 retreat served as a collaborative working meeting where the DART teams worked on their strategic plan revisions, and discussed how to better integrate various aspects across teams. Other factors contributing to the revised milestones also include the expected closure of the Central Office when the DART award ends, due to the programmatic changes at NSF EPSCoR. No changes in scope are planned.

We have made a decision to remove the middle school coding block effort from the Education efforts. This is due to the loss of the project champion at our partner institution, the Arkansas School for Mathematics, Sciences, & the Arts, as well as the increase in resources that are available to educators since the proposal submission in 2019. After the project champion Dr. Daniel Moix left for another state, the team reached out to the Arkansas Department of Education and other stakeholders. The consensus from the people surveyed was that this was no longer a high need for schools in Arkansas. We plan to rebudget the funds allotted for this to support other education efforts that will be reported next year.

## Personnel Changes

The project experienced turnover in the SSC including the loss of Xintao Wu, who stepped down to focus on the research activities; Xiuzhen Huang, who is on sabbatical; and Justin Zhan, who left the state. Qinghua Li has stepped forward to represent the SA team on the SSC, and we are considering participants that could fill the other gaps. Below is a table summarizing the other recent personnel changes, most of which are not reflected in the current strategic plan but will be updated in the revision. All but 2 of these individuals left the state for other offers.

Justin Zhan who was serving as co-lead for SM left the state for a position in Ohio. After a search, a candidate was identified to pick up some of the effort on SM goals, Susan Gauch at UARK. Susan met with leads from SA and SM to review the planned activities and will have request some updates to milestones based on differences in expertise.

Olcay Kursun and Paul Schrader left the state, LP3 will have updates in the strategic plan revision to reflect personnel changes. Ahmad Al-Shami and Zachary Stine have been recruited to the team. Zachary Stine was previously supported by the project as a graduate student at SAU, and upon graduation received a faculty position at UCA. As a result of personnel changes and feedback from advisors, the team plans to request minor changes to some milestones but no changes in the goals.

During Year 3, SA SSC co-lead Xiuzhen Huang went on sabbatical, and co-lead Xintao Wu asked to step down from the SSC to focus on research progress. The team also faced a number of other personnel losses, resulting in a large number of projects and activities outlined in the strategic plan for the smallest team in the overall project. At the January retreat, SA team member Qinghua Li volunteered to serve on the SSC and represent the SA team moving forward. Zenghui Sha left the state, and the SA team's milestones and activities in the strategic plan will be updated to reflect these personnel changes as well as the incorporation of feedback from the EAB and RSV panelists.

| Lost Personnel, Campus, DART Team | Replacement |
|---|---|
| Justin Zhan, UARK, SM | Susan Gauch, UARK, SM/SA |
| Zenghui Sha, UARK, SA | None |
| Xiuzhen Huang, A-State, SA | Jonathan Stubblefield, A-State, DC |
| Paul Schrader, SAU, LP | Ahmad Al-Shami, SAU, LP |
| Olcay Kursun, UCA, LP | Zachary Stine, UCA, LP |
| Elizabeth Pierce, UALR, DC | Daniel Berleant, UALR, DC |
| Sean Young, UAMS, Seed Grant Recipient | None |
| Esther Mead, UALR, Seed Grant Recipient | None |
| Emily Bellis, A-State, Seed Grant Recipient | None |

# Research & Education Program - Year 3 Accomplishments

In this section of the report, we will provide updates according to each project team. Each subsection will begin with a summary of highlights from the team, as well as details on personnel changes and responses to EAB/RSV feedback as applicable. The milestone and activity tables (or stoplight tables) are presented under each subsection. Please refer to the following legend for the stoplight tables:

| **Objective No.** | **Specific Milestones** | | | |
|---|---|---|---|---|
| | Year 1 | Year 2 | | Year 3 |
| **Activity #**<br><br>**CELL COLOR LEGEND:** | **Milestone is Complete** | **Milestone will be completed by end of reporting year** | **Milestone is delayed or will not be completed** | **Changes to milestone requested in strategic plan update** |

## Coordinated Cyberinfrastructure

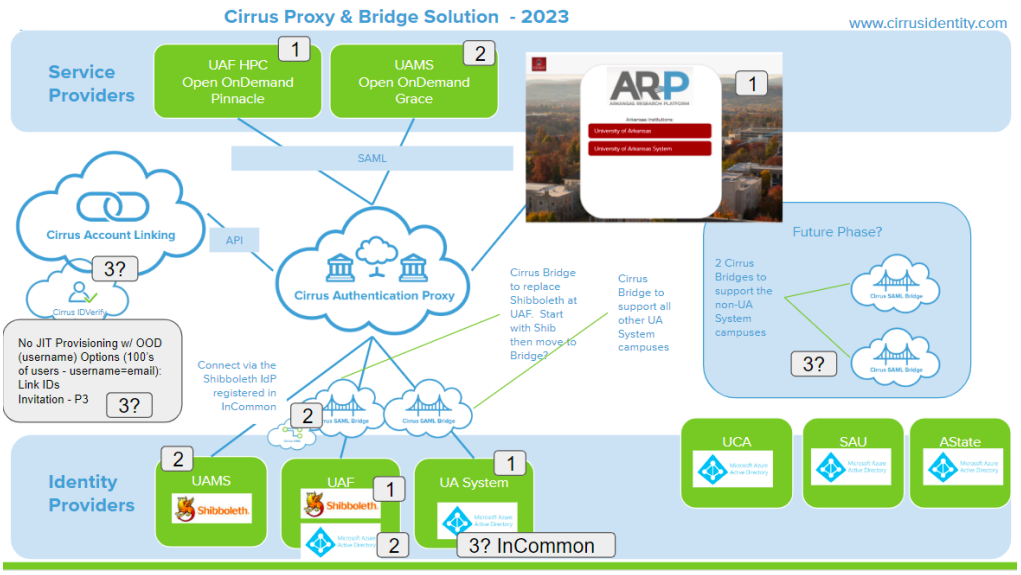
In the vRSV report, the panelists noted some significant challenges that the project was facing in implementing the vision of the Arkansas Research Platform. In our response to the report, we described some progress under a new collaboration with the University of Arkansas Systems office, and participants in the CC* CIRA: Shared Arkansas Research Plan for Community Cyber Infrastructure (SHARP CCI) project. The SHARP CCI award has enabled additional support to comprehensively assess cyberinfrastructure, cybersecurity, and computational research at participating campuses. As stated in the vRSV response, ARP needs a federated identity solution that accommodates both Azure Active Directory (AD) and on-premises AD installations (as well as identity providers that are supported by InCommon), which covers the majority of DART participating institutions, and most campuses in Arkansas. UCA incorporates a different Shibboleth-based solution. Among DART campuses, only UAF and UAMS are registered with InCommon, but ARP must also allow this service as an identity provider (IdP) to access non-DART institutions and researchers.

DART leadership has been working steadily with the UA Systems Office CIO and CISO and the vendor Cirrus to implement a solution. The primary issues we are working to resolve are not only identity federation of users from multiple institutions, but also the protocols that high-performance computers (HPC) like Grace and Pinnacle use to authenticate users for the HPC. In layman's terms, the HPCs don't speak the same language as the active directories that each campus is using to manage user access. This means that until we have a solution in place, anyone who wants to use the HPC resources at UARK and UAMS has to be manually vetted and their user credentials had to be manually entered to the HPC. This is not scalable, and with interest and demand growing for powerful computational infrastructure, it places undue burden on our limited HPC staff. The Cirrus Bridge is a product that will act as a translator between the HPC and AD, which will solve one of the key technical challenges. Contracts with

Cirrus are underway, and our team has been working directly with the CEO and a whole team from Cirrus to run a pilot, connecting both the UA Systems Azure AD and the UARK AD to Pinnacle, which will hopefully result in any authenticated user from either AD being able to log in and use the HPC with their standard credentials from their home institution. UAMS has configured Grace a bit differently with Shibboleth, and we are working on another similar solution involving Keycloak. The goal is that with these two solutions, any UA system authenticated user will be able to log in with their home institution's credentials, and would see an icon (right next to Workday and other applications) for Pinnacle Portal or Grace. This will also enable secure data sharing of large datasets and code files between DART researchers at different campuses across the state, which we currently do not have a robust/scalable solution for.

After these solutions are in place, the team has identified two campuses outside the UA system to work with as our test cases. We have a group of DART researchers and students at SAU Magnolia who are involved in the LP team. We've already met with the CIO for SAU, and determined their technical configurations which include an on-premise AD. When the Cirrus Bridge is effective, we will work with the SAU CIO Mike Argo to connect their AD to either the UA System AD or the UARK AD, which will enable SAU students and faculty to access Pinnacle and share data with their UARK collaborators in the LP team. The second test case will be UCA, which uses a Shibboleth configuration similar to UAMS. When the Keycloak solution is effective, we will work with DART SSC member Addison and the UCA CIO Trevor Siefert to pilot a connection between UCA and UAMS.

These topics were a major part of the focus of our January 2023 Retreat, and we were able to convene most of the key stakeholders during that meeting. In addition to the DART CI team, Eric Wall (UA Systems CISO), Steve Krogull (UARK CIO), Ian Czarnezki (UARK Associate CIO), James Deaton (EAB member), Scotty Strachan (new EAB member), David Chaffin and Pawel Wolinski (UARK HPC staff), Lawrence Tarbox and Kaleb Abram (UAMS HPC staff), and others were present to contribute to ARP strategic planning and technical discussions.




We are thrilled to report that as of March 2023, we have a working prototype where users within the UASYS Azure Tenant can automatically authenticate and access UARK's tenant. This is using a new UA System Bridge that Cirrus configured and is connected to the UARK UAT Proxy from the previous work last fall which fronts the HPC's Open OnDemand portal. A demonstration was recorded, showing the ability to log in to Pinnacle's Open OnDemand portal using credentials from the UASYS and UA Phillips County Community College. The demo can be viewed at this link with the following Passcode: 98PWx3?H

This breakthrough has marked a major step in the implementation of ARP. The next technical challenge will be to develop a workflow for approval of requesting accounts to the HPC user registry. This process cannot be securely automated, but the team is working to make the process as efficient as possible. We look forward to providing future updates on this to NSF and also the larger computing community. Presentations are planned at the upcoming OAK Supercomputing Conference in Oklahoma City, and the PEARC Symposium in Portland, OR this summer. Below is a summary of specific accomplishments and the stoplight tables for the team.

**CI1, Establish the Arkansas Research Platform as a shared data science resource across the jurisdiction.** The milestones for Objectives 1.1.a – 1.1.e are generally on schedule, with some minor changes. An updated MOU between UAMS and UARK is being considered for the ARCC advisory board, and roles and responsibilities are planned for further development in the remainder of Year 3. The milestones for the planned expansion of ARCC to other DART campuses will be updated in the requested strategic plan changes. Objective 1.1.c milestones are complete, both Grace and Pinnacle now have Hadoop environments. The DC team is currently using the Grace-based environment in its scaling activities. The equipment related to 1.1.c. has been delivered and installed. The 100Gb switch is expected to be installed before the end of the

reporting year, however, the direct link may not be ready until Fall. Both UAMS and UAF are working continually with their respective IT departments to ensure that the twin object stores and clusters are ready for the test. It is possible the link will go smoothly once the switch installed and configured and this activity will be complete by the end of the reporting year.

The Enterprise GitLab solution was not viable due to licensing and network restrictions at UAF. DART instead established a private GitHub Organization to act as a project-wide repository to appropriately share non-proprietary code and other resources. The CI team is in the process of working with the other participants to copy or mirror existing repositories to the DART GitHub Organization. Globus Data Management (GDM) was intended to provide various resources including file sharing and identify management. However, with the purchase of Cirrus bridge services and commitments from UAF and UAS to fund these services in perpetuity, and the ability of no-cost Globus services for file-sharing, a project-wide GDM was viewed as unnecessary and less sustainable. UAF will be able to host 5 carpentry workshops during the year, however, the goal of training 2 instructors has proven difficult to achieve. One instructor, Hanna Ford, has been trained but we are still working to identify potential instructors at other institutions. We will continue to work with our participating campuses to identify potential instructors but will modify this goal in the revised Strategic Plan.

Finally, for Objective 1.1.e., procedures for managing CUI have been implemented and tested with a non-DART-funded project. Procedures for developing and implementing System Security Plans are in place and defined on project. Similar plans are in place for HIPAA procedures are in place at UAMS. Because these plans are project specific and the institution is responsible for adherence, a DART-wide capability (e.g., a researcher at ASU would not host CUI or HIPAA data on the UAF or UAMS campus) they cannot be used directly on shared clusters. However, the procedures and templates can be made available across the DART institutions.

| **Objective 1.1.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Create ARCC advisory board with regional partners (GPN) | Form CI advisory board | Establish roles and responsibilities consistent with MOU for the advisory board | |
| **Activity 2**: Establish ARCC governance, operations and staff between UA and UAMS | | Document defining organizational structure, roles, and responsibilities of ARCC | |
| **Activity 3**: Expand ARCC to include UALR as a provider and other DART participants as consumers | | | Amended MOU among three institutions signed by Chancellors. |

| | | | |
|---|---|---|---|
| **Activity 4**: Create UAF CI Plan to support DART (prior to 1.1.b and 1.1.c) | | Publish a design and configuration document on DART website | |
| **Objective 1.1.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1**: Specify and purchase data science cluster based on document from 1.1.a | Issue UA purchase order for additional equipment | Receive and install new data science nodes on Pinnacle | |
| **Activity 2:** Test and deploy hardware elements for Pinnacle expansion for DART | | Install, configure, and make available data science nodes on Pinnacle | |
| **Activity 3:** Install and configure data science cluster to work with existing resources at UA, UAMS, UALR resources | | Collect testbed specifications and software/platform needs | Create containerized Hadoop-based testbed for DC |
| **Objective 1.1.c** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Specify and purchase 100Gb switch | | Issue UAMS purchase order for 100 Gb switch | Receive 100GB switch at UAMS |
| **Activity 2:** Install 100Gb switch | | | Install and configure new 100GB switch |
| **Activity 3**: Establish ScienceDMZ at UAMS | Create UAMS CI Plan | Specify and acquire additional DMZ components | Establish and validate 100Gb link to UAF and integrated DMZ |

| | | | |
|---|---|---|---|
| **Objective 1.1.d** | **Specific Milestones** | | |
| | Y1 | Year 2 | Year 3 |
| **Activity 1:** Create/identify federated identify or other authentication mechanism for all sites that provides access to core ARP resources | | | Establish federated ID for all project participants |
| **Activity 2**: Setup dedicated GitLab repository | | | Create and publish document outlining GitLab user guidelines, minimum standards for code repositories, and best practices. |
| **Activity 3:** Setup Globus Data Management Services | | | Globus Data Management contract executed |

| | | | |
|---|---|---|---|
| **Activity 4:** Engage other research themes to develop research-specific training modules in e.g. Python, R, Git, HPC, Singularity | | Host 5 software carpentry workshops; Train 2 software carpentry instructors | Host 5 software carpentry workshops; Train 2 software carpentry instructors |
| **Activity 5:** Develop and deploy training materials for code sharing, large data transfer protocols | | Host 2 online ARP-specific training sessions | Host 2 online ARP-specific training sessions |

| **Objective 1.1.e** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1**: Identify the number and type (HIPAA, proprietary economic, CUI, etc.) of private and secure data sources that will be need to be accessed by DART researchers. | Collect research theme needs | | |
| **Activity 2:** Setup capacity for storing and managing CUI and HIPAA data at UAF | | Deploy restricted access storage | Draft user guidelines and policy |

**CI2, Visualization for complex data in diverse data-analytics application domains**. The objectives and activities for this goal have been impacted by numerous personnel losses, leaving the team unable to complete the activities as initially planned. This goal will be revised based on current resources and expertise in the pending update to the project's strategic plan. Due to the disruption of the team, no progress was made on this goal and the tables are not included below.

## Data Curation and Life Cycle Analysis

The milestones and objectives for Year 3 are all complete or will be complete by the end of the year. The DC team has successfully established the clean ESKAPE genome database. Se-Ran Jun & Zulema Dominguez were awarded Best Representation of Informatics Supported Research, at the UAMS TRI Summer Writing Challenge. Below is a summary of specific accomplishments and the stoplight tables for the team.

**DC1, Automate heterogeneous data curation.** Objective 2.1.a. milestones for Year 3 are all complete or will be complete by the end of Year 3. Using large-scale language models, we can create knowledge graphs based on the data collection objectives and the selected datasets. This would allow for knowledge discovery in supervised contexts and extrapolation of this knowledge in unobserved datasets. Using Bayesian Neural networks, we have developed a need-based sampling methodology to train surrogate models for all-terminal network reliability estimation. We are extending this framework to other supervised learning contexts by changing the objective functions and data preprocessing pipelines. Some activities under Objectives 2.1.b. and 2.1.c. are slightly behind schedule, though we anticipate catching up by Year 4. We

completed the development of a new data imputation method that combines statistical modeling and machine learning for missing values recovery, a resulting manuscript was submitted for publication. We employed linear and nonlinear dimensionality reduction methods, such as PCA, tSNE, to visualize and reveal distinct clusters in high dimensional datasets.

| **Objective 2.1.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Define metrics for data quality to measure impact of unsupervised data cleansing on data standardization and reference clustering | Define at least one metric for completeness, standardization, and clustering quality of unstandardized reference data; Design and implement an unsupervised algorithm for each metric | Design and implement an algorithm using ML or Graph techniques for **one** metric | Design and implement a scalable algorithm in HDFS for **one** metric |
| **Activity 2:** Set baseline data quality for initial test datasets used in prior research and acquire additional test datasets | Establish baseline quality using supervised methods for existing datasets; Compare results of unsupervised quality metrics developed in Activity 1 to supervised results | Add 5 new person and 5 new business reference datasets for testing, at least 2 real-world; Add 5 new product reference datasets | Add 3 new person and 2 new business reference datasets with more than 1 million records for testing HDFS code |
| **Activity 3:** Curate test datasets and make available to other researchers | Establish a repository for the reference datasets and make available to other researchers | Add new reference datasets to the repository, as needed | Add new reference datasets to the repository, as needed |
| **Activity 4:** Develop a framework for collaborative data collection and cleansing for knowledge discovery | Formulate a hierarchical and as-needed data collection and cleansing strategy | Refine the formulation by including various practical constraints and test on small-scale problems; Formulate a collaborative data collection strategy involving multiple teams | Solve large-scale problems by considering a tree-based tool for data clustering and cleansing |

| | | | |
|---|---|---|---|
| **Activity 5:** Develop a need- and prediction-based feedback mechanism for future data collection and making scalable decisions | Formulate a framework for sequential data collection on an as-needed basis; Refine the formulation by including various practical constraints and test on small-scale problems | Formulate a Bayesian framework for sequential data collection based on predictive models; Investigate analytical approaches for using large datasets for different levels of decision making | Refine the Bayesian framework for sequential data collection based on predictive models and medium-sized data |
| **Objective 2.1.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Improve the unsupervised frequency-based data cleansing method used in prior POC; Explore and test alternative methods and models for unsupervised data cleansing including ML, AI, and graph approaches | Document and train team on data cleansing methods developed in prior research<br>-- Design and implement in Python or Java improvements to the prior frequency-based approach | Design and test an ML or Graph implementation to the prior frequency-based approach<br>-- Design and test new techniques for unsupervised data cleansing in Python, Java, ML, or Graph | Continue to design and test new techniques for unsupervised data cleansing in Python, Java, ML, or Graph |
| **Activity 2:** Migrate successful data cleansing models into a scalable process | | Refactor and migrate prior frequency-based approach into a scalable process | Refactor and migrate most successful of new data cleansing techniques into a scalable process |

| | | | |
|---|---|---|---|
| **Objective 2.1.c** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Improve the unsupervised frequency-based data integration method used in prior POC and explore and test alternative methods and models for unsupervised data integration including ML, AI, and graph approaches | Document and train team on reference clustering method developed in prior research; Design and implement in Python or Java improvements to the prior frequency-based approach | Design and test an ML or Graph implementation to the prior frequency-based approach; Design and test new techniques for unsupervised data cleansing in Python, Java, ML, or Graph | Continue to design and test new techniques for unsupervised data cleansing in Python, Java, ML, or Graph |

| | | | |
|---|---|---|---|
| **Activity 2:** Migrate successful reference clustering models into a scalable HDFS processes | | Refactor and migrate **prior** frequency-based approach into a scalable HDFS process | Refactor and migrate most successful of new reference clustering techniques into a scalable HDFS processes |

**DC2, Explore secure and private distributed data management.** Significant progress has been made under Objective 2.2 since the last report. The proof of concept for the positive data control system was successfully validated, and the team is incorporating some feedback from the project's EAB and RSV reports to update the trajectory of this effort and expand the focus to include data governance and other aspects of the data life cycle. The original Year 3 milestone will be updated in the strategic plan revision based on this feedback.

| **Objective 2.2** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Build a POC and demonstration code for a Positive Data Control system layer forcing all of the tools read/write operations to synchronize with the platforms metadata tool | | -- Setup a test platform with at least one processing function (e.g. Hive), metadata function (e.g. Atlas), and security function (e.g. Ranger);<br>-- Build POC with a simple PDC layer where Hive user is forced to go through PDC layer for all read/write operations | Modify POC to synchronize Hive operations with metadata layer (Atlas) and security permissions (Ranger) |

**DC3, Harmonize multi-organizational and siloed data.** All of this goal's milestones and activities are generally on schedule, and are complete or will be by the end of Year 3. Progress under Objective 2.3.a. includes the development of a clean ESKAPE genome database using reference free genomic data cleaning methodology (called GRUMP). This database will be updated every 6 months, and a manuscript was published in BioRxiv. For 2.3.b., we published the pangenome structure for *Pseudomonas aeruginosa*. We have generated about 200 ESKAPE genomes collected in Arkansas (UAMS and ADH) and published three resulting papers. We will add 11 more isolates by end of Year 3. All Arkansas genomic pathogens are annotated with species information obtained by comparing them with known pathogens at the global scale. Finally, a collaboration with LP faculty at A-State and the DC team at UAMS resulted in major progress under 2.3.c. We explored using XGBoost for antibiotic resistance prediction focusing on Enterococcus faecium and vancomycin. We developed a well-tuned application which is better or at least comparable to the existing ML approaches. We also developed and published an approach that integrates transcriptomes, protein-protein interactions and drug-protein interactions to study drug resistance in CML (chronic myelogenous leukemia), as well as a

machine learning model to infer transcription factors of target genes based on transcriptomics datasets.

| **Objective 2.3.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Define and download datasets to be curated | Build genomics database, including quality scores, gene/protein annotation | Extend to proteomics database - all for fast characterization of proteins (links to SwissProt) | Update databases (every 6 months) |
| **Activity 2:** Optimize data storage and retrieval | Use Elastic Cloud Storage for fast retrieval | Develop integrated database for proteomics & genomics, including annotations | Update databases (every 6 months) |
| **Activity 3:** Develop visualization methods | Prototype of R-BioTools for visualizing genomes | Publish one (1) R-BioTools paper for visualizing genomes | Develop visualization methods for very large trees |
| **Objective 2.3.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Develop pan-genome and Pan-proteome databases | Develop architecture / structure for rapid storage/retrieval of taxa-specific pan- and core-genomes | | Develop rapid storage/retrieval for core- and pan-proteomes |
| **Activity 2:** Develop taxonomy links to downloaded genomes/proteomes | Compare duplicate, known type strain genomes using ANI, Mash, 16S rRNA | Use Mash and other methods to assign nearest neighbors in phylogenetic space. | Update taxonomy (every 3 to 6 months) |
| **Activity 3**: Develop a genomic database for Arkansas genomic pathogen surveillance of antimicrobial resistance | | | Build Arkansas pathogen database, with links to known pathogens; develop GPU-based methods for fast calculation of genomic distances |
| **Objective 2.3.c** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Define training sets to be used for ML | Identify key datasets and problems for ML | Develop ML models for known toxins | Develop ML models for antibiotic resistance |

| | | | |
|---|---|---|---|
| **Activity 2:** Integrate multi-omic models for ML | Integrate genomic / microbiome / taxonomy datasets (petabytes) | Integrate genomic, transcriptomic, proteomic, and metabolomic datasets (petabytes) | Integrate model organisms (mouse, rat, human, yeast, etc.) as well as microbial |
| **Activity 3**: Benchmark ML results | | Develop Benchmarking Standards for ML of pathogens | Publish one (1) ML Benchmarking papers |

## Social Awareness

The SA team experienced the highest number of personnel changes, and will submit updated milestones and activities for the remainder of the project in the coming weeks. The overall goals and objectives will remain the same, but specific deliverables and research areas have shifted a bit due to changes in the experience of the team. In addition, the team plans to consolidate some of the activities to reduce granularity, and incorporate feedback from the EAB and RSV panelists who encouraged more crossover with the SM team. Aside from the milestones that were dependent on lost personnel, the team is on track and making progress. All of the remaining milestones are completed on schedule or are scheduled to be completed by the end of the reporting year. Below is a summary of specific accomplishments and the stoplight tables for the team.

**SA1, Privacy-Preserving and Attack Resilient Deep Learning.** Objective 3.1a activities and milestones were all completed since the last report. The Year 2 and Year 3 milestones for Objective 3.1b have also been completed. For Activity 1 and 2, the team explored tradeoffs among privacy, resilience, and utility in deep learning models. Deep learning models such as graph neural networks are susceptible to privacy inference attacks, given their ability to learn joint representation from features and edges among nodes in graph data. To prevent privacy leakages, we developed a novel heterogeneous randomized response mechanism to protect nodes' features and edges against privacy inference attacks under differential privacy guarantees. Our idea is to balance the importance and sensitivity of nodes' features and edges in redistributing the privacy budgets since some features and edges are more sensitive or important to the model utility than others. As a result, we achieved significantly better randomization probabilities and tighter error bounds at both levels of nodes' features and edges departing from existing approaches, thus enabling us to maintain high data utility for training graph neural networks. We also proposed a learning framework that can provide node local differential privacy at the user level, while incurring low utility loss and apply randomization mechanisms to perturb both feature and label data at the node level before the data is collected by a central server for model training. Specifically, we investigated the application of randomization mechanisms in high-dimensional feature settings and propose a reconstruction-based protocol with strict privacy guarantees. We also formulate this learning framework to

utilize frequency estimates of graph clusters to supervise the training procedure at a sub-graph level. Finally, we conducted an extensive theoretical and empirical analysis using benchmark datasets and results show that our approach significantly outperforms various baselines in terms of model utility under rigorous privacy protection for both nodes' features and edges. No milestones are scheduled for Objective 3.1.c.

| **Objective 3.1.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Research existing attacks including model inversion attacks and data poisoning attacks and capture mechanisms behind the threat models | Document literature research of attack models and mechanisms behind attacks. | | |
| **Activity 2:** Study the potential risks due to correlations among input data features, parameters, output, target victims, and latent feature space in deep learning algorithms | Initiate theoretical investigation on the risks of deep learning algorithms | Disseminate the findings of both theoretical and empirical studies on risks of deep learning algorithms | |
| **Activity 3:** Study the sensitivity and impact of input data features, parameters, and the objective functions on the model output and identify appropriate differential privacy preserving mechanisms for different computational components in a variety of deep learning models | Initiate theoretical investigation of privacy preserving mechanisms. | Disseminate the findings of both theoretical and empirical studies on privacy preserving mechanisms used for deep learning algorithms | |

| **Objective 3.1.b** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Investigate the tradeoff of achieving privacy, resilience to adversarial attacks, and utility | | Research the tradeoff of privacy, resilience, and utility. | Complete both theoretical and empirical studies on privacy, resilience, utility tradeoff in the deep learning setting. |
| **Activity 2:** Study the mechanisms of redistributing injected noise across input data features, model parameters, and coefficients of objective functions based on their vulnerability and impact on the model output | | Examine the noise redistribution mechanism. | Complete both theoretical and empirical studies on the noise redistribution mechanism in the deep learning setting. |
| **Activity 3:** Develop and implant threat- and privacy-aware deep learning models | | Design algorithms of threat- and privacy-aware deep learning models | Complete initial implementation |

**SA2, Socially Aware Crowdsourcing.** Progress was made towards milestones and activities for Objectives 3.2.a and 3.2.b, some Year 2 milestones were completed since the last report. The team plans to make some minor revisions to the Year 3 milestone language, particularly 3.2.c, and will report on that next year.

| **Objective 3.2.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Allow uncertain labels in crowdsourcing data collection | Selected the approaches through literature review | Implemented and tested | |
| **Activity 2:** Aggregate raw labels after label collection | Computational schemes are identified | Implemented and tested | |
| **Activity 3**: Filter out possible noises to further improve data quality | Identified possible sources of noises | Filtering algorithms designed | Algorithms implemented and tested |
| **Objective 3.2.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Build theoretic foundations | Specified mathematical requirements | | |
| **Activity 2:** Develop learning models and inference algorithms | | Algorithms designed to meet specification | Algorithms implemented and tested |
| **Activity 3**: Test and apply these learning models and algorithms | | Testing dataset selected | Initial tests completed |
| **Objective 3.2.c** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Establish additional evaluation metrics | | | Quality metrics established |
| **Activity 2:** Develop algorithms to calculate the metrics | | | Quality metrics are quantified |
| **Activity 3**: Verify and validate computational results | | | |

**SA3, User-centric Data Sharing in Cyberspaces.** Progress was made towards milestones and activities for Objectives 3.3.a and 3.3.b, some Year 2 milestones were completed since the last report. The team plans to make some minor revisions to the Year 3 milestone language, particularly 3.3.c, and will report on that next year.

| **Objective 3.3.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Research state-of-art entity identification techniques for non-structure data | Document and disseminate the findings on personal identifying information and their privacy issues | | |
| **Activity 2:** Investigate appropriate techniques for identifying context-aware sensitive information | Identify and disseminate the findings on the sensitivity of information in different context | Identify appropriate techniques for identifying context aware sensitive information | |
| **Activity 3:** Develop appropriate text analysis techniques to identify sensitive information from unstructured data | Research appropriate text analysis techniques to identify sensitive information from unstructured data | Develop appropriate techniques for identifying sensitive information from unstructured data | |

| **Objective 3.3.b** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Research state-of-art multimodal deep learning techniques for identifying private sensitive information | Study and document state of art multimodal deep learning techniques | Document and disseminate the findings on the determination of appropriate multimodal techniques for detecting sensitive information | |
| **Activity 2:** Investigate appropriate techniques for identifying discriminating and stigmatizing information | | Document and disseminate the findings of state-of-art techniques for identifying discriminating information | Document and disseminate the findings on the determination; Development of appropriate techniques for identifying stigmatizing information |
| **Activity 3:** Develop appropriate deep learning text analysis techniques to accurately remove discriminating and stigmatizing information | | Design deep learning techniques for removing discriminating and stigmatizing information; Implement deep learning techniques for removing discriminating and stigmatizing information | Test and document the efficiency of the techniques and improve them when necessary |

| **Objective 3.3.c** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |

| | | | |
|---|---|---|---|
| **Activity 1:** Develop appropriate techniques for monitoring personal information disclosure on the Internet such as those from government records, news reports, and online documents | | | Research the key ventures for sharing/publishing/releasing PII |
| **Activity 2:** Develop a risk assessment method for possible privacy breach given the amount of personal identifying information disclosed/published | | | Research the relationship among PII attributes and how the release of one attribute would affect the overall privacy |
| **Activity 3:** Develop appropriate techniques for safeguarding sensitive information by helping end users monitor and proactively control the release of their personal information | | | |

**SA4, Deep Learning for Preventing Cross-Media Discrimination.** The milestones for Objective 3.4.a, Years 1 and 2 have been completed. We conducted a literature study on hate speech detection in multimodal publications. We developed a testbed for evaluating recently proposed unimodal and multimodal models for hate speech detection. The performance of these baseline models is evaluated in three different settings: (1) image plus tweet text plus image text; (2) image plus tweet text; and (3) image plus image text. One technical report was produced. Other milestones for this goal and objective will be modified in the upcoming revision of the strategic plan. Progress has been made under Objective 3.4.b, though some minor changes to the remaining milestones will be submitted in the revision as well. We studied the coded hate speech detection problem, where coded words have been used to represent the targeted groups in hate speech to evade detection. We develop a coded hate speech detection framework, called CODE, to detect hate speech by judging whether coded words like Google or Skittles are used in the coded meaning or not. Based on a proposed two-layer structure, CODE is able to detect the hateful texts with observed coded words as well as newly emerged coded words. Experimental results on a Twitter dataset show the effectiveness of our approach. One paper was produced and published at PAKDD 2022. Finally, the milestones for Objective 3.4.c will also see minor revisions in our pending requested changes to the strategic plan. Progress here includes a novel robust hate speech detection model that can defend against both word- and character-level adversarial attacks. We identified the essential factor that vanilla detection models are vulnerable to adversarial attacks as the spurious correlation between certain target words in the text and the prediction label. To mitigate such spurious correlation, we described

the process of hate speech detection by a causal graph. Then, we employ the causal strength to quantify the spurious correlation and formulate a regularized entropy loss function. We show that our method generalizes the backdoor adjustment technique in causal inference. Finally, the empirical evaluation shows the efficacy of our method. One paper was produced and published at AACL-IJCNLP 2022.

| **Objective 3.4.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Use deep convolutional neural networks (CNN) to recognize discrimination-sensitive objects from images | Initiate theoretical investigation on using CNN to recognize discriminatory objects | Complete exploration and comparison of different multimodal hateful image-text detection models | |
| **Activity 2:** Adopt long short-term memory (LSTM) network to model the text | Initiate theoretical investigation on using LSTM to model discriminatory text | Complete design and implementation of the LSTM-based model | |
| **Activity 3:** Utilize bilinear model to capture the implicit relationship between the detected discrimination-related objects and the text | | Initiate the theoretical investigation on the implicit relationship between the detected discrimination-related objects and the text | Complete design and implementation of the bilinear model |

| **Objective 3.4.b** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Adopt mixture Generative Adversarial Nets framework for generating perceptually similar and discrimination-free image patches | | | Initiate the theoretical investigation on using GAN to generate discrimination-free image patches |
| **Activity 2:** Applies the encoder-decoder mechanism to automatically generate the discrimination-free text based on the original text | | | Complete preliminary theoretical investigation on using encoder-decoder to generate discrimination-free text |

| Objective 3.4.c | Specific Milestones | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 2:** Test and evaluate the proposed techniques and models from available data sources in social networks like Facebook, Instagram, and Foursquare | | -- Complete social media data collection<br>-- Compete evaluation of CNN and LSTM models | Complete evaluation of the bilinear model |

**SA5, Marketing Strategy Design with Fairness.** The milestones for this objective will be updated in the strategic plan revision, due to personnel changes and RSV feedback. The loss of Zenghui Sha left an expertise gap we were not able to match. Prior to the departure progress had been made against Year 1 and 2 milestones, so the changes requested will impact the milestones for Years 3-5 and will incorporate additional crosslinks to other SA projects and the SM team.

| Objective 3.5.a | Specific Milestones | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Data collection from social media data, e.g., Amazon and Facebook, using vacuum product as the application context | Document and disseminate the findings of literature research and evaluation of the target products for case study | Document and disseminate the findings of data collection from social media including both review data and production information | |
| **Activity 2:** Data collection and analysis of consumer panel data from Nielsen datasets at the Kilts Center for Marketing | Document and disseminate the findings of processing the data from Nielsen datasets and extract the information needed (e.g., demographics, product market segment, etc.) for this project | Document and disseminate the findings of the data analysis obtained from Nielsen dataset to complement the data from social media | |
| **Activity 3:** Collection and analysis of marketing cases and/or ads with unfairness and exclusions | | Document and disseminate the findings of the analysis of unfair marketing cases and extract the features/forms of biases | |

| | | | |
|---|---|---|---|
| **Activity 4:** Text mining and sentiment analysis of the collected data for the development of metrics of fairness in marketing, and the identification of customer-descried product features | | Document the text mining and sentiment analysis for the data collected from social media | Establish and publish the method to identify the deciding factor/features that influence customers' judgement |
| **Objective 3.5.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Quantification and rating of bias and unfairness of marketing strategies and relate it to customer-desired product features | Document and disseminate the findings of researching the existing methods for fairness quantification and adverting parameterization | Define and document the quantification methods for the features identified from the data collected | Establish and publish the association/mapping between the marketing features and the product features |
| **Activity 2:** Network-based approach for choice modeling by incorporating customer preferences and perception to marketing (e.g., price) (un)fairness | | Define and document the network-based model for choice prediction and demand forecasting | Incorporate customer-related attributes into the modeling |
| **Activity 3:** Validation of the network-based choice modeling via demand prediction | | | Validating the model with simulation studies on choice prediction |
| **Objective 3.5.c** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Parameterize marketing strategies and incentive design for improved advertisement with fairness consideration | | Document and disseminate the findings of the literature study on computational marketing and computational advertising | Create a computational design approach for marketing strategies with the inclusion of fairness parameters |
| **Activity 2:** What-if scenario analysis with the simulation of recommended marketing strategies and quantify the impact of those strategies on brand value and demand | | | Finish the model testing with simulation studies |
| **Activity 3:** Validation through human-subject experiment and surveys | | | Create experimental protocols and the design of experiment; finish the pilot studies |

**SA6, Privacy-Preserving Analytics in Health and Genomics.** Some activities for Objectives 3.6 were recently delayed due to personnel changes, and some milestones will be updated in the revision of the strategic plan. This project was primarily driven by Xiuzhen Huang. Initial progress was made, but the replacement for Huang does not have the same expertise and instead will contribute to the DC team.

<table>
<tr><td rowspan="2">Objective 3.6.a</td><td colspan="3">Specific Milestones</td></tr>
<tr><td>Year 1</td><td>Year 2</td><td>Year 3</td></tr>
<tr><td>Activity 1: Design and develop machine learning and deep learning algorithms and software for privacy-preserving data analytics; Data and Infrastructure request and preparation</td><td colspan="2">Document and disseminate the findings of literature research of privacy-preserving data analytics alights and software</td><td></td></tr>
<tr><td>Activity 2: Develop privacy-preserving analytics algorithms, which will be based on high-dimensional tensor mathematical optimization model and combinatorial models</td><td colspan="3">Initiate investigation on mathematical optimization models</td></tr>
<tr><td>Activity 3: The optimization models will be incorporated with machine learning and deep convolutional neural network models</td><td></td><td colspan="2">Document and disseminate findings on theoretical investigation of privacy preserving algorithms</td></tr>
</table>

**SA7, Cryptography-Assisted Secure and Privacy-Preserving Learning.** Some activities for Objectives 3.7.a. and 3.7.b. were recently delayed due to personnel changes, and some milestones will be updated in the revision of the strategic plan. A privacy-preserving distributed learning scheme named divide-and-conquer learning (DCL) was designed that allows a resource-limited user to offload neural network model training to multiple distributed participants. The work was published in the 2022 ACM/IEEE Symposium on Edge Computing (SEC), Workshop on Edge Computing and Communications. A minor deviation from the original plan is that blockchain was not used in the solution, since we found that the communication cost would be too high if blockchain were used. The Year 3 milestone will be removed and two new milestones will be added to Year 4 (Empirical study of privacy-preserving and secure learning solutions) and Year 5 (Analytical study of privacy-preserving and secure learning solutions). Under 3.7.b., we designed a privacy-preserving machine learning model training and classification method for face recognition-based access control systems. It allows training of model over perturbed facial feature data and also classification of data sample over perturbed form to protect face privacy. The work generated a paper submission to the IEEE PerCom 2022. No milestones were scheduled for Year 3 under 3.7.c., and some minor updates to this will be requested.

| Objective 3.7.a | Specific Milestones | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Research the hybrid use of existing cryptography techniques and differential privacy in federated machine learning | A survey of existing cryptography techniques and their applications in differentially private federated learning | | |
| **Activity 2:** Develop new applied cryptography techniques to use in combination with differential privacy for federated machine learning | Design of preliminary new cryptography techniques used for differentially private federated learning | Design of blockchain-based private distributed learning | |
| **Activity 3:** Develop unified security models for theoretical analysis of hybrid solutions | | Preliminary united security models for analysis of hybrid solutions | United security models for analysis of hybrid solutions |
| **Objective 3.7.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Develop methods for building/perturbing the model so that it can respond to encrypted or perturbed classification input | | Methods for building/perturbing model to support perturbed classification input | Preliminary methods for building/perturbing model to support perturbed classification input |
| **Activity 2:** Study whether and how differential privacy can be achieved for classification input | | | Complete exploration of whether/how differential privacy can be achieved for classification input |

## Social Media and Networks

SM Co-lead Agarwal received a number of Best Paper awards at conferences during this reporting period. Progress against the strategic plan milestones and objectives for SM activities are included below.

**SM1, Mining cyber argumentation data for collective opinions and their evolution.** While progress was made under this goal early in the project, the loss of Justin Zhan towards the end of Year 2 halted these efforts, and progress towards the Year 2 and 3 milestones has not kept pace. After identifying Zhan's replacement as Susan Gauch, the team is working to update the activities and milestones for the SM1 goal which will be requested in the upcoming strategic plan revision. This goal is one area that was also discussed by RSV panelists as a good target for integration with the SA team, which we plan to pursue. The milestone tables for this goal are not included below, as there are no additional updates to report at this time.

**SM2, Socio-computational models for safer social media.** Planned milestones for Objectives 4.2.a were all completed on schedule, though some minor changes will be submitted for revision due to the loss of Justin Zhan. Progress here includes development of a taxonomy to characterize social media (including multimedia rich online information environments) which was published.

Planned milestones for Objectives 4.2.b were all completed on schedule. We reviewed and collected various cyber campaigns datasets including 2021 US Capitol riots; COVID-19 anti-mask, anti-lockdown, anti-vaccine campaigns, 2023 Brazil Capitol riots, 2023 Peru protests, and other cyber campaigns in the Indo-Pacific region. Data was published in usable online tools viz., blogtracker, vtracker, and COVID-19 misinformation tracker, publicly available at www.cosmos.ualr.edu. Behavioral traits of individuals and groups were identified (collective identity emergence, mobilization, and network organization) and published. One of the publications about these findings received top 25% full paper award at AMCIS 2022. A model was developed to identify contextual focal structures (CFSA) that are fundamental to coordinating/mobilizing cyber campaigns. Research findings were published in the leading social network journals and mathematical organization theory journal. Finally, a refined model was explored, which considers a multiplex network representation of individuals and contexts to offer explainability and interpretability of the findings.

For Objective 4.2.c, Year 3 milestones are complete. Several campaigns have been identified that utilize multimedia rich content for coordination and mobilization including 2023 Brazil Capitol riots and 2023 Peru protests. Multimedia-intensive platforms under study include YouTube, TikTok, and Instagram along with text-intensive platforms such as Twitter. Identified and documented newer forms of TTPs on multimedia rich social platforms increasingly gaining traction in coordination/mobilization of cyber campaigns, particularly of deviant/adversarial nature. Analytics are being developed to computationally detect and measure the impact of such TTPs. This is a year 3 + 4 activity. The activity is on schedule according to the strategic planning document. No milestones have been scheduled for Objective 4.2.d during this reporting period.

| **Objective 4.2.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Study social media spaces and cyber campaigns to identify characteristics and features | Social media platforms identified; Cyber campaigns identified; Characteristics and features identified | | |
| **Activity 2:** Create a taxonomy of dimensions to characterize social media spaces | Taxonomy developed | | |

| | | | |
|---|---|---|---|
| **Activity 3**: Revisit and adjust taxonomy as social media space evolves | | Revised taxonomy developed and published based on new social media, campaigns, features, and characteristics | |
| **Objective 4.2.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Review cyber campaigns and social media data | Data sources identified; Data acquisition procedures established Database setup | Data reviewed and modifications incorporated; Data collected and shared with DART teams | Cyber campaign data reviewed, additional data collected if needed; Data published according to NSF's data sharing policies and social media platforms' terms and agreements |
| **Activity 2:** Identify behavioral traits for key actors and key groups by leveraging OIE characterization | | Key actors and key groups identified empirically | Behavioral traits of key actors and key groups identified |
| **Activity 3**: Develop computational model(s) for key actor and key group discovery | | Model(s) developed | Model(s) refined; Published in peer reviewed forums; Model transitioned to usable web-based application |
| **Activity 4:** Evaluate model(s) | | | Model(s) evaluated and refinements proposed |

| **Objective 4.2.c** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Review campaigns, social media platforms, and involved actors and groups | | | Campaigns identified for further review |
| **Activity 2:** Identify and document tactics, techniques, and procedures (TTPs) (e.g., platform orchestration, botnets, inorganic behaviors, stalking, pacing, leading, threadjacking, hashtag latching, boosting, echo chambers) | | | TTPs identified and documented/published |

**SM3, Auto-annotation of multimedia data**. The activities and milestones for this goal are generally on schedule, despite some minor delays due to hiring and human resources issues for graduate students working on this effort. Multimedia data characteristics towards target applications have been identified for Objective 4.3.a. Activity 2 is slightly delayed, with planned completion over the next few months. Under Objective 4.3.b., activity 2 is also scheduled for

completion in Fall 2023. Progress here includes integration of learning objectives from SM4 into disaster management scenarios. Results from SM4 have also been incorporated into the activity for Objective 4.3.c, which will be completed during Year 4.

| **Objective 4.3.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Define priorities and characteristics for multimedia data on social platforms | Key characteristics defined | | |
| **Activity 2:** Design and build algorithms for efficient retrieval of nontraditional data | | Image, video and 3D retrieval methods defined and tested | |
| **Objective 4.3.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Define learning objectives for social data from multimodal sources | Identify and define three major learning objectives document | | |
| **Activity 2:** Develop detection and classification methods | | Object and event detection methods implemented | |
| **Activity 3**: Deep learning applied to multimedia data and related indexing mechanisms | | | Two learning methods developed and tested |
| **Objective 4.3.c** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Define key applications for the implementation and testing of the indexing and retrieval mechanisms | Three key applications defined | | |
| **Activity 2:** Integration with disaster response and other applications defined in Activity 1 | | | Integration methods developed |
| **Activity 3**: Define ethical and legal perspectives for the use of multimedia data | | **Delayed**: Use and access-based ethics principles defined for multimedia social data | |

**SM4, Informing disaster response with social media.** As noted in the paragraph above for SM3, some of the cross-linked activities between these two goals were slightly delayed, but the team plans to catch up and does not intend to revise anything in the strategic plan. Under Objective 4.4.a., multimedia data from two major SM sources have been identified, collected and analyzed. Two major scenarios one based on water depth estimation, and another based on speed estimation have been covered in publications. The team is building the methodology to index the transportation infrastructure belonging to the identified relevant categories, for a

given geographical area. The current approach uses publicly available records from official repositories such as data.gov and alternative private or crowdsourced repositories such as Google Street View and OpenStreetMap databases. We have obtained data for Hurricane Harvey, including satellite imagery from Planet as a proof of concept, and social media data from Twitter. The team has done experiments to combine multiple data sources and modes into valuable insights regarding the status of road conditions and disaster relief requirements. This encompasses text, image, and video data. This is still an ongoing effort under Objective 4.4.b.

The team has evaluated multiple approaches for using images sourced from social media for identification of flood afflicted areas. This includes building Deep Learning classifiers and hierarchical flood models that determine the relevancy of said images and evaluate the quality of the prediction by using Bayesian methods. Using the developed Deep Learning and Bayesian models, we can assign scores for assessing road conditions conditional to the observed evidence from social media. We will develop a simulation environment that uses Multi-Agent Deep Reinforcement Learning to develop optimal policies for post-disaster assistance under the identified constraints and relevant stochastic conditions.

The team will complete most of the Objective 4.4.c. milestones by the end of this reporting year. A journal article synthesizing qualitative data will be submitted in summer 2023. Literature reviews for identified routing problem variants are included in a dissertation being completed in May 2023. Routing algorithms were developed, validated and tested for single and multiple agent and destination variations of a routing problem on an uncertain and disrupted network. For activity 4, the milestone will be revised, and the team will focus on proof-of-concept testing of the overall framework and methodologies that comprise it, rather than displaying the outputs from those methodologies in a GIS testbed. Some Year 3 milestones under this goal will be updated to reflect these changes.

| **Objective 4.4.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Study social platforms to identify types of content that describe transportation infrastructure status | Identify and define social platform content types of interest (e.g., image, video, text, etc.) | | |
| **Activity 2:** Develop and implement extraction techniques for identified types of social platform content | | Develop social platform extraction techniques for content types of interest and pilot test on at least two disaster scenarios | |

| | | | |
|---|---|---|---|
| **Activity 3**: Develop and implement indexing techniques for extracted social platform content | | Develop and implement indexing techniques for extracted social platform content and pilot test on at least two disaster scenarios | |
| **Objective 4.4.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Identify other data sources that contain real-time information regarding transportation infrastructure status | Identify and define content types of interest (e.g., satellite imagery, traffic cameras) from sources other than social platforms | | |
| **Activity 2:** Obtain and index transportation infrastructure data from other data sources | | Obtain and index identified content types for at least two disaster scenarios | |
| **Activity 3**: Develop and implement data fusion techniques to combine data from social platforms and other sources | | | Fuse data from social platforms and other sources for at least two disaster scenarios |
| **Objective 4.4.c** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Develop and implement machine learning classifiers to detect quality of information | Obtain testing data from social platforms | Develop machine learning classifiers to detect false or low-quality information | |
| **Activity 2:** Develop and implement schema to map credibility/quality scores for data to probabilistic inputs of transportation infrastructure status | | | Develop schema for mapping each datum to a probability describing its credibility |
| **Objective 4.4.d** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Identify critical routing problems with application in disaster response | Select at least two disaster response routing problem variants using existing qualitative interview data | | |
| **Activity 2:** Develop models of identified disaster response routing problems and assess state of the literature | Conduct literature review for identified routing problem variants and publish journal article synthesizing review with qualitative data from 4.4.c.1 | | |

| | | | |
|---|---|---|---|
| **Activity 3**: Develop and implement routing algorithms for identified routing problem variants | | For at least two routing problem variants, develop, validate and test at least one solution algorithm each on randomly generated test networks | |
| **Activity 4**: Implement GIS testbed capable of displaying and analyzing real-time road status and routing algorithm outputs | Define GIS system requirements | Develop GIS system to display real-time road status inputs | |
| **Activity 5**: Demonstrate models and solution approaches via pilot study of one or more disaster scenarios | | Select one or more disaster scenarios for pilot | Obtain test data for selected disaster scenario |

## Learning & Prediction

Emre Celebi and colleagues guest-edited a special issue entitled "Skin Image Analysis in the Age of Deep Learning" in the IEEE Journal of Biomedical and Health Informatics, one of the most prestigious journals in its field, with an impact factor of 7.021; and another special issue entitled "Image Analysis in Dermatology" in the Medical Image Analysis journal, with an impact factor of 8.545. The issue received 74 submissions from 29 countries. Twelve articles were accepted for publication, resulting in a 16% acceptance rate. Ahmad Al-Shami received the Faculty Excellence award for research at Southern Arkansas University. Below is a summary of specific accomplishments and the stoplight tables for the team. LP Seed Grant Recipient Han Hu received the 2023 Faculty Gold Medal award at UARK, as well as the MEEG Outstanding Service Award.

**LP1, Statistical Learning – Random Forests for Recurrent Event Analytics**. Milestones for all objectives under goal 5.1 are on schedule, and the team does not intend to revise anything in the strategic plan. A random forest-based model for recurrent event data has been established, and the model was implemented in R. A paper was submitted for publication describing the model. Additionally, a gradient-boosted tree-based model for recurrent event data has been established. No milestones were scheduled for Year 3 in many of the 5.1 activities.

| **Objective 5.1.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Establish a preliminary model, and complete the theoretical investigation | Complete the preliminary theoretical investigation on the proposed modeling approach | | |

| | | | |
|---|---|---|---|
| **Activity 2: C**omplete the coding and numerical examples; write, submit, revise paper | | Complete the numerical studies, and submit a research paper | |
| **Activity 3:** Revise paper and research outcomes dissemination through conferences | | | Complete the model improvement and paper revision |
| **Objective 5.1.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Establish a preliminary model, and complete the theoretical investigation under Obj 5,2 | | | Complete the preliminary theoretical investigation on the proposed modeling approach |

**LP2, Statistical Learning – Marked Temporal Point Process Enhancements via Long Short-Term Memory Networks.** The approach for 5.2.b has focused on remaining time until failures predictions using existing technique. This shift has allowed us to begin benchmarking our work against other RNN architectures and leverage results available via common data sets (C-MAPSS). Under 5.2.c, the main work has been tied to a civil architecture dataset consisting of oil production equipment failure information. The efforts in this project to study this data has revealed interesting limitations in aggregated time-series datasets. We are using the lessons learned to study a NASA system dataset. Some minor changes will be requested in the strategic plan revision to the Year 3-5 milestones for this goal. Some Year 2 milestones were completed since the last report.

| **Objective 5.2.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Formally define approach integrating intensity function of MTTP into LTSM | Submit conference paper with initial model | | |
| **Activity 2:** Establish proof-of-concept implementation of MTTP/LTSM approach | Present conference paper with preliminary results of implementation | Submit journal article with conceptual findings and initial implementation of approach | |
| **Activity 3:** Perform benchmark of MTTP/LSTM tests on small simulated data sets | | Publish white paper and GitHub repository with benchmark tests/results | |

| Objective 5.2.b | Specific Milestones | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Assess data collected from Activities 1 and 2 of Objective 5.2c to define methodology computation performance requirements | | Produce system requirements document for V2 implementation | |
| **Activity 2:** Create version 2 implementation of approach using lessons learned from Activity 2 of Objective 5.2a | | Submitted conference paper | Present conference paper on V2 implementation progress |

| Objective 5.2.c | Specific Milestones | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Acquire healthcare IoT datasets | Publish curated data to GitHub | | |
| **Activity 2:** Acquire civil infrastructure datasets | Publish curated data to GitHub | | |
| **Activity 3**: Establish baseline performance of predictions made by existing approaches applied to datasets from Objective 5.2c Activities 1 and 2 | | Publish white paper and GitHub repository with benchmark predictions | Make conference presentation on baseline performance of predictions compared against existing approach |

**LP3, Deep Learning – Novel Approaches.** Olcay Kursun and Paul Schrader left the state, LP3 will have updates in the strategic plan revision to reflect personnel changes. Ahmad Al-Shami and Zachary Stine have been recruited to the team. Zachary Stine was previously supported by the project as a graduate student at SAU, and upon graduation received a faculty position at UCA. All milestones and activities for this goal are complete on time or will be by the end of the reporting year.

Under Objective 5.3.a, the team developed a novel self-supervised domain adaptation method in crowd counting and published two novel tools- Fairness Domain Adaptation (FREDOM) Approach to Semantic Scene Understanding and Self-supervised Spatiotemporal Transformers (SPARTAN) Approach to Group Action Recognition. These are available on GitHub. Under 5.3.b, the team published a research paper in IEEE ACCESS, detailing the development of novel features for classifying malware using various machine learning models, such as ensemble models and MLP. We published an image preprocessing model based on topological data analysis in the 26th International Conference on Image Processing, Computer Vision, & Pattern Recognition proceedings. A paper was submitted to Future Technologies Conference (FTC) 2023, on comparative analysis of object localization using topological data analysis, and a master's thesis will be published this summer on problem specific selection of

topological features and distances. Finally, under 5.3.c, multiple causal inference models have been studied and investigation is underway on their use in DRL with a thesis planned for completion this summer as well.

| **Objective 5.3.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Development of novel self-supervised and flow-based deep learning approaches | Development of the first unsupervised convolutional area and the first flow-based deep learning approach | Adaptation of the novel methods for application to the tactical agility dataset | Development of the areal postprocessing of the self-supervised model and the Flow Autoencoder as deep graphical model with dual objectives |
| **Activity 2:** Developing a library of classifiers for benchmarking | Development of linear dimensionality reduction methods | Development of the standard autoencoder method | Development of signal feature extraction tools |
| **Activity 3**: Application of the developed methods on real-world datasets | Application of the developed methods with applications on natural images and textures, and classification of a malware dataset | Comparisons of the developed methods/libraries on the tactical agility dataset and malware dataset | Exploratory studies with the datasets of the researchers in the DC thrust of DART. Evaluate various deep learning models on the malware dataset |

| **Objective 5.3.b** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Investigate group theoretical and topological properties of generalized neural network architectures | Investigate group theoretical approaches to generalized NN architecture design within the context of interpretability for Objectives 5.3a and 5.3c | Exploration of internal topologies in generalized NN structures using TDA and PH to identify architectures which address high dimensionality and enhance the developments in Objectives 5.3a and Objective 5.3c | Study of probabilistic spaces associated to generalized NN from the perspectives of group theory and topology for identifying structure that enhances inferential capabilities and address concerns in the activities of Objectives 5.3a and 5.3c (e.g., stochastic spiking) |

| **Objective 5.3.c** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Design an improved reward process for DRL | Development of a generalized model of reward function in DRL addressing the issues with both sparse and dense feedback | Development of use cases of the developed reward model | |

| | | | |
|---|---|---|---|
| **Activity 2:** Explore PH based filtering to optimize scenario space | | | Development of a PH based scenario-space optimization algorithm |

**LP4, Deep Learning – Efficiency and Specification.** Significant progress was made toward the 5.4 Objectives, with all milestones met for Year 3. We developed and published a number of novel and best-in-class self-supervised tools for a variety of applications, such as 3D capsule networks for medical segmentation on less labeled data, multimodality multi-lead ECG arrhythmia classification, image deblurring, and a new self-supervised domain adaptation deep learning method to deal with limited training data. Other accomplishments include a vision-language with contrastive learning (VLcap) for coherent video paragraph captioning, and a Visual-Linguistic Transformer-in-Transformer (VLTinT) for Coherent Video Paragraph Captioning. We published a model for CLIP-assisted temporal self-attention for weakly-supervised video anomaly detection, and an amodal instance segmentation with transformer (AISFormer). Under 5.4.b., we implemented meta-learning of NAS for few-shot learning in medical imaging, and developed the (2+1)D Distilled ShuffleNet: A Lightweight Unsupervised Distillation Network for Human Action Recognition. A comprehensive review on spiking neural networks and their applications was conducted. Relevant milestones are included below.

| **Objective 5.4.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Develop and demonstrate new low-cost deep neural network algorithms. | Develop Teacher - Student Distillation Deep Learning Algorithms; Develop Light-weight Deep Learning Algorithms | Develop Deep Network Compression Algorithms; Develop Deep Network Pruning Algorithms | |
| **Activity 2:** Develop new objective loss functions in deep neural networks | | | Develop auto and semi-auto deep network searching algorithms to discover optimal deep neural networks given a particular application and data training sets. |
| **Objective 5.4.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Mathematically analyze the proposed deep learning methods | Develop analytic approaches to the proposed methods in Activities 1.1 | Develop analytic approaches to the proposed methods in Activities 1.2 | |

| | | | |
|---|---|---|---|
| **Activity 2:** Improve the computational time and accuracy performance | | | Improve the computational time and accuracy performance on the standard databases and challenges |
| **Objective 5.4.c** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** The developed deep learning algorithms will be optimized and implemented in two applications, including natural images and medical imaging. | Develop Low-cost Deep Learning Approaches in Image Classification | Develop Low-cost Deep Learning Approaches in MRI Segmentation | |
| **Activity 2:** The developed deep learning algorithms will be further optimized and deployed in high dimension data, such as: videos and medical MRI volumetric data | | | Develop Low-cost Deep Learning Approaches in Automatic Human Activity Recognition in videos |

**LP5, Harnessing Transaction Data through Feature Engineering.** We developed a temporal clustering optimization model to determine the optimal patient-specific time-window size for irregularly-sample multi-variate transaction data. Under 5.5.b., we integrated the optimal time-window size to LSTM and reinforcement learning models, and the performance is shown improved considering these patient-specific time windows. All activities and milestones are progressing on schedule and the Year 3 milestones are planned for completion by Fall 2023. No changes will be requested in the strategic plan revision.

| **Objective 5.5.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Extract and process APCD data | Obtain and prepare cleaned data for research | | |
| **Activity 2:** Extract and engineer features from the high-dimensional temporal data | Acquire features that are highly representative | | |
| **Activity 3**: Explore and test automation of feature engineering in transaction data | | Achieve automotive feature engineering | Optimize feature engineering with transaction data |
| **Objective 5.5.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |

| | | | |
|---|---|---|---|
| **Activity 1:** Develop deep learning prediction models and algorithms with feature engineering | Complete selection and testing of deep learning models | Improve the predictive models | |
| **Activity 2:** Incorporate representation learning in prediction with engineered features | | Implement and test autoencoders | Compare and test representation learning methods |
| **Objective 5.5.c** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Extract and process business transaction data | | | Obtain and prepare business transaction data |
| **Activity 2:** Employ feature engineering and prediction in the business transaction data | | | Employ and test various models for prediction |

## Education

We are happy to report that in contrast with the number of approved bachelor's programs in data science when the project began (zero), we now have three bachelor's of science in data science degree offerings in three regions of the state- UARK, A-State, and UCA. Work is ongoing to solidify plans at SAU, and the first associate's offering at North Arkansas College is expected to accept enrollment as early as Fall 2023. The first cohort of graduates matriculated from UARK in May 2023, three students total. Once the matriculation quota is reached, UARK will initiate ABET accreditation of the degree offering. Below is a summary of specific accomplishments and the stoplight tables for the team.

**ED1, Developing a combination of model programs, degrees, pedagogy, and curriculum including a 9-week middle school coding block; a technical certificate, certificate of proficiency, and associate of science in data science; and a Bachelor of Science in data science with minors or concentrations.** Objective 6.1.a will be replaced in the pending revision of the strategic plan. The project champion left the state and recent feedback from educators indicates that this is no longer a significant challenge. The team has collaborated with the Arkansas Department of Education and partners at EAST Initiative to determine a more appropriate strategy. The tables for this objective are not included due to lack of progress on these milestones. The team has developed multiple proposals for two-year schools and has placed focus on engaging two-year campuses across the state during Years 3 and 4 of the project. Some minor adjustments are planned regarding outreach to other campuses and how we are supporting the participants on other campuses that don't have approved degree offerings yet. We plan to host "Data Science for Arkansas" branded events across the state during Year 4 to raise awareness and interest, including a series of broadcast lectures, data

science career fairs, and asynchronous campus-based events. Most of the Objective 6.1.b milestones are complete or will be completed by the end of Year 3. The only milestones that are delayed are related to the accreditation process, which cannot begin until students matriculate from the program.

| **Objective 6.1.a** | **Specific Milestones** | | |
|---|---|---|---|
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Hold a two-day workshop to include K20 computer science educators to outline the curriculum and pedagogy and establish the project timeline, roles, and deliverables | Workshop completed, plan finalized and disseminated to stakeholders | | |
| **Activity 2:** Develop curriculum | | Curriculum 50% complete | Curriculum 100% complete |
| **Objective 6.1.b** | **Specific Milestones** | | |
| | Year 1 | Year 2 | Year 3 |
| **Activity 1:** Create the 5-year Plan to meet the Objective | Plan disseminated to stakeholders | Review 5-yr plan & update as needed | Review 5-yr plan & update as needed |
| **Activity 2:** Identify the level of involvement and timing by academic institutions within the State | Cohorts identified, all collaborators assigned | Begin Cohort 1 | Begin Cohort 2 & Complete Cohort 1 |
| **Activity 3:** Review UA-Fayetteville and UCA Data Science Programs with the Teams | 1 meeting complete | | Update meeting complete |
| **Activity 4:** Convene workshops annually of engaged academic and government institutions to establish baseline | 3 Workshops completed | 3 Workshops completed | 3 Workshops completed |
| **Activity 5:** Define Data Science Objectives and Outcomes base for defined degrees and certificates | Info disseminated to stakeholders | | Updated Info disseminated to stakeholders |
| **Activity 6:** Define Data Science Courses Objectives, Learning Outcomes, and applicability to the defined degrees and certificates | | Info disseminated to stakeholders | |
| **Activity 7:** Dissemination of developed program details with collaborating institutions, government, and industry partners | Info disseminated to stakeholders | Updated Info disseminated to stakeholders | |

| | | | |
|---|---|---|---|
| **Activity 8:** Ensure defined programs are in line with appropriate accrediting bodies | Identify "Wave 1" of accreditation candidates | | Review "Wave 1" for accreditation readiness |
| **Activity 9:** Prepare and submit program proposals of each type at each level for appropriate approval | Begin "Cohort 1" Proposal Preparation | Submit "Cohort 1" Proposals | Begin "Cohort 2" Proposal Preparation |
| **Activity 10:** Evaluate progress and iteratively improve for future programs as appropriate | | | Evaluation report disseminated to stakeholders |
| **Activity 11:** When accreditation is available, propose first-pass consultative reviews by those bodies for ready institutions | | | "Wave 1" consultative review |
| **Activity 12:** Prepare those institutions which do not have accredited programs in closely aligned areas for the pre-accreditation visit year | | | Non-accredited Cohort 1 pre-accreditation review |
| **Activity 13:** Identify appropriate accreditation by program and provide visibility to academic institution administrations | | | Cohort 1 Appropriate Accreditation bodies identified and reviewed with appropriate academic institutions |
| **Activity 14:** Create and maintain clearing house for course materials | Create shared resources with UAF UCA existing materials and establish cataloging methodology | Add Cohort 1 developed materials | Update contributed materials |
| **Activity 15:** Connect students, courses, problems, data, etc., with the Research Themes | Identify "Opt-In" Research Theme Researchers & Collaboration Types & Timing | Group 1 Collaboration | Group 2 Collaboration |

# Other Project Elements

## Partnerships and Collaborations

The project team has maintained a number of successful partnerships and collaborations during Year 3, including a new partnership with IEEE-USA. In early 2023, PI Fowler was connected to the current Director of IEEE-USA, Russell Harrison, through an AR EPSCoR entrepreneur, Matt Francis who founded Ozark Integrated Circuits. The first major collaboration with IEEE-USA will be the upcoming Innovation, Workforce, and Research Conference (IWRC) which will be held September 13-15 in Little Rock. Several speakers from NSF have already confirmed participation in the event. More information can be found on the event website: http://iwrc.ieeeusa.org/

The Arkansas Center for Data Sciences (ACDS) has continued to grow and currently hosts apprenticeship programs for 21 various professional roles in information technology, data analytics, cybersecurity, and related fields. ACDS is serving more than 100 employers in Arkansas, including businesses like Hytrol Conveyor, JB Hunt, Arvest, Simmons Bank, Arkansas Blue Cross Blue Shield, First Orion, Walmart, Metova, and the City of Little Rock. The project has continued collaborating with ACDS through co-hosted events and webinars, placing interns, and collecting feedback from employers to inform education and training.

More details regarding the partnerships with the Arkansas School for Mathematics, Sciences, and the Arts (ASMSA), and the EAST Imitative, are included below in the report sections pertaining to those specific activities.

## Technology Transfer, Innovation, & Entrepreneurship

In 2022, Arkansas climbed to No. 1 in the number of national I-Corps teams among all NSF EPSCoR jurisdictions. Previously, Arkansas was ranked No. 16, climbing 15 positions to the No. 1 spot in the past two years. Among EPSCoR-eligible states and at the time we issued a press release on this subject, Arkansas had nearly twice as many teams as No.-2 ranked Alabama and was ranked No. 15 among all states, regardless of EPSCoR eligibility. This ranking was up from No. 43 in recent years and surpasses states such as Colorado, Arizona and Virginia. This huge increase in I-Corps participation was due in large part to the collaboration with Weston Waldo at UARK who worked with PI Fowler to form teams with DART participants and DART IAB members. During Year 3, the project issued three industry mentor stipends to industry representatives to serve as the mentor role on national I-Corps teams.

Unfortunately for us, Weston recently moved to Texas and joined UT Austin upon their award for an I-Corps hub. To maintain this momentum, the central office is working with the Arkansas Small Business and Technology Development Center, UAMS Bioventures, the Catalyst at Arkansas State University, and other partners in the entrepreneurial support ecosystem to finalize a DART tech transfer roadmap and prepare to support the faculty coming out of the I-Corps program for the next steps in commercialization.

## Seed Grants

Below are progress reports from each of the first round of seed grant awardees.

**Emily Bellis, A-State; "AgAdapt: An evolutionarily-informed algorithm for genomic prediction of crop performance in novel environments"**

The first year of our work on this grant has focused primarily on development of our machine learning approach to predict crop traits in novel environments. The key innovation in our strategy is the idea that feature selection informed by crop evolutionary history will improve prediction performance in new environments. So far, we have carried out preliminary testing of our strategy using a subset of single nucleotide polymorphisms (SNPs) associated with adaptation of maize landraces to variation in altitude and latitude (data published as part of a previous study [Romero Navarro et al. 2017, Nat. Genetics]). We have also finished all data pre-processing and imputation for environmental, soil, and weather data from the Maize G2F dataset. We have trained preliminary tree-based genomic prediction models, and are in the process of larger-scale evaluation and testing. Fortuitously, an international data science competition led by the Maize Genomes2Fields Initiative (which produced the publicly available dataset we are using to evaluate our method) was just launched in December 2022 (see https://www.maizegxeprediction2022.org), allowing us to benchmark our strategy against models developed by other data scientists. The competition closes Jan. 15, 2023, and we are working on improving our model to increase our standing on the leaderboard.

One challenge has been that very few of the genetic polymorphisms published as part of Romero Navarro et al. 2017 study lifted over to the genome coordinates in the G2F dataset due to the competition dataset using a more recent and improved version of the maize reference genome. In 2023, we will improve our SNP feature set of adaptation-associated polymorphisms by realigning the genomic data from the maize landraces in Romero Navarro et al. 2017 to the updated reference genome and also by exploring a reference-free (k-mer based) approach. We will also evaluate an improved approach for dimension reduction of environment features, using genome associations with the modeled ancestral niche of Zea mays subspecies parviglumis and Zea mays ssp. mexicana, the two wild teosinte species which contributed to maize domestication. This is an approach we developed previously (Bellis et al. 2020, PNAS). The more comprehensive SNP feature set, and reduction of environmental features, will allow us to further evaluate our hypothesis that using SNP feature subsets of polymorphisms associated with plant adaptation to environment will improve performance of genomic prediction models in new environments.

This project has also contributed directly to the training of four undergraduate students majoring in Math or Computer Science, three of whom are from groups underrepresented in STEM. Results from this seed grant served as preliminary data for Aim 2 of a grant submission I led to the USDA Data Science for Food and Agriculture Systems program in November 2022. I proposed to build on the work supported by the DART Seed Grant, which focused on genomic prediction of 'static', low-dimensional phenotypes from multimodal data. The submitted

proposal builds on this through development of methods for genomic prediction of high-dimensional phenotypes (i.e. multispectral, multitemporal images from unmanned aerial vehicles that capture dynamic stress response trajectories).

In 2023, I plan to lead preparation a manuscript describing the work we have carried out in the first year of the grant period, with intended submission in Summer 2023, along with three undergraduate student coauthors who have contributed substantively to the project. I will also hire a Master's student starting in summer to lead develop and deploy the genomic prediction model on our lab website.

**Fig. 1.** Figure from my November 2022 grant proposal submission to the USDA Data Science for Food and Agricultural program, showing proposed next steps in development of the work supported by this seed grant.

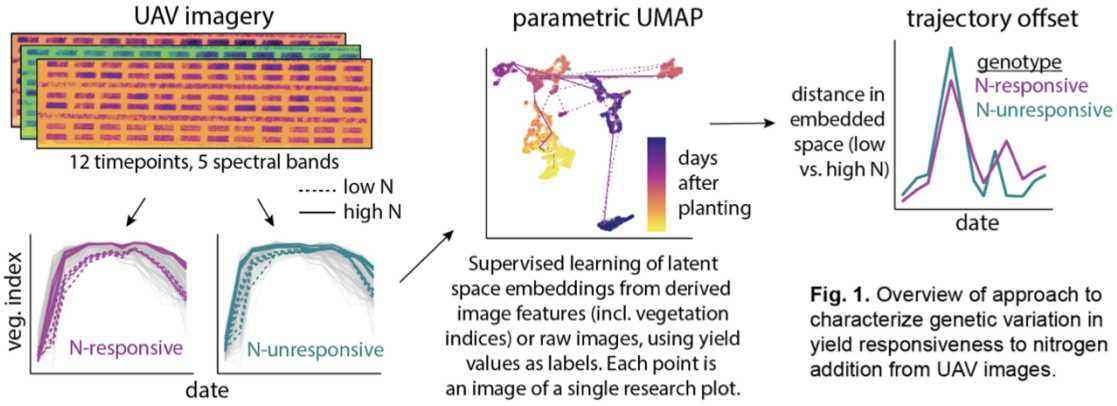



**Fig. 1.** Overview of approach to characterize genetic variation in yield responsiveness to nitrogen addition from UAV images.

**Dongyi Wang, UARK; "Toward fair and reliable consumer acceptability prediction from food appearance"**

Illumination estimation is a fundamental prerequisite for many computer vision applications. Unnatural illumination would influence human perceptions of essential characteristics of goods. For example, humans will feel differently in response to food products in retail stores under different lighting conditions, which can affect consumers' consumption decisions, and potentially causing customers' economic losses. To address this, we have created a novel FoodVisionDataset (FVD) [1] describing the relationship between food (lettuce)_appearance and consumer acceptability under different illumination conditions as shown in Table 1 (illumination temperature and power).

Table 1: Illumination temperature and power settings

| | | | | |
|---|---|---|---|---|
| 3500 K, 17.5 W | 4000 K, 17.5 W | 4250 K, 17.5 W | 4500 K, 17.5 W | 4750 K, 17.5 W |
| 3500 K, 20 W | 4000 K, 20 W | 4250 K, 20 W | 4500 K, 20 W | 4750 K, 20 W |
| 3500 K, 22.5 W | 4000 K, 22.5 W | 4250 K, 22.5 W | 4500 K, 22.5 W | 4750 K, 22.5 W |

The fresh lettuce samples were collected from local grocery stores and then were cut and stored at the 4°C conditions. FVD contains images from 9 lettuce samples. Each sample was imaged at 5 different days utilizing Basler acA1920-40gc camera. The lighting condition is generated from a professional photography light box with adjustable brightness and illumination temperature, and there are total 675 images (9 samples * 15 images/sample/day * 5 days) generated. The images were firstly graded internally by three lab members. t-test results show that illumination temperature and power affect the graders significantly ($p<0.05$). Meanwhile, the inter-grader reliability is low quantified by Cohen's kappa value (<0.3) and three-way ANOVA ($p<0.0001$). Additionally, an external grading happens this fall at University of Arkansas Sensory Center with 88 participants involved. Each participant was assigned to grade 75 random images/day * 9-day sessions, by giving purchase intent, overall liking, and freshness ratings score.

As the first step to build an illumination robust machine learning model, an illumination parameter prediction network has been tested, serval commonly used deep learning network architecture has been tested with best results in mean absolute error 386.1 for temperature predictions and 3.17 for power predictions. The next steps will evaluate the illumination effects for 88 panelists, and an illumination robust deep learning model will be developed with the guidance of predicted illumination parameters to benefit broader Artificial Intelligence and image processing audiences, which can also be transferred to general industrial manufacturing and inspection applications.

Our team also filed a disclosure of the dataset at the Division of Agriculture commercialization Office at the University of Arkansas. Our project was also selected for the National NSF I Corps award with a funding of $50,000 and we participated in the fall 2022 cohort to investigate the commercialization aspects of our technology, which is built upon the success completion of Regional I-Corps program with the Southwest Regional I-Corps Node award #1740705 – Southwest Innovation Corps Node course held in August 2022. Adewale Obadimu, a member of the DART industry advisory board, served as the team's industry mentor. The team's poster won second place at the DART Annual Conference and Poster Competition in May 2022, and Swarna Sethu won a travel award to attend and present her poster at the National NSF EPSCoR Conference in Maine in November 2022. Two publications have been submitted for review.

**Rob Coridan, UARK; "Machine Learning-based emulation and prediction in ensembles in disordered photocatalytic composites"**

The focus of our proposed research is to develop neural network emulators for predicting light concentration in a model disordered photonic composite. My experimental research group builds composites out of chemically-synthesized nanospheres that we use as a structured photoelectrode for solar energy conversion. One of the primary functions of the structuring is to concentrate light in "hot spots", then add functional materials such as light absorbing semiconductors, chemically synthesized photocatalysts, or plasmonic nanocrystals.

We aim to design the nanostructure of these composites by choice of spheresize, diameter, and the relative refractive indices of the components. Due to the complex randomness or a composite, modeling the combinatorial number of structures relevant to a single choice of these parameters (an ensemble) is an intractable computational problem. Developing methods to approximate them becomes exceedingly important.

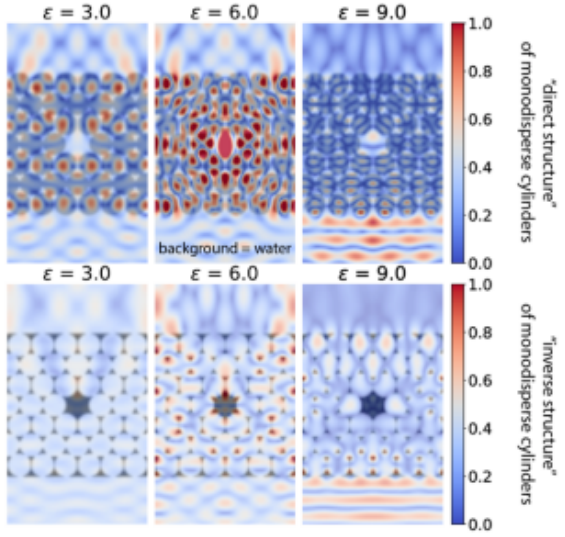



**Figure 1** – Variations in the steady state electric field, $|E(x,y)|^2$, in the central cavity of an omission glass electrode without additional omissions. The intensity at the functional omission in the center for a given wavelength of light depends on the index of refraction of the scattering material, the diameter of the scatterers, whether or not the structure is the direct one (top) or the templated inverse (bottom), among other issues. These electrodes are representative of structures that we can synthesize in the lab.

Our solution for this is to use neural network emulators of the ensemble. First, we generate a relatively small training set of physical (finite-difference time domain, FDTD) simulations representing the steady-state light-field intensities in the targeted composite ensemble. This set is then used to train a neural network to predict the steady-state light field for all possible configurations. This allows us to explore the structure of light localization throughout the ensemble with orders of magnitude improvement in efficiency. In the abstract sense, this approach is simple. However, the size and scale of the data limits this approach and requires new efficient ways to represent the data. In our DART SEED project, we have proposed to research a model system called an omission glass, a 2D photonic system with a regular lattice of light scatterers and a binary representation of site occupancy (1 if occupied, 0 if unoccupied). Figure 1 is an example of an omission glass with a single omission at the center in a variety of dielectric environments (inverse and direct structures with varying values for the real dielectric constant, ε). In recent work, we have focused on developing efficient representations of the output of the light field simulations, which are grids of real-valued data like images. In previous work, we used the raw data as the output feature of the neural network emulator, though limited to a particular volume of interest (Coridan, Chem Comm 56, 10473 (2020)). This uses the output data without approximation but ignores the spatial relationship between pixels –since the data represents real, physical fields, the voxels are spatially correlated in this image. To solve this problem, we have developed an approach based on the principal component analysis (PCA) methods used for facial recognition. In facial recognition, an algorithm is trained to find a basis set of 'eigenfaces' from a training set of facial images. Each real image of a face can therefore be represented by the linear combination of these eigenfaces, and similarities are captured in the vector representing the coefficients in this linear combination.

We used the same approach to capture the spatial relationship between pixels in field data for the light localization in a cavity at the center of an omission glass system. The digital basis set 'eigenfaces' are for a 40-element basis is shown in Figure 2. In this case, the simulation detail can be represented as a 40-element vector rather than a 16x16 array while also capturing the spatial correlations in the output. We tried a variety of basis-set sizes, from 5 to 200 –more than that would not compress the representation of the data beyond the 16x16 images, which we found to still be rather successful at emulating the absorption in a photocatalyst. Based on empirical considerations, the 40-elementbasis struck the correct balance between having sufficient structure to represent the variety of optical profiles in the cavity and overfitting the training set. Additionally, this is a different problem than we have used in the past, where the data is not terminated at the edges of the volume as they would in an absorption profile.

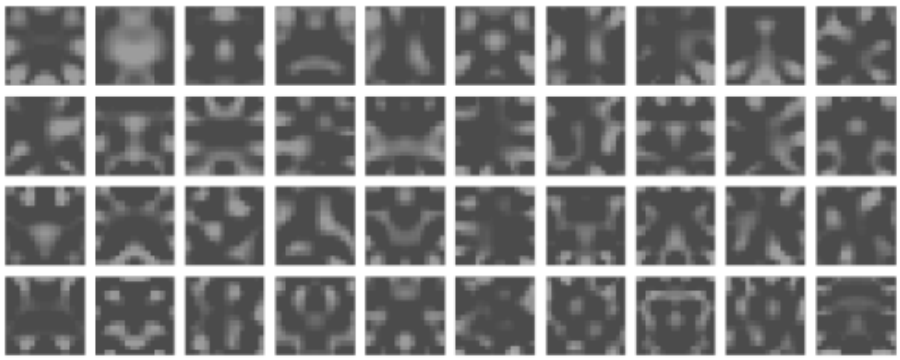
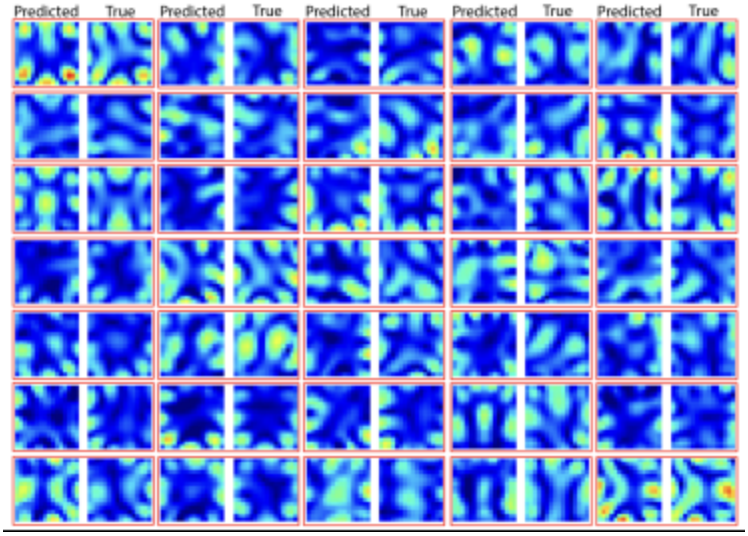


**Figure 2** – An example of a 40-element PCA "eigenface" digital basis set for representing the steady-state energy density, $|E(x,y)|^2$, in a pore in the center of an omission glass photoelectrode.

We used the 40-element basis set to decompose a training data set into the appropriate vector representation, then trained the emulator to map between the binary omission glass structure representation and the PCA-reduced basis representation for the fields. After applying the reconstruction routine, we can use the emulator to predict the steady state electric field energy density, $|E(x,y)|^2$, in a test set of simulations that had not been used in the training (Figure 3). The qualitative predictions of topology/structure of $|E(x,y)|^2$ match each "true" simulation rather nicely. Quantitative matching isn't strictly necessary, but this worked well in most cases as well.

**Figure 3** – Neural network emulation of $|E(x,y)|2$ in an optical cavity at the center of an omission glass. The prediction space used the 40-element PCA digital basis shown in Figure 3 to predict the




representative vector from the omission glass structure. The representative vector was then transformed into an array representing a prediction of the electric field.

We are currently working on similar PCA analyses based on larger scale electrodes. The main issue with the omission glass as described above is that it 'chooses' the cavity at the center of the electrode, which is not an easy thing to do experimentally. We are currently generating a much larger data set (540 scatterers removed at random) with no central one to focus on. An example of these simulations is shown in Figure 4. We want to include the full relevant volume of the simulation and translational symmetry (periodic boundary conditions) to maximize the training capacity of a small number of simulations. This also requires a new mathematical representation to reduce the dimensionality of the output space. We are currently learning how to apply FFT methods and autoencoders to compress the data for representing large-scale fields in a neural network emulator.

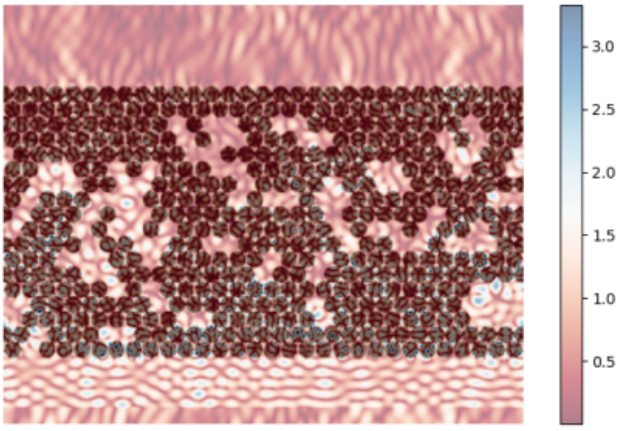



**Figure 4** – A large-scale omission glass (540 sites in the lattice) with pairs of neighboring scatterers removed to 90% density.

**Weijia Jia, Christopher Kellner, Jacob Grosskopf, Xinli Xiao, Matthew Wilson, Wan Wei, Arkansas Tech University; "Development of Interdisciplinary Research Collaborative to Provide Datasets in Support of Education Research in Data Science"**

Stage I – Preparation and publicity: The research team worked together to contact the University Relations Office to advertise the research project in Tech News and on the ATU OneTech platform, prepare information for the Tech news report, design the detailed guidelines for dataset collection, design flyers and share with the faculty members at ATU, solicit input from the faculty colleagues, and recruit undergraduate students for presenting in the Data Science Workshops (in Stage II).

Stage II – Dataset Collection and Data Science Workshop: In Fall 2021 and Spring 2022, eight sessions of ATU Data Science workshop were offered through Webex to undergraduate students, graduate students, and faculty and staff members at ATU. In Fall 2021, four sessions on data visualization were offered. In Spring 2022, the four sessions were on data wrangling.

Ten undergraduate students were recruited and trained on both R coding and presentation skills before the workshop sessions. The research team organized the study groups to offer these trainings to the student presenters and helped the students prepare the R codes, slides, and practice problems in Rmarkdown files. The workshop sessions were highly rated by the workshop participants.

The dataset collection procedure was first advertised in October 2021. The DART Seed research team worked closely on the collection by soliciting input from their respective faculty colleagues, answering questions from researchers, and organizing meetings with the researchers to discuss the possible datasets. By the end of Stage II, there were 25+ datasets collected. After screening and discussion, we selected 11 datasets.

Stage III – Data Cleaning and Data Wrangling: During Summer 2022, five undergraduate students at ATU, Faith Walker, Jiajie Yi, Mary-Ashley Qualls, Stanley Pham, and Tristan Caja, were selected to work on the data wrangling of the datasets collected in Stage II. Every student was assigned two datasets to work on. Weekly meetings on Tuesdays were organized and hosted by the research team. Students were asked to give 10~15 minutes to report the work done in the week. The research team answered questions from the students, provided guidance and advised the students on the next research step. Additional help sessions on Thursdays were offered to provide individual training as needed. Students were guided to contact the researcher who contributed the datasets (in Stage II) for necessary information about the data. Students also prepared the final reports to summarize their findings and the data wrangling results under the guidance of the research team.

Stage IV and Stage V research are on-going in Year II.

Stage IV – Course Project and Solution/Programming Code Preparation: The Data Science Course Project Design Team was recruited at the end of Spring 2021. The team will meet and work on the project design in Fall 2022 and Spring 2023.

Stage V – Project Wrap-Up and Dissemination of Result: The PI of the project, Dr. Weijia Jia, will submit a proposal to the 2023 Symposium on Data Science and Statistics to present the project in May 2023.

This project involved 10 undergraduate students on campus, including several female and underrepresented minority students, many of whom attended the 2022 ASRI. The students all completed presentations on their work and are planning to participate in the 2023 ASRI.

**Suzan Anwar, Philander Smith College; "Generating Big Radiogenomic Data of Cancer Using Deepfake Learning Approach"**

The research started on 10/1/2021 with 3 students who were enrolled in the data science using Python course. The students needed to learn about deep learning methods and specially Deepfake approach. The teaching process of deep learning approaches continued until July 2022. In June, we started exploring datasets for the project and selected the Breast Histopathology Images datasets, we did data preprocessing and we are building the model to generate more of this data for AI researchers to use. We will develop a Generative Adversarial Network GAN to generate deepfakes. A Generative adversarial network (GAN) is a special type of deep learning, designed by Goodfellow et al. (2014), which is what we call convolution neural networks (CNN). How a GAN works is that when given a training set, it can generate new data with the same information as the training set, and this is often what we refer to as deep fakes. CNN takes an input image, assigns learnable weights and biases to various aspects of the object, and is able to differentiate one from the other. This is similar to what GAN does, it creates two neural networks called discriminator and generator, and they work together to differentiate the sample input from the generated input.

Next, we will develop a survey to determine if participants are able to identify authentic versus deep fake images. The survey employs a questionnaire asking participants their perception of AI technology based on their overall familiarity with AI, deep fake generation, reliability and trustworthiness of AI, as well as testing to see if subjects can distinguish real versus deep fake images. This project supported research experience for five undergraduate students, all URM, two of whom graduated in May 2022. The other three students are still working on the project, and participated in the 2022 ASRI.

**Kevin Phelan, Tiffany Huitt, UAMS, & Annice Steadman, Little Rock School District; "Piloting Big Data Science in Arkansas Middle School Classrooms"**

Year 1 of the project was a great success! The Arkansas Big Data Science (ARBDS) program staff developed and piloted a data science focused middle school curriculum in several schools across the state. The two specific aims of the Seed grant are: 1) Increase teacher knowledge & confidence in using big data science in the classroom, and 2) Increase student knowledge, skills and efficacy in using big data science. We have addressed each aim and met our stated goals for year one of the grant.

Developed Curriculum: The data science piloted program uses online short videos, slides, and datasets from NOAA and other sites to introduce students to the data science field while also promoting basic data literacy targeted to the middle school level. The curriculum that we have developed consists of three separate modules. The first two modules are based on NOAA data-in-the-classroom website while the third module is a data science focused module that uses a roller coaster dataset and introduces students to the Orange 3 data mining software. We have developed teacher resource guides, student worksheets, slides and active weblinks for all modules and posted this to our google drive for the pilot program. All of the activities in our curriculum have been aligned to Arkansas and NGSS standards. Our program introduces teachers and students to the Common Online Data Analysis Program (CODAP) which is a free

HTML program designed specifically for middle through high school students. Our curriculum has been designed to work within the limitations of the chrome books that students typically use in their classrooms. If better laptop computers are available then we conduct one or two sessions on data mining using the free online program called Orange 3 that has a graphical interface (python block coding).


**Figure 1:** Top row – Composite showing the NOAA website organization; sample teacher guide; two sample student worksheets. Middle row – Students obtaining water model measurements in the classroom; student generated graph with box plot; Student plot of deviation from mean values; Student generated color bar for visualizing deviations from mean. Bottom row – NOAA tidal sites for data downloading; CODAP plotting of tidal data for hurricane IDA; sample Orange flow diagram for machine learning sorting of photos; and dendrogram of sorted photos using inception V3.

Recruitment of teachers: The Arkansas Big Data Science team met with teachers at various schools in person and on ZOOM in order to introduce our program. We also conducted a virtual statewide Arkansas middle school/junior high teacher workshop during the summer on June 29th, 2022. We had a total of 3 teachers attend this event. The workshop was successful in recruiting teachers at Vilonia Middle School into the second group of classes piloting the program (2022-2023 academic year). We are currently working with the Arkansas Department of Education to setup meetings with teachers, administrators, and Education Coop Directors across the state to spread the word about our program so we can recruit more teachers and schools.

Summary of Activities-Year 01: Table 1 summarizes our classroom piloting outreach activities for year one of the grant. In the 2021-2022 academic year (August – May) of year 01, we implemented the curriculum in 2 schools in the Little Rock School District (Forest Heights STEM Academy and Pinnacle View Middle School) and 1 school in the Cabot School District (Cabot Junior High South). The LRSD schools included 7th and 8th grade science classes while the Cabot JHS classes were all 8th grade science. All of these were preAP/advanced classes. This initial pilot included 5 teachers, 16 classes, and reached 365 students. To date, the team has engaged more than 770 students across 32 classrooms in 5 schools.

In the first part of the 2022-2023 academic year (August – September) in year 01 of the grant, we began implementing the curriculum in the Vilonia School District (Vilonia Middle School), the LRSD (Mabelvale Middle School), and again in the Cabot School District (Cabot Junior High South). The Vilonia classes are 8th graders (2 of which are advanced classes), the Cabot JHS classes include 5 classes of advanced 8th graders, and the Mabelvale class is a combined class of advanced 7th and 8th graders. This second pilot includes 4 teachers, 16 classes and over 388 students. These activities are being continued in year 02 of the grant which started on October 1, 2022. We also have plans to expand the number of schools reached in year 02 of the seed grant.

Program Evaluation: The students have provided generally positive feedback. We are still in the process of evaluating the first year surveys and will have more specific student evaluation data available later. The teachers involved in the program have rated the program highly and indicated that it is engaging their students. The teacher evaluation data from the five teachers involved in the 2021-2022 academic year are summarized in the table below. We have posted short teacher testimonials from three of these teachers on our website at: https://medicine.uams.edu/neurobiology/outreach/arkansas-big-data-science/

TABLE. Teacher evaluation of the data science pilot program – year 01 of the grant.

| **Teachers reported that they were:** | **Likert Rating Mean ± SEM (n=5)** |
|---|---|
| more likely to discuss data science careers with their students | 5.0 ± 0.0 |
| more confident in using large data sets in their class | 5.0 ± 0.0 |
| more likely to use CODAP in their classroom to demonstrate how to analyze and/or graph data | 5.0 ± 0.0 |
| I am more likely to have students use CODAP to analyze and/or graph data | 5.0 ± 0.0 |
| **Teachers reported that their students:** | |
| learned about data science as a career | 4.8 ± 0.2 |
| benefitted from participation in the program | 4.8 ± 0.2 |
| learned new skills from participation | 4.6 ± 0.4 |

| | |
|---|---|
| were engaged during the data science modules | 4.0 ± 0.5 |
| **Teachers reported that:** | |
| the data science program was a valuable use of classroom time for their students | 5.0 ± 0.0 |
| they were interested in using the data science modules in their classroom next year | 4.4 ± 0.4 |

(5-strongly agree; 4-agree; 3-neither; 2-disagree; 1-strongly disagree)

**Han Hu, UARK; "Interpretable Multimodal Fusion Networks for Fault Detection and Diagnostics of Two-Phase Cooling Under Transient Heat Loads"**

The seed grant was used to support research on machine learning analysis of boiling heat transfer data and led to advances in the following three research directions. The team applied and participated in the national NSF I-Corps program during Year 3.

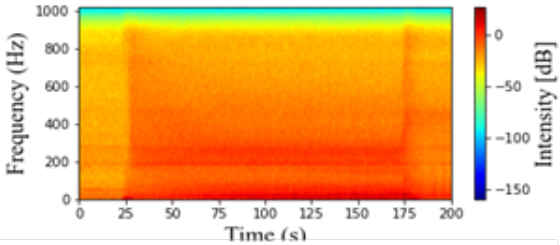
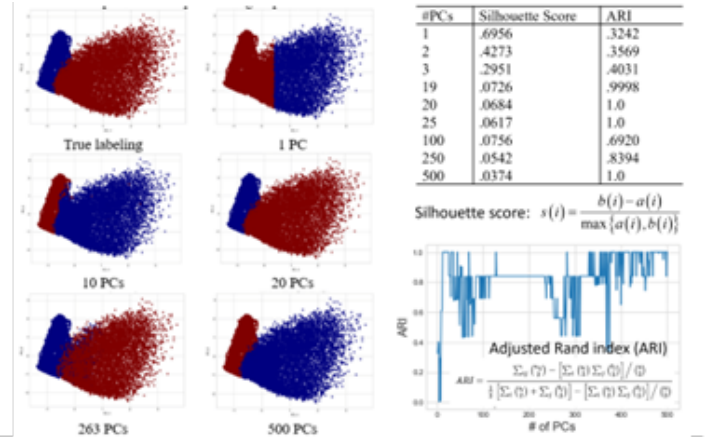



**Figure 1.** Spectrogram during a transient boiling test (ramp up heat load) on a polished copper surface.

Non-Intrusive Cooling System Fault Detection and Diagnostics Using Acoustic Emission: Power intensification and miniaturization of electronics and energy systems are causing a critical challenge for thermal management. Single-phase heat transfer mechanisms including natural and forced convection of air and liquids cannot meet the ever-increasing demands. Two-phase heat transfer modes, such as evaporation, pool boiling, flow boiling, have much higher cooling capacities but are limited by a variety of practical instabilities, e.g., the critical heat flux (CHF), aka departure from nucleate boiling (DNB) in the nuclear industry, flow maldistribution, flow reversal, among others. These instabilities are often triggered suddenly during normal operation, and if not identified and mitigated in time, will lead to overheating issues and detrimental device failures. For example, when CHF is triggered during pool boiling, the device temperature can ramp up in the order of 150 °C/min. It is thus critical to implement real-time detection and mitigation algorithms for two-phase cooling. In the present work, we have developed an accurate and reliable technology for fault detection of high-performance two-phase cooling systems by coupling acoustic emission (AE) with multimodal fusion using deep learning. We have leveraged the contact AE sensor attached to the heater and hydrophones immersed in the working fluid to enable non-invasive fault detection. Fig. 1 shows the spectrogram during a transient boiling test from single-phase convection to nucleate boiling and the CHF condition. Two major frequency shifts are observed as the onset of nucleate boiling and CHF, respectively, demonstrating that acoustic emission frequencies can be used as signatures of boiling states.

Unsupervised Learning Models for Detection of Critical Heat Flux During Pool Boiling: Unsupervised learning is used to extract and identify the key features of boiling images Using




| #PCs | Silhouette Score | ARI |
|---|---|---|
| 1 | .6956 | .3242 |
| 2 | .4273 | .3569 |
| 3 | .2951 | .4031 |
| 19 | .0726 | .9998 |
| 20 | .0684 | 1.0 |
| 25 | .0617 | 1.0 |
| 100 | .0756 | .6920 |
| 250 | .0542 | .8394 |
| 500 | .0374 | 1.0 |

**Figure 2.** k-means clustering of boiling images in different regimes with varied number of PCs.

different numbers of principal components (PCs), K-means clustering was performed. Fig. 2 illustrates the clustering results using different numbers of PCs. Each point on a plot represents one image (The point is made of the first 2 PCs). Each color represents which cluster the image belongs to. The true labeling plot shows the two boiling image groups, post-CHF and pre-CHF. The other plots show how the unsupervised model grouped the images with various numbers of PCS. Ideally, the clusters will look like the first plot. This is the case for 20 PCs and 500 PCs. With just 20 PCs, the clusters are the same as the pre-CHF and post-CHF groups. In order to analyze the clusters formed through K-means a few different metrics are used. Some such as silhouette score and adjusted rand index (ARI) are used in clustering. The silhouette score is a measurement of how dense the clusters are and requires no knowledge of the true labels. A score of 1 implies dense and distanced clusters while a lower score implies more sparse clusters. The true labels are also used to evaluate the performance of clustering. When 500 PCs are used, k-means clustering can separate the boiling images from these two regimes without any error. The adjusted rand index (ARI) is used to determine how similar two clusters are to each other. The ARI for various numbers of PCs are shown in Fig. 2

Nonintrusive Heat Flux Quantification Using Acoustic Emissions During Pool Boiling:

Monitoring two-phase cooling systems is crucial to avoid thermal runaways and device failures. Nonintrusive monitoring methods using remote sensing, e.g., optical and acoustic sensors are desired to avoid interfering with bubble dynamics and ease replacement. Compared to image-based technologies, sound-based sensors are cheaper and do not require the same environment as cameras. Acoustic signals during pool boiling have been used to identify boiling states, but acoustic-based quantitative predictions have been challenging. We have developed a machine learning framework to determine the heat flux during pool boiling using acoustic signals captured through a hydrophone (Fig. 3). This framework investigates and compares the performance and computational cost of six machine learning models by coupling two feature extraction algorithms (fast Fourier transform and convolution) and three different regressors (multilayer perceptron, random forest, and Gaussian process regression). The fast Fourier transform-Gaussian process regression model is found to be the most promising with high accuracy and the lowest computational cost. A parametric study is performed to investigate the effect of the temporal length and sampling rates on the model predictions. It is found that the model's performance is improved with increasing temporal lengths of the acoustic sequences for all sampling rates. Acoustic features below 512 Hz are found to be most significant for heat

flux predictions. For sampling rates beyond 512 Hz, the model performance is dictated by the temporal length of the acoustic sequences.

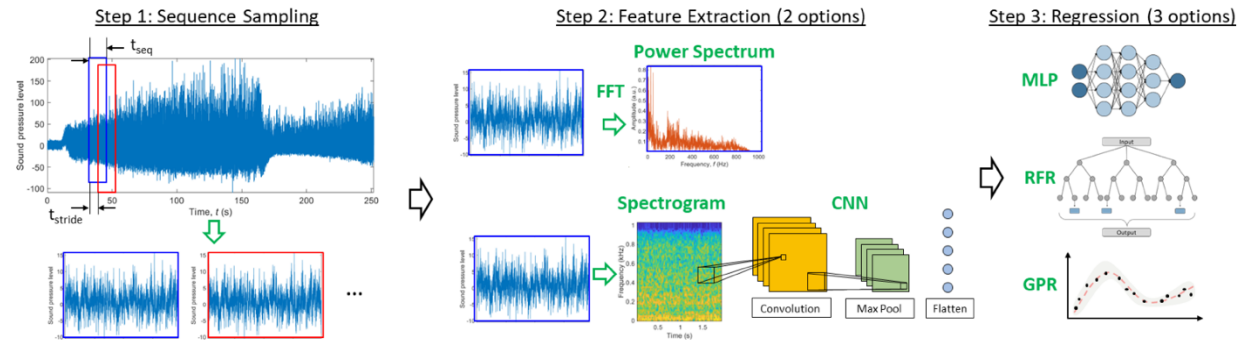



Figure 3. Illustration of the machine learning model framework for acoustic-based boiling heat flux quantification.

**Yeil Kwon, Nesrin Sahin, UCA; "Crying Out Data Science in the Center of Arkansas-Invitation for High School Students to the World of Data Science"**

The project provided high school students in Arkansas with a one-week long workshop on data science. The workshop took place in June 2022 on the UCA campus and was attended by 15 high school students from central Arkansas. The schedule is shown below.

| | | 6/06 | 6/07 | 6/08 | 6/09 | 6/10 |
|---|---|---|---|---|---|---|
| Begin | End | MONDAY | TUESDAY | WEDNESDAY | THURSDAY | FRIDAY |
| 9:30 | 10:30 | Registration Pre-Survey | Intro. to R | Data Summary & Graphs in R [1] | Data Summary & Graphs in R [2] | Data Summary & Graphs in R [3] |
| 10:30 | 12:00 | Descriptive Statistics | R - Data Structures | Descrive Our World with Data | Explore the U.S. using Data | Sketch Arkansas in Numbers and Graphs |
| 12:00 | 1:00 | Lunch | Lunch | Lunch | Lunch | Lunch |
| 1:00 | 2:30 | Into. to CODAP | R - Data Manipulation | Probabilistic Simulation in R | Predictive Modeling | Team Project with R/CODAP |
| 2:30 | 3:30 | Data Visualization with CODAP | Team Project with R/CODAP | Team Project with R/CODAP | Team Project with R/CODAP | Presentation Post-Survey |

Lectures in the data science workshop included data summarization and visualization with fundamental knowledge of descriptive statistics; the basic syntax of R programming and usage of functions; Probabilistic Simulation and Broken Stick Problems; and Introduction to Simple Linear Regression Model. The hands-on training component included Data summarization & visualization with CODAP; Data summarization & visualization with R using real-world datasets; GDP, infant mortality, fertility, and life expectancy data from the World Bank; Solving Probability problems based on the simulations using R; and Linear Regression

Model with R. Each of the students delivered a presentation about a selected topic they studied during the workshop.

The investigators presented about the workshop at the Arkansas Council of Teachers of Mathematics in November 2022, and plan to present at the Research Council on Mathematical Learning conference in March 2023. One graduate student and one undergraduate student, both female, were hired to assist with the workshop and help with instruction. The graduate student completed her master's degree in August of 2022 and is now employed with Arvest as a statistical analyst. The undergraduate will complete her bachelor's in May 2023 and plans to apply for a graduate degree in Arkansas.

**New Seed Awards**

Five seed grants were awarded in 2022, after a very competitive round with 13 applicants. The proposals were reviewed for scientific merit and DART alignment by the project's external advisory board members. One change to the solicitation this time was that instead of requiring a letter of support from any DART faculty participant, the proposals were required to include a letter of support from an SSC member. This resulted in better alignment and potential for integration of the applicants. The awards are summarized in the table below, and updates from these projects will be provided in the Year 4 report. Some funds were reserved for a third and final round of seed funding, to be held in 2023. We plan to utilize a letter of intent mechanism and conduct outreach at PUIs and encourage applications from underrepresented minority faculty in several disciplines.

| Project Title | Institution | PI Name | Amount |
|---|---|---|---|
| LP: MoDaCoM-TL: Model and Data Compatibility Metric for Transfer Learning | A-State | Jason Causey | $ 92,686.00 |
| DC/LP: Smart curation and deep learning-based enhancement of social risk data | UAMS & A-State | Melody Greer, Sudeepa Bhattacharyya | $ 118,190.00 |
| CI: AI-Supported Cyberinfrastructure for Scalable Flood Resilience Assessment | UARK | Xiao Huang | $ 95,459.00 |
| LP: Machine Learning Approaches for Remote Pathological Speech Assessment for Parkinson's Disease | UAMS | Yasir Rahmatallah | $ 100,000.00 |
| DC / LP: Developing Machine Learning Models to Improve the Effectiveness of Automated Data Curation Processes | UALR | Ahmed Abu Halimeh | $ 99,414.00 |

## Workforce Development

### Graduate Student Training

During Year 3, 117 graduate students participated in DART as research assistants and in other roles. As the project progresses, we are experiencing more matriculation and sending

students off into the workforce or other professional roles. Seven graduate research assistants completed their doctoral degrees during Year 3, and seven additional students completed master's degrees. Yanbin Ye defended his doctoral dissertation under Dr. Talburt at UALR and then volunteered to serve on the project's IAB from his role as Data Science Director at Walmart. Marcel Nwaukwa, Shishila Awung Shimray, Beiimbet Sarsekeyev, Regan Harper Hodges, Kazi Tanvir Islam, Daysi Guerra, Adeola Adesoba all secured full-time positions in industry upon graduating at companies like J.B. Hunt, Walmart, Dell, Cerner, NVIDIA, and others. Billy Spann received a promotion at Windstream upon graduating, from staff to director. Duah Alkam, Sangam Kangel, Visanu Wanchai, John Michael Schonefeld, and Maryam Alimohammadi continued their careers in academia either as research scientists or instructors at Arkansas institutions.

Some graduate participants have received awards, scholarships, and other honors. Graduate students Winthrop Harvey and Jose Azucena from UARK attended and presented at the NSF EPSCoR workshop for AI & No-Boundary Thinking in April 2022. Azucena also won best student paper and was awarded the Hans Reiche RAMS scholarship through the Society of Reliability Engineers (SRE) at the 69th Annual Reliability and Maintainability Symposium.

### Undergraduate Student Training

During Year 3, 50 undergraduates participated in DART through research assistantships, summer research experiences, and other roles. Of the 40 undergraduates who completed their bachelor's degrees since the last report, eight were bridged into advanced degree programs including some in Arkansas and some in other states. Makenzie Spurling, Knia Williams, Owen Habeger, Khue Dang, Maeyonna Done, Isaac Woollen, Layla Holloway, Annelise Koster, Natalie Garcia, and Bret Timme all received full time positions (mostly as developers) in industry (mostly within Arkansas), including at companies like First Orion, ArcBest, Mastercard, FedEx, Goldman Sachs, and Murphy USA.

Among notable accomplishments from DART undergraduates, a team from Philander Smith College led by Dr. Anwar, a seed grant recipient, won first prize at the UIDP 2023 Pitchfest in Nashville, TN after being one of only two finalists among all 20 teams representing HBCUs at the event. The team proposed a Smart Prosthetic Knee for Above Knee Amputees and received a cash prize of $10,000.

### Arkansas Summer Research Institute

The 2022 Arkansas Summer Research Institute (ASRI) was held June 3-17 virtually and was attended by 102 students. Over half (52%) of the participants reported as Male, 45% Female and 3% Non-binary or non-conforming. The ethnic distribution was over one-third (40%) Asian; more than one-fourth (27%) White; more than one-fifth (22%) Black; 7% Hispanic and 4% 'other'. The ASRI team iteratively improves the ASRI experience based on evaluation data and

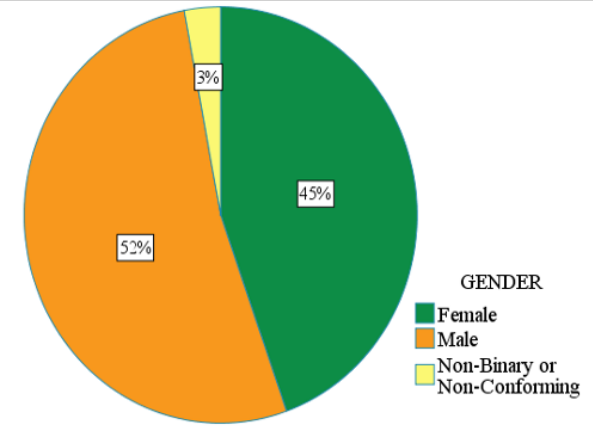
other information from student and faculty participants. In 2022, students were given the option to choose a dataset from the curated ASRI repository and choose either a Python track or R track. External evaluation was conducted by Minnick & Associates.

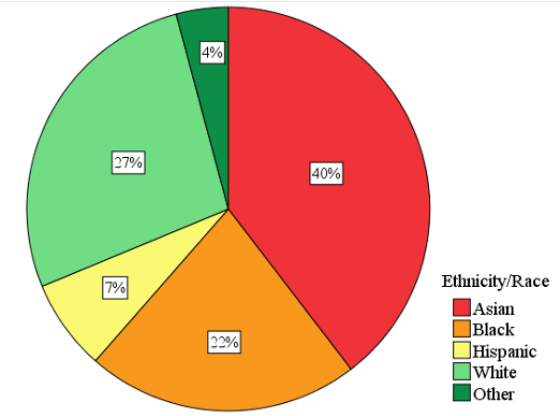
**State and Institutional Affiliation of ASRI Participants**

| State/Country and Institution | Number | Percent |
|---|---|---|
| AR-Arkansas State University | 5 | 4.9 |
| AR-Arkansas Tech University | 7 | 6.9 |
| AR-Bentonville High School | 1 | 1.0 |
| AR-Harding University | 1 | 1.0 |
| AR-Henderson State University | 2 | 2.0 |
| AR-Hendrix College | 1 | 1.0 |
| AR-Ignite Professional Studies | 1 | 1.0 |
| AR-Northwest Arkansas Community College | 2 | 2.0 |
| AR-Ouachita Baptist University | 1 | 1.0 |
| AR-Philander Smith College | 2 | 2.0 |
| AR-Southern Arkansas University | 1 | 1.0 |
| AR-UAMS | 1 | 1.0 |
| AR-University of Arkansas at Fayetteville | 32 | 31.4 |
| AR-University of Arkansas at Little Rock | 11 | 10.8 |
| AR-University of Arkansas at Pine Bluff | 1 | 1.0 |
| AR-University of Central Arkansas | 5 | 4.9 |
| LA-University of Louisiana at Lafayette | 1 | 1.0 |
| MS-Mississippi State University | 19 | 18.6 |
| NC-Wake Forest University | 1 | 1.0 |
| OK-University of Oklahoma | 2 | 2.0 |
| PA-Pennsylvania State University | 1 | 1.0 |
| TN-University of Memphis | 1 | 1.0 |
| TX-Southern University and A&M College | 2 | 2.0 |
| N/A | 1 | 1.0 |
| Total | 102 | 100.0 |

**Self-Reported Gender and Ethnicity of Participants**

**Gender (n=102)** **Ethnicity (n=102)**




An introductory session on the first day describes expectations, which continue to be reiterated throughout. Students are led through interactive lessons in data exploration, literature search, reading research papers, hypothesis formulation, experimental design, choosing the correct statistical tests for their dataset, data analysis in R or Python, visualization (graphing) of data, and how to develop a slideshow and present their findings professionally. Time is built in ahead of each day's first session for questions and instructions. To maximize and individualize the participant experience, up to 4 breakout sessions are used simultaneously to diversify learning by interest and level. Multiple technical sessions teach skills in Python and RStudio. Panel discussions have included STEM Careers, Data Science in Today's World, Entrepreneurship in Research, and Diversity, Equity, and Inclusion. A number of faculty present "research stories" that exhibit their diverse backgrounds, career paths, and the reward derived from challenging research endeavors. Each day includes a lunch break with virtual office hours. Students sign up for general office hours or more targeted working groups during these times. Breaks and wrap-up times are used to avoid student frustrations when sessions go over scheduled times.

Each day ends with daily evaluations and "reflection and prep" homework based on that day's sessions to support the experiential learning process. Examples of homework assignments include reviewing a data repository, signing up for a dataset, signing up for specific breakout rooms, reviewing or designing a résumé, constructing experimental design and sharing it with others for feedback, and developing a presentation for the final day. Below is a three-day excerpt from the 2022 ASRI schedule showing multiple concurrent sessions at beginner and intermediate levels with networking and help sessions, research stories, and panel discussions. Synchronous (9am to 5pm) and asynchronous (Reflection and Prep) portions are also shown.

| DAY 6: Monday June 13 | | | DAY 7: Tuesday June 14 | | | DAY 8: Wednesday June 15 | | |
|---|---|---|---|---|---|---|---|---|
| 8:45 | Log In & Announcements | | 8:45 | Log In & Announcements | | 8:45 | Log In & Announcements | |
| 9:00 | Starting your project in R (Beginner #2) | Starting your project in Python (Beginner #2) | 9:00 | Overcoming the Odds- Perseverance in Research | | 9:00 | Justice, Equity, Diversity, & Inclusion in STEM | |
| | Exploratory Data Analysis in R (Int #2) | Exploratory Data Analysis in Python (Int #2) | 10:30 | Feature Selection (Beginner) | Feature Selection (Intermediate) | 10:30 | Individual Consultations | Stats & Viz Working Rooms |
| 11:30 | Support Rooms | | 11:30 | Support Rooms | | 11:30 | Support Rooms | |
| 12:00 PM | Lunch + Virtual Office Hours | | 12:00 PM | Lunch + Virtual Office Hours | | 12:00 PM | Lunch + Virtual Office Hours | |
| 1:00 | Giving a Compelling Research Presentation | | 1:00 | Intro to Arithmetic & Matrix Operations | Working Group- Intermediate Feature Analysis & Engineering | 1:00 | Individual Consultations | Python Stats & Viz Support |
| 2:00 | Designing your Data Science Experiment (Beginner) | Designing your Data Science Experiment (Intermediate) | | | | | | R Stats & Viz Support |
| 3:15 | Afternoon Break | | 2:45 | Afternoon Break | | 3:15 | Afternoon Break | |
| 3:30 | Data Science Case Study - Gathering Data: *Personal* Inputs | | 3:00 | Panel: STEM Careers | | 3:30 | Translational Bioinformatics: A Research Story | |
| 4:30 | Wrap Up | | 4:30 | Wrap Up | | 4:30 | Wrap Up | |
| **Reflection (by 7PM)** | (1) Complete Daily Evaluation (2) Tweet or post on LinkedIn and tag #ArkansasSRI with your response: **What is your experimental hypothesis for your project?** | | **Reflection (by 7PM)** | (1) Complete Daily Evaluation (2) Tweet or post on LinkedIn and tag #ArkansasSRI with your response: **What advice would you give to someone who's interested in data science but hesitant to learn?** | | **Reflection (by 7PM)** | (1) Complete Daily Evaluation (2) Tweet or post on LinkedIn and tag #ArkansasSRI with your response: **What is the best piece of advice you got today?** | |
| **Prep (by 7PM)** | Post a screenshot of the slide that shows your experimental design in the UpSquad social page for asynch prep points. | | **Prep** | Have draft of slides and list of questions ready to share with coaches during consultation tomorrow. | | **Prep (by 7PM)** | Post a screenshot of the slide that shows your chosen statistical method in the UpSquad social page for asynch prep points. | |

In addition to the undergraduate student participants, the ASRI involved 52 presenters and panelists, which included graduate students, faculty, staff and entrepreneurs.

**Institutional Affiliation of ASRI Presenters/Coaches**

| Institution | Number | Percent |
|---|---|---|
| AR EPSCoR | 1 | 1.9 |
| AR Lyon College | 3 | 5.8 |
| Arkansas State Crime Laboratory | 1 | 1.9 |
| Arkansas State University | 4 | 7.7 |
| Arkansas Tech | 7 | 13.5 |
| ASMSA | 4 | 7.7 |
| Industry-ABF Freight | 1 | 1.9 |
| Industry-Blue Cross Blue Shield of Arkansas | 1 | 1.9 |
| Industry-Center for Toxicology & Environmental Health | 1 | 1.9 |
| Industry-JB Hunt | 1 | 1.9 |
| Industry-Walmart Global Tech | 1 | 1.9 |

| | | |
|---|---|---|
| National Park College | 1 | 1.9 |
| Non-Profit-Community Health Centers of Arkansas | 1 | 1.9 |
| Non-Profit-Heartland Forward | 1 | 1.9 |
| University of Arkansas for Medical Sciences | 7 | 13.5 |
| University of Arkansas-Fayetteville | 11 | 21.2 |
| University of Arkansas-Little Rock | 3 | 5.8 |
| University of Central Arkansas | 3 | 5.8 |
| Total | 52 | 100.0 |

About one-fifth (21%) were from the University of Arkansas at Fayetteville; 14% from Arkansas Tech; 14% from the University of Arkansas for Medical Science; 8% from the Arkansas School for Mathematics, Sciences, and the Arts; 8% from Arkansas State University and 10% representing industry. The post-event evaluations showed that the event was a success, and has contributed to students' confidence and ability in data science and research skills. Additional evaluation data can be provided upon request.

ASRI organizers Holden & Krakowiak attended and presented about the ASRI at the 2023 National Association of African American Studies & Affiliates Conference in Arlington, TX. The title of the presentation was "*Arkansas Summer Research Institute: The Evolution of a Hands-on Experiential STEM Training Program in Response to COVID-19 with an Emphasis on Supporting Students from Underrepresented Populations."* This conference presentation addressed program details, innovations, and outcomes of the Arkansas Summer Research Institute (ASRI), particularly in regard to underrepresented minority students. The ASRI improves confidence in data science skills in participating students who also indicate broad satisfaction with the online model of program delivery. While students of any race or ethnicity may participate, targeted recruitment and retention methods improve representation of students traditionally underrepresented in STEM careers. The NAAAS conference attendees were interested in this information, provided valuable feedback, and may also be helpful in future recruiting efforts, as the majority of participants at the conference were individuals from underrepresented groups. The team is also preparing a paper for submission in an academic journal for STEM education.

The 2023 event is planned for June 1-19, 2023 and is accepting applications now. At the time of this report, more than 300 students from 45 campuses have applied. We will provide updates on the 2023 event in the Year 4 report.

## K20 Educator Professional Development

**LET'S Prepare for the Future, in collaboration with EAST Initiative**

This program conducted a variety of professional development activities during Year 3. This program is one of our primary engagement strategies to support K12 educators. In June of 2022, EAST hosted an in-person workshop for 3D printers, followed up by two webinars for troubleshooting and support. The workshop is offered for continuing education credits to Arkansas educators, and teachers who attended the workshop were provided a 3D printer and related supplies to take back to their classrooms. The workshop was focused on operational basics of the machines and consumables, publicly available curricula and design resources, the full print process from finding the template to the end product, and printer maintenance. 28 teachers representing 22 school districts, including 22 female teachers and 3 URM teachers, participated in the 3D printing sessions. Post-survey data from attendees is available on request.

In June 2022, EAST offered another workshop under this program to teach educators how to design and program with Pi-Tops. Pi-Tops are a small computer interface designed to facilitate use of raspberry pi devices. Attendees received Pi-tops to take to their classrooms. 17 teachers representing 14 school districts, including 13 female teachers and 2 URM teachers, attended the in-person training and follow-up webinars. Finally, three virtual reality workshops were held in different locations around the state under this program. A total of 57 teachers attended and learned how to integrate virtual reality and augmented reality in the classroom. 35 of the attendees were female, race and ethnicity were not reported.

The online community for participants in the LET'S program is being established currently, the vendor contract is in place and the community will be available for all participants by August of 2023. Educators who attend workshops under this program will be invited to the community to share feedback, update curricula, share resources, and troubleshoot.

## Career Development Workshops (CDW)

We do have some workshops to report that took place since the last report, and a number of events scheduled for Summer 2023. Through the partnership with the Arkansas Department of Education and Division of Higher Education, DART co-hosted a Computing and Data Science Ecosystem workshop on May 20, 2022 which was attended by approximately 60 people from across the state. These are attended by industry professionals, educators, and representatives from educational and government institutions in computing and data science fields. A workshop on AI in Healthcare was held on March 14, 2023 and hosted by PI Fowler along with Fred Prior and Dave Ussery. Approximately 100 people attended for the presentations about crowdsourcing AI solutions in healthcare and to learn how clinicians are incorporating AI tools. On February 22, 2023, The ED team co-hosted a virtual workshop with the Arkansas Department of Education for educators statewide on K12 Data Science competencies and careers, which was attended by approximately 30 people. On February 15,

2023, DART hosted a grantsmanship workshop for DART faculty with NSF program officers Wendy Nilsen and Sylvia Spengler from the CISE directorate.

Future workshops that are being scheduled for this summer include a grantsmanship workshop with the NSF CBET directorate, and a series of events that are being organized in collaboration with the NSF-funded STEM Learning and Research Center (STELAR) focused on the ITEST program, as well as a workshop about CHIPS+ related opportunities including the Economic Development Administration Tech Hubs program and the upcoming NIST Incentives for Semiconductor R&D facilities.

## Communication & Dissemination

In January 2023, the project held its first in-person faculty and graduate student retreat at Winthrop Rockefeller Institute. The

DART participants gave at least 60 presentations including invited talks and posters at local, regional, and national events. Some examples include:

- Hong Cheng, presented "Modified Topological Image Preprocessing for Skin Lesion Classifications" at the 26th International Conference on Image Processing, Computer Vision, & Pattern Recognition (IPCV'22).
- Ningning Wu, presented "A Study on Personal Identifiable Information Exposure on the Internet" at the International Conference on Computational Science and Computational Intelligence.
- Student Xiatong Sun, presented "Significance level effects on variable selection and information criteria" at the 2022 INFORMS Annual meeting.
- Student Adetola Odebode, presented "Big Data Framework for Improving Disaster Response Logistics and Assessing Transportation Infrastructure in Near Real Time" at the Institute of Industrial and System Engineers (IISE) Annual Conference & Expo 2023.
- Stephanie Byrum, presented "Multi-omics data integration reveals correlated regulatory features of triple negative breast cancer" at the Intelligent Systems For Molecular Biology (ISMB) conference.
- Se-Ran Jun, presented "Bioinformatician-initiated translation research: real-time surveillance of antibiotic resistance and antibiotic resistance mechanisms" at the 35th Annual US-Korea Conference on Science, Technology and Entrepreneurship (UKC).
- Nitin Agarwal, presented a keynote talk at the 2022 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM 2022).

## Broadening Participation

The project team has some concerns about our ability to effectively implement broadening participation strategies due to contemporaneous political changes in the state. We are in the process of evaluating risks and developing mitigation strategies. EOD Hillyer intends

to compile a list of partners and organizations who are not as directly influenced or regulated, and we will figure out some strategies to direct resources to those partners for impactful programming. We also are aware that we may need to rethink some of the language that we use around DEIA efforts. In January 2023, PI Fowler and EOD Hillyer met with the recently appointed Chief Diversity Officer for NSF, Dr. Chuck Barber. Dr. Barber provided some great feedback and talking points to help us navigate this climate. He also committed to sharing materials and resources developed with the Department of Navy, that provide actionable frameworks and metrics for equity.

As noted in previous reports and communications with Dr. Small, we as a project try to foster an inclusive environment to all participants. The demographic categories used in Table B, which are the same as the US Census, are not very inclusive and can be confusing to foreign participants. When we collect the demographic information presented below, we offer a wider selection for participants to choose and identify themselves. When the self-selection matches one of the census categories, we report that participant as such on Table B. When the self-selection does not match or is not available as an option on the census, we report the participant as "Prefer not to say" or "Other," depending on the available information. Screenshots of the selections we offer participants for self-reporting demographics are included below for reference. Graphical representations of participant demographics are also included below, as well as details regarding progress towards the goals set in the project's broadening participation plan. The plan itself is included as an Appendix to this report. This information is aligned with NSF's definition of URM to include: Black or African American, Hispanic, Latino, Chicano, Native American or other Indigenous peoples, Native Alaskan, Native Hawaiian, Pacific Islander, Filipino.

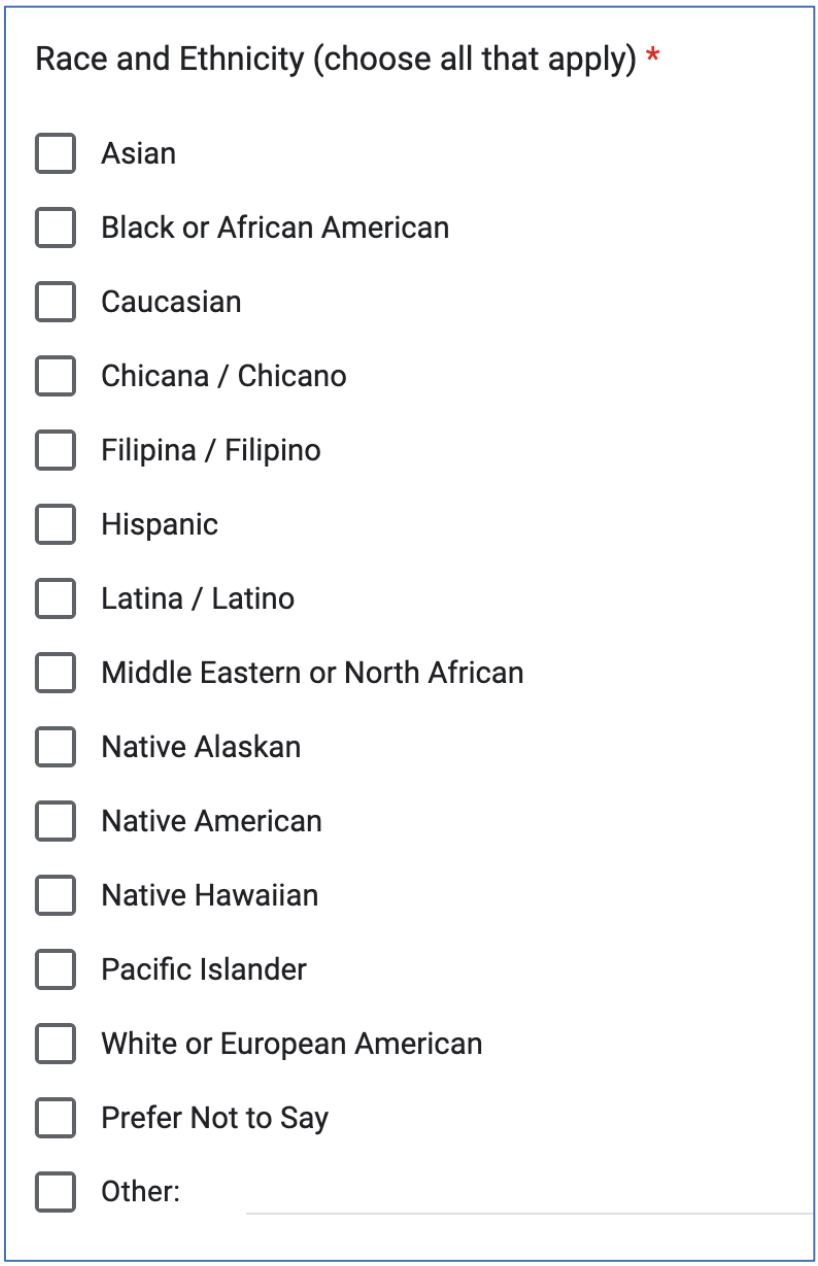



***Figure 1*** Offerings presented to participants for self-reporting race and ethnicity.

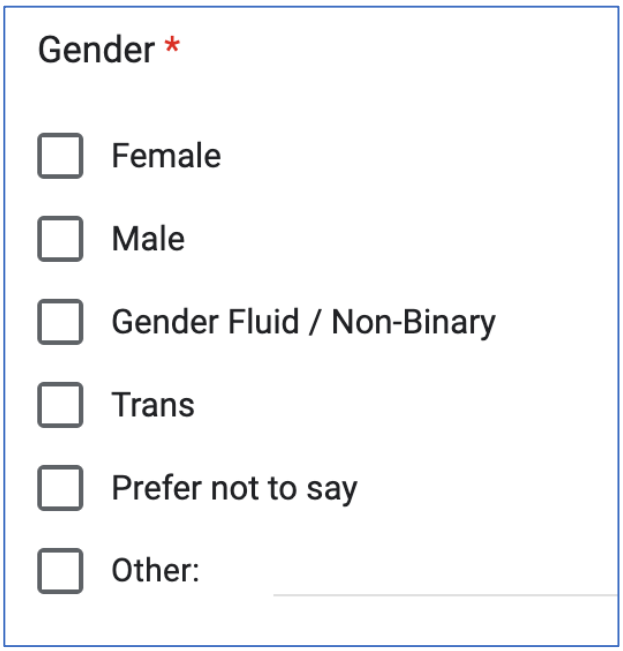
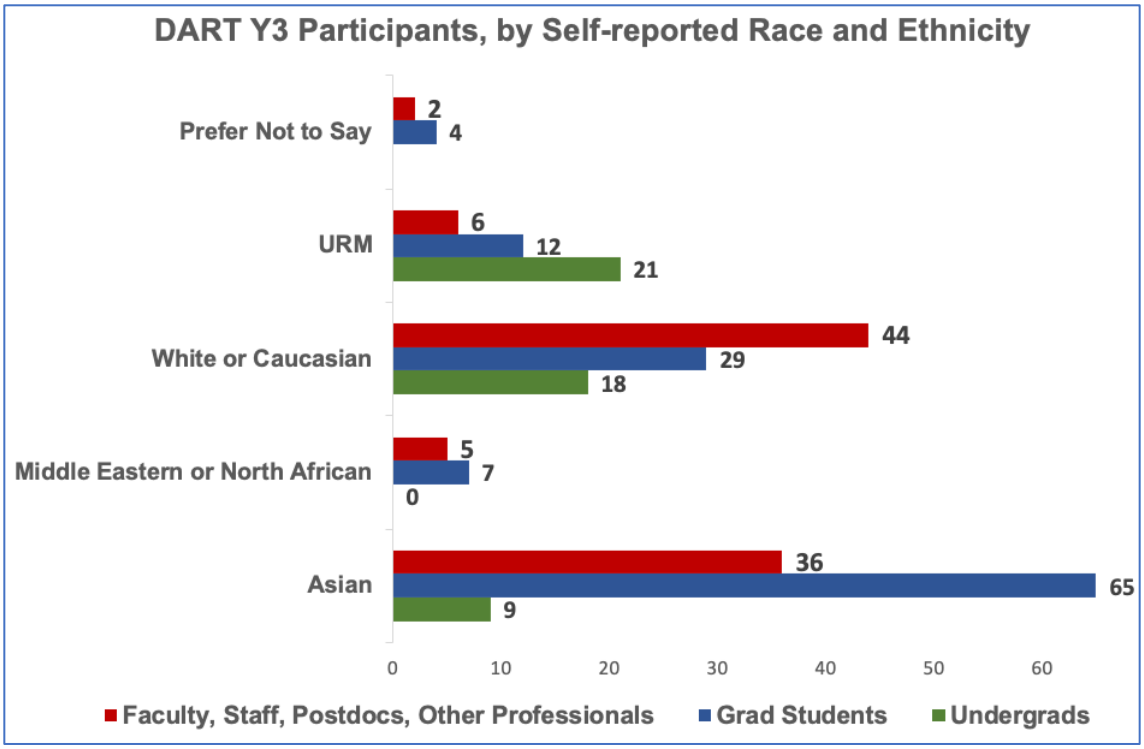
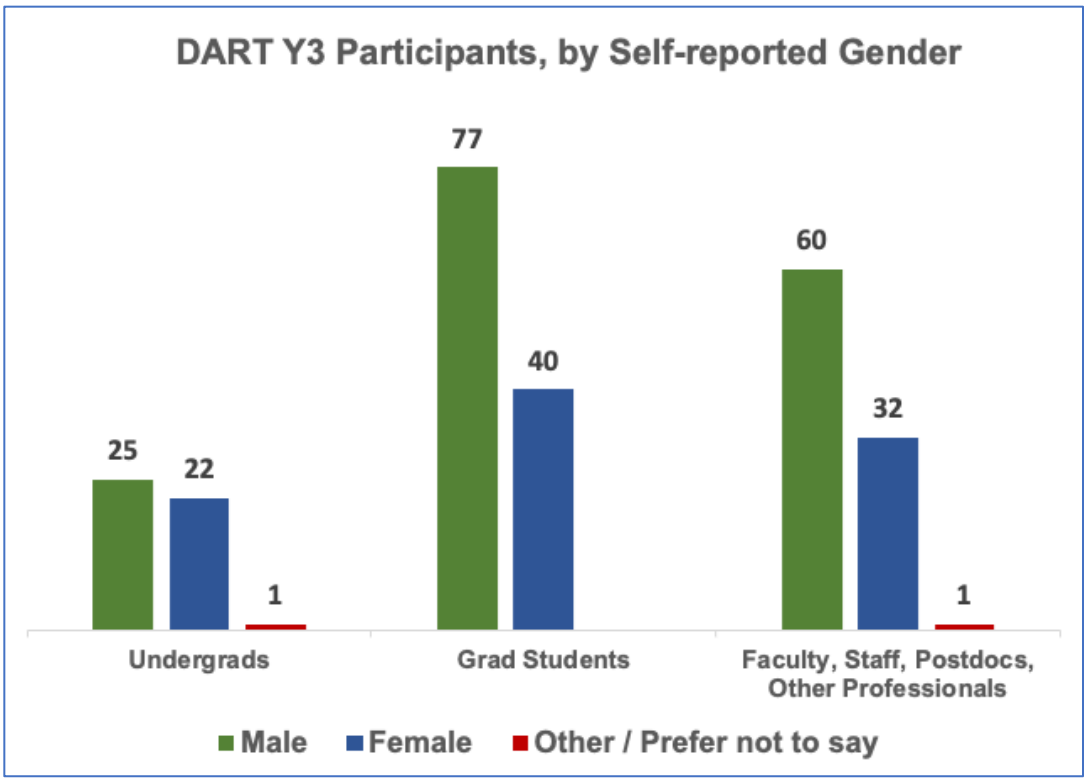


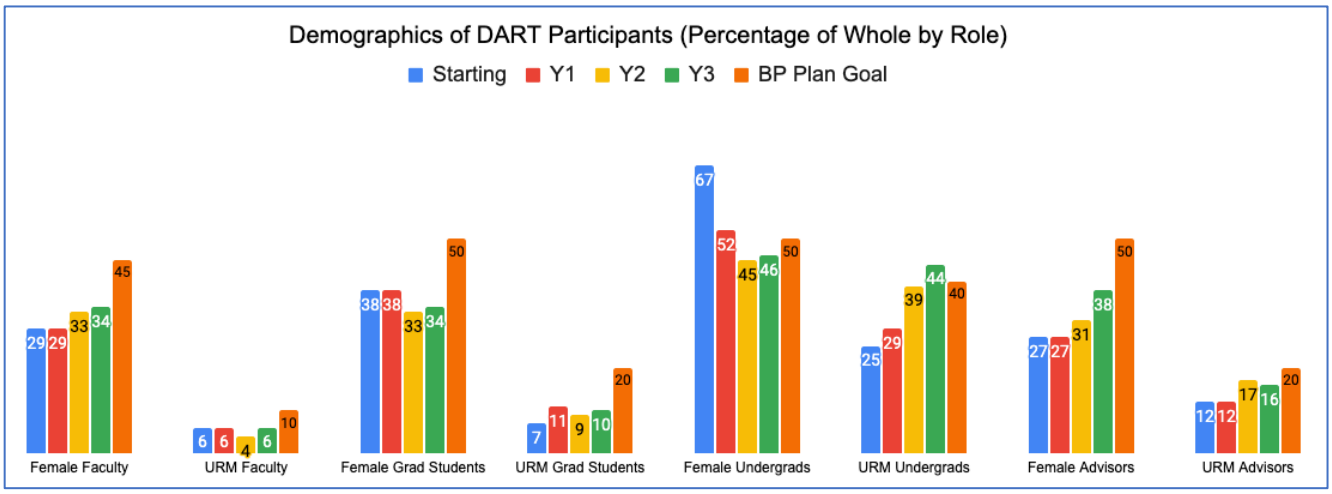

***Figure 2*** Offerings presented to participants for self-reporting gender.

The project participants during Year 3 included 6 underrepresented minority (URM) faculty, staff, or postdoctoral associates, 12 URM graduate students, and 21 URM undergraduate students. The male-to-female ratio of undergraduate participants is nearly even, though the graduate student and faculty participants are still largely male.






The chart below shows the starting and yearly participant demographic reporting in percentages, compared to the goal set for that role type in the project's broadening participation plan. Progress has been made at increasing the number of female faculty and advisors, with those goals nearly met. 34% of DART faculty are female, and 38% of DART advisors are female. The percentage of female graduate students and undergraduates both experienced a drop in Year 2, with some progress in Year 3. The goal of 40% URM undergraduates was met in Year 3, and the goal of 20% URM advisors is nearly met, with 16% currently.



# Special Conditions

## Jurisdiction-Specific PTCs

Details regarding the JSPTCs have been addressed in the earlier report sections, pertaining to cyberinfrastructure and Broadening Participation.

## Feedback from External Advisory Board

The feedback provided by our project's EAB has been addressed in numerous report sections, and will be noticeable in the pending request to revise the strategic plan.

## Tabular Representation of Progress to Date

The stoplight tables have been included embedded within each respective report section above.

## Expenditures and Unobligated Funds

As reported in Table F, the Year 3 expenditures reached 93%of the budget, and cumulatively for the project we have reached 80% expenditures.

# Appendices

## Appendix 1, DART Broadening Participation Plan

# RII Track-1: Data Analytics that are Robust and Trusted (DART): From Smart Curation to Socially Aware Decision Making

## Broadening Participation Plan

**Introduction**

The Track-1 project DART has convened a large group of participants including undergraduate and graduate students, early career and tenured faculty, administrators, staff, and K12 educators. We firmly believe in the importance of diversity, equity, and inclusion, and acknowledge the difference in those three terms.

To remain updated on best practices in broadening participation and mentorship, DART's leadership team and external evaluator will conduct quarterly reviews of leading organizations such as Advancing Research in Society (ARIS) and the Center for the Improvement of Mentored Experiences in Research (CIMER), as well as resources such as the National Academies Science of Effective Mentorship in STEMM guide and the Institute for Broadening Participation Mentor Manual. These resources were used in the development of this plan.

**Participant Diversity**

DART's leadership will take into consideration the institutional, gender, ethnic, and other forms of diversity of all participant groups and role types, including established panels. DART does not plan to hire new faculty or postdoctoral associates but will utilize the planned activities described here to broaden participation in the project. We will also encourage recruitment of women and underrepresented minority (URM) faculty and students for open positions. We will commit to the following targets for each role type:

- Faculty- 45% female, 10% URM
- Graduate Students- 50% female, 20% URM
- Undergraduate Students- 50% female, 40% URM
- Advisory Boards- 50% female, 20% URM

Starting Participant Diversity*

- Faculty- 29% female, 6% URM

- Graduate Students- 38% female, 7% URM
- Undergraduate Students- 67% female, 25% URM
- Advisory Boards- 27% female, 12% URM

** As self-reported by participants*

Please see Appendix 1 for starting participant diversity by role, gender, and race/ethnicity.

**Mentorship Program**

We recognize the importance of mentorship in the formation of science identity and retention of students and early career faculty in STEM. DART plans to utilize individual development plans for summer (SURE) students, undergraduate research assistants (UGRA), graduate research assistants (GRA), and early career faculty. An Individual Development Plan (IDP) is a personal action plan designed to help students and postdocs clarify their academic responsibilities and expectations and take more ownership of their professional development. IDPs can be a useful advising tool, helping mentors and mentees align their goals and expectations, identify areas for improvement, and track progress.

During Year 1, DART will adapt versions of IDP templates for each of the four roles listed above. The IDP will be implemented starting in Year 2 at the beginning of the project participation for each role and reviewed at the end of each person's participation period, culminating in a survey. The SURE students will complete a skills assessment before and after the experience, and develop their IDPs with mentors during the ASRI. Graduate students will complete their IDPs at the Annual Retreat with their mentors, and evaluate it annually. Early career seed grant faculty will complete theirs at the beginning of their awards and evaluate them annually and/or at the end of the seed project. Interviews and focus groups will also be facilitated by the external evaluator to obtain qualitative feedback from both mentors and mentees regarding the mentoring experience.

Mentors will also complete an annual survey to reflect on their growth as mentors. DART leadership will review assessment data to continually improve mentee and mentor support. The central office will also support campus level efforts to implement or strengthen relevant mentorship programs.

**DART Research Seed Grant Program**

DART will fund research seed grants in emerging or transformative areas of research that align with, but do not overlap DART planned research activities. Seed grant proposals will be reviewed by external advisory board and industry advisory board members. Applications will be encouraged from underrepresented minority faculty, postdoctoral associates, and early career faculty. DART will utilize a number of communication channels to raise awareness of the seed grant program.

Each seed grant applicant will be required to submit a letter of support from an existing DART participant. Upon award, the SSC will assign mentors as appropriate to the awardees. Awardees will be invited to present their work during DART meetings such as the monthly webinar, Annual Conference, and Annual Retreat. The central office will foster collaboration discussions among seed grant awardees and DART participants. Milestones have been included in the strategic plan, and metrics have been included in the evaluation plan.

**Career Development Workshops**

DART will host at least three career development workshops annually with three rotating topics: mentorship, grantsmanship, and science communication. These workshops will be open to all project participants and free to attend. The central office will recruit presenters for these workshops from national leading organizations in the fields, or agency program officers as appropriate. Information on these workshops will be distributed through the DART faculty and student listservs, as well as social media and other channels. The workshops will be recorded with recordings posted on the @arepscor YouTube channel. Surveys will be distributed to workshop attendees which will provide feedback to inform the program. Milestones have been included in the strategic plan, and metrics have been included in the evaluation plan.

**DART Summer Undergraduate Research Experiences (SURE) Program**

In addition to the 15+ undergraduate research assistantships that are funded through the project, DART will fund summer undergraduate research experiences (SURE), for students belonging to groups that are underrepresented in computer science, information science, and data science related fields (as defined by NSF CISE). Students will be recruited through established campus-based programs such as the UARK Engineering Career Awareness Program (ECAP), UAPB STEM Academy, UAMS diversity office, and by faculty at participating institutions. DART faculty will apply for funds to host these students for 8 weeks, with a limit of $8,000 per award. $80,000 annually has been budgeted for this program. Funds will support student stipends, housing, student-specific supplies, and in-state travel. At the beginning of each project, the student will complete an individual development plan (IDP) in collaboration with their faculty host. Upon completion of their projects, students will review their IDPs and complete a survey about their experience. Each SURE student will be invited to participate in the DART Student poster competition and present their projects at DART monthly seminars. The SURE students will remain in contact with the DART project and will receive invitations to all future professional development opportunities. Milestones have been included in the strategic plan, and metrics have been included in the evaluation plan.

**Arkansas Summer Research Institute (ASRI)**

The ASRI will be hosted in partnership with the Arkansas School for Mathematics, Sciences, and the Arts (ASMSA). The ASRI is an intensive professional development experience for STEM students (seniors in high school up to graduate students). This 2-week event is of no cost for attendees and will provide training on technical skills and career skills. The main goals

of ASRI are to a) build a diverse support network of peers for STEM students in Arkansas and b) provide summer bridge professional development to increase retention in STEM. Each year's program will be evaluated by the external evaluator. Participants will be added to ASRI alumni Facebook and LinkedIn groups to facilitate longitudinal tracking and communications. DART will conduct targeted outreach to recruit URM students to attend the ASRI. Students will be recruited at all Arkansas higher education campuses, on social media, through email, and in-person recruiting events. DART will also leverage existing and new relationships with the Arkansas Louis Stokes Alliance for Minority Participation (LSAMP), the McNair Achievement Program, local diversity offices, and other organizations that are connected with URM students. DART will utilize participant survey data to inform and improve the ASRI each year. Milestones have been included in the strategic plan, and metrics have been included in the evaluation plan.

**Broadening participation Seed mini-grants**

Broadening Participation Seed Grants: DART will solicit proposals for project related mini-seed grants for education, outreach, and broadening participation. Eligible entities will include school districts, post-secondary institutions, educational service co-ops, non-profits, or other entities supporting data science and computer science education and outreach activities in Arkansas. The central office will manage the solicitations which will be posted on the AEDC website, DART website, and emailed through higher education channels. The proposals will be reviewed by the central office, and outside experts as needed. Review criteria will include the audience or participants served and evaluation of the impact of the proposed activity. One awardee will be selected each year to present their project at the annual All-Hands meeting. Reports from these projects will be maintained in ER Core and the central office will foster collaboration discussions among awardees and DART participants. Milestones have been included in the strategic plan, and metrics have been included in the evaluation plan.

**DART Starting Participant Demographics: Role Type, Gender, Race/Ethnicity**

DART leadership has reported the participant demographics below according to how the participants identified, which does not necessarily match the official NSF demographic categories in Participant Table B that is submitted with the annual report. During reporting, if a participant identifies with a category that is present on Table B, they will be assigned to that category. In past projects, we felt confined to rigid U.S. concepts of race and ethnicity, and were not able to report the true diversity of the participants. Many foreign-born participants don't know which options to choose or are not represented on Federal demographic surveys. We recognize we must report in accordance with NSF guidelines and will provide information both in Table B and in the report narrative with the self-reported demographic terms.

Various populations are considered underrepresented in computing by NSF CISE including women of any race or ethnicity, Black or African Americans, Hispanic or Latinos, American Indians, Alaska Natives, Native Hawaiians, Native Pacific Islanders, and persons with disabilities. DART participants also include one faculty member and one graduate student

who are veterans, nine first-generation graduate students, and one first-generation undergraduate student.

| | | Faculty | Graduate | Undergraduate | Category Total | Percent of Whole |
|---|---|---|---|---|---|---|
| GENDER | Female | 15 | 11 | 8 | 34 | 36.56% |
| | Male | 36 | 18 | 4 | 58 | 62.37% |
| | Prefer Not to say | 1 | | | 1 | 1.08% |
| RACE AND ETHNICITY | Asian | 22 | 11 | 2 | 35 | 37.63% |
| | Asian, Caucasian | | 1 | | 1 | 1.08% |
| | Black or African American* | 1 | 1 | 3 | 5 | 5.38% |
| | Caucasian | 11 | 3 | 1 | 15 | 16.13% |
| | Caucasian, Native American* | 1 | | | 1 | 1.08% |
| | Caucasian, White or European American | | 1 | 3 | 4 | 4.30% |
| | Hispanic* | 1 | | | 1 | 1.08% |
| | Hispanic, Latina / Latino* | | 1 | | 1 | 1.08% |
| | Middle Eastern or North African | 2 | 3 | | 5 | 5.38% |
| | Middle Eastern or North African, White or European American | | 1 | | 1 | 1.08% |
| | Prefer Not to Say | 3 | | | 3 | 3.23% |
| | White or European American | 11 | 7 | 3 | 21 | 22.58% |
| | TOTAL | 52 | 29 | 12 | 93 | |
| | | | | **Total Female** | **34** | |
| | | | | **Total URM*** | **8** | |