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

External Evaluation Report

July 1, 2020 to February 28, 2022

Prepared for:
Arkansas Economic Development Commission

Author:
Kirk F. Minnick

Minnick & Associates, Inc.
PO Box 820 Tijeras, New Mexico 87059
505-889-9358 Fax: 505-212-5842
Kminnick@EvalTeam.com

This document is supported by federal funds from the National Science Foundation (NSF), Office of Integrative Activities (OIA), contract number (OIA-1946391). Its contents do not necessarily reflect the views or policies of NSF, and no official endorsement of material should be inferred.

Introduction

Arkansas Data Analytics that are Robust and Trusted (DART) is funded by the National Science Foundation (NSF) and is designed to help fulfill the foundation's mandate to promote scientific progress nationwide. The EPSCoR program is directed at those jurisdictions that have historically received lesser amounts of NSF Research and Development (R&D) funding. Twenty-five states, the Commonwealth of Puerto Rico, Guam, and the U. S. Virgin Islands are currently eligible to participate. NSF establishes partnerships with government, higher education and industry that are designed to effect lasting improvements in a state's research infrastructure, R&D capacity and its national R&D competitiveness.

DART was awarded funding of $20 million for five years beginning July 1, 2020. The overarching goals and focus are outlined in the proposal and reiterated below:
DART will develop:

1) the means to increase the speed and efficiency of data curation and labeling,
2) techniques to protect privacy and impartial content,
3) methods for harnessing the predictive power of machine learning while increasing the interpretability of the processes behind the predictions, and
4) data science curricula that are more inclusive and better prepare students for a data-centric future.

The project is organized into nine components:

1. Coordinated Data Science Infrastructure (CI)
2. Data Life Cycle and Curation (DC)
3. Social Awareness (SA)
4. Social Media and Networks (SM)
5. Learning and Prediction (LP)
6. Education (ED)
7. Workforce Development (WD)
8. Communications & Dissemination Component (CD)
9. Broadening Participation (BP).

Evaluation Overview

Success of DART will be evaluated using a comprehensive assessment plan with both quantitative and qualitative methods. The plan includes independent external experts who will monitor progress and regularly review and report on each of the programs to the Project Management Team. The Management Team (MT) will ensure the program components are collecting the data needed by EPSCoR and the external experts to assess and evaluate the impacts and achievements of the award. Report recommendations will be used by the MT during the annual review of the strategic plan.

The evaluation will assess the progress of DART in reducing the fundamental barriers to the application and acceptance of data analytics in using the growing array of tools used in data analytics. These include reducing the barrier to the practical application and finding solutions to:

1) Big data management,
2) Security and privacy, and
3) Model interpretability.

Kirk Minnick, Minnick & Associates, Inc., will oversee the overall project evaluation. Formative data will be collected in a variety of ways. Participant surveys will provide "customer satisfaction" feedback on project activities and will help identify participant needs for future events/activities. Most surveys are administered online through email invitations to participants. Observations by the evaluator are another means by which formative data are collected. The evaluator will observe and participate in many of the project activities;

including, but not limited to, attending annual meetings and planning meetings, as well as outreach activities. These observations provide the context for participant survey feedback and can help activity leaders make real-time changes during the activity.

Progress/summative data will be collected from a variety of sources. Project data on participants, proposals, awards, presentations, publications, collaborations (individual and institutions), products, and patents are collected through a secure web-based online portal developed for project participants to report EPSCoR activities. These data will be used for project monitoring, to provide some of the basic output and outcome measures used in tracking progress and summative results, to assess the success of researchers in competing for funding, and to track faculty and student outcomes.

Evaluation Process

Data for evaluation metrics and for reports to NSF-EPSCoR will be collected electronically. DART has implemented the EPSCoR Reporting Core (ER-Core) which provides secure access to project participants for reporting and assessment.

Both the quantitative and qualitative data will be collected using a variety of methodologies. The formative evaluation will focus on the component development and the process of implementation, including feedback from students, faculty/researchers and our own observations. The Evaluation Team will collect and evaluate formative data to assist project leadership in assuring quality of program management and effective project development and implementation. An effective formative evaluation is essential to identifying organizational and structural areas that may enable or inhibit progress towards project goals. Data such as meeting minutes, communications/correspondence, project documentation, interviews/observations, surveys and participant feedback will help inform the formative evaluation.

The program leadership, External Evaluator, External Advisory Board (EAB) and Industry Advisory Board (IAB) will monitor and assess program activities. Project leadership is responsible for ensuring that data is collected on milestones, participants, proposals/awards, publications, and other project outputs, as well as implementation of recommendations from the EAB, IAB and evaluator. The evaluator will develop and administer surveys, observe project activities, analyze data, and produce evaluation reports. The EAB and IAB will assess overall results and progress towards objectives and identify problem areas. They will conduct annual site visits and produce a report that will be forwarded to the NSF Program Officer.

Minnick & Associates, Inc. will conduct the external evaluation using formative and summative evaluation strategies with both quantitative and qualitative data. The formative evaluation will focus on the development and implementation process, using feedback from researchers and workforce program participants and observations of project events. The formative evaluation will enhance program quality and project development and implementation. Formative or implementation questions modeled after those in the NSF User-Friendly Handbook for Project Evaluation (2010) will guide the formative evaluation: Are the activities being implemented as planned with adequate resources/materials /equipment? Is there a diversity of participants, institutions, disciplines and regions involved? Are the activities being developed and implemented according to the proposed timeline?

The progress/summative evaluation data will assess data collected from a variety of sources, including project data on participants, collaborations (individual and institutions), new software tools, methodologies, presentations, publications, proposals, awards, and patents. These data will support project monitoring; provide output and outcome data for tracking progress and summative results; and allow the team to track faculty and student outcomes. Annual evaluation reports will summarize project outputs and outcomes, as well as findings and recommendations. Data will be presented and analyzed longitudinally by year and cumulatively, so that any problems in achieving annual and final project targets will be apparent.

The summative evaluation will assess key progress and outcomes questions: Are the research areas advancing the efficacy of data analytics research? Are collaborations being expanded among people, institutions and disciplines within and across Arkansas? Is the project broadening participation of people, institutions, and organizations in STEM? Are the workforce activities producing a diverse group of next generation data scientists that can apply these new technologies in academic, industry and government?

The outcomes or summative evaluation is designed to assess the effectiveness of the project at attaining its stated goals, as well as those that were not intended. The evaluator looks at indicators or metrics that can inform the project and NSF whether the project is achieving its outcomes. Indicators or metrics are used because project outcomes are often difficult to measure directly.

Outcome evaluation indicators/metrics for research consortiums typically involve looking at change over time in terms of research competitiveness; degree and type of leveraging; participation of under-represented groups; workforce development; and the generation of new tools and discoveries. In addition, the evaluator attempts to assess the degree to which the project activities were responsible for any changes that occurred and to identify any unintended outcomes that resulted from the project activities.

The evaluation is designed to answer the following progress and outcomes questions:

- Are DART researchers becoming more competitive for R&D funding?
- Is the DART research generating knowledge that is being disseminated and applied in academia, industry and government?
- Are state and regional collaborations being fostered that promote research, innovation, and benefit society?
- Is the DART broadening participation of its people (especially those historically underrepresented), institutions and organizations in STEM?
- Is Arkansas capitalizing on the investment to further develop experimental programs in data analytics?

Evaluation and Data Collection Methods

Data for the evaluation will come from a variety of sources and will be collected using a variety of methodologies, including both formative and summative evaluation strategies. The following provides a sample of the data sources and methodologies that will be used.

Table I: Evaluation Data Sources

Data Source	Description	Purpose
Feedback Surveys	Participants in DART activities/events will be asked to provide feedback on the activity in which they were involved. Most surveys will be administered online through email invites to the participant.	These provide "customer satisfaction" feedback for project activities, help the project identify participant needs for future events/activities and assess the efficacy of materials and displays.
Observations	The evaluator will observe and participate in many of the project activities. These will include attending student outreach and faculty workshops, All Hands Meetings and project meetings.	These observations help provide context for participant survey feedback and can help activity leaders make real time changes during the activity.
Project Documents	Project reports and other written documents, such as the strategic plan; provide the evaluator with the research plans and accomplishments.	These documents provide the plans and accomplishments in a more holistic manner, which can be compared to the reported project data.
Project Data (ERCore)	Project data on participants, proposals, awards, presentations, publications, collaborations, etc. are collected by DART through the ERCore portal.	These data provide some of the basic outputs and outcome measures used in tracking summative results and reviewed for consistency and reliability.
Group/individual interviews	Interviews will be conducted with individuals and groups of faculty to gather feedback on the project.	These interviews provide a perspective on the project from the field.
NSF Awards Database	This database provides the proposals funded in AR, along with the funding amounts and duration. It also contains the names of the PI and Co-PIs.	These data provide the data to assess the success of Arkansas researchers in competing for NSF funding and track AR EPSCoR faculty and student success.
Web of Science: Citation database	Web of Science provides access to a database of scientific publications and citations that can be searched by individual author, institution, or state	These data provide the number/citations of publications by participants. These data can be examined by time period as well as by eywords.

Formative/Implementation Evaluation

The formative or implementation evaluation examines how the project proposed activities are being implemented. The evaluators examine how the activities are conducted; including the timing, participants involved, location, venue, content, project resources utilized, alignment with proposal, as well as the immediate outputs from those activities. This aspect of evaluation is designed to help improve future project activities by soliciting feedback from participants, conducting observations and assessing the alignment between the proposed activities and implementation. Process evaluation metrics typically involve number and diversity of participants; participation satisfaction; number of proposals, presentations and papers submitted; number of collaborations; amount of equipment purchased and when; as well as other immediate outputs resulting from project activities. Formative evaluation feedback to project leadership is designed to help improve future activities and identify any discrepancies between proposed activities and timelines and implementation, including the involvement of the appropriate participants. The purpose of the formative evaluation is to help the project have the best implementation of its planned activities, so that it has the best opportunity to achieve its stated goals.

Table II
Formative/Implementation Evaluation

Evaluation Question	Data Sources	Frequency
Are the individual components being implemented as planned?	ER Core Annual Report	Annual
Are the appropriate staff/faculty/partners involved and working together towards the component goals(s)?	ER Core Annual Report	Annual
Are there adequate resources/materials/equipment available?	ER Core Annual Report	Annual
Are the appropriate participants selected and involved in activities/programs?	ER Core Annual Report	Annual
Do the activities/strategies match those described in the strategic plan/proposal? If not, are the changes in activities justified and described?	ER Core Annual Report	Annual
Are activities being conducted according to the proposed timelines? By the appropriate personnel?	ER Core Annual Report	Annual
Do project participants report high satisfaction in their participation in the activities?	Faculty and Graduate Student Survey	Annual

Progress/Summative Evaluation

The progress/summative evaluation will assess data that will be collected from participants through a secure web-based portal, including data on participant numbers and demographics, collaborations (individual and institutions), new software tools, methodologies, presentations, publications, proposals, awards, and patents. These data will be used to track outputs and outcomes for faculty, students, and the project, and will be analyzed both by year and cumulatively to ensure that the project is on-track to achieve annual and final project targets. The summative evaluation will assess key progress and outcomes questions, such as:

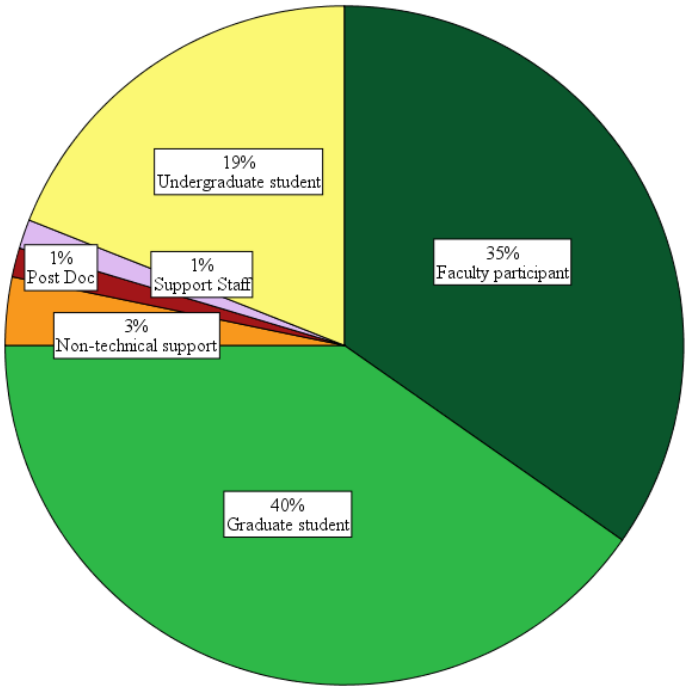
- Are the research components reducing the barriers of big data management, security and privacy and model interpretability and expanding the use of data analytics in industry and government?
- Are collaborations being expanded among people, institutions, and disciplines within Arkansas, as well as outside the state?
- Is the project broadening participation of people, institutions, and organizations in STEM with specific emphasis on data analytics? and finally,
- Are the workforce activities producing a diverse group of next generation data scientists that can apply these new technologies in academia, industry, and government?

People (Human Infrastructure, Collaboration and Diversity)
Diversity, and expanding the involvement of women and underrepresented minorities in the sciences, is one of the objectives of NSF, EPSCoR and DART. This section explores the degree of involvement of women and underrepresented minorities by level of participation, participant role and participating institutions in the initiative. The role of collaborators and the nature of their collaborations with the project will also be examined.

In this report, there are two types of participants: direct and indirect. There are direct project participants, those that have participated in a significant way in accomplishing the goals of DART; either in one of the research areas or education outreach/external engagement. Direct participants are all individuals who have been individually identified. Indirect project participants are the students of the K-12 teachers who have received professional development, members of the public or education institutions who have participated in outreach events. These latter participants are reported in the external engagement section

Participants
DART has involved 216 direct participants in the project so far. Figure 1 shows the percentage of participants by their role.

Figure 1
Participant Roles
(N=216)



Over half (59%) of the project participants are students with over one-third (40%) ‘Graduate students’ and one fifth (19%) 'Undergraduates’. Over one-third (35%) are ‘Faculty participants’ with Post Docs representing 1% of the total number of participants. The remaining participants were 'Support staff’ (4%).

It is important to examine the ethnic and gender diversity of participants by their role, as well, to ensure that diversity goals are being addressed across the different levels of participants. Figure 2 shows the number and percentage of under-represented ethnic minorities and females by participant role for all program years, as well as the unduplicated count of participants across all years. Demographic data are self-reported and do not represent all participants.

Figure 2
Participants by Role and URM* and Female*
for Program Years 1 and 2

Role	Year 1 Baseline	Year 2	Unduplicated Count
Faculty	56 (41%)	75 (35%)	75 (34%)
URM-Ethnic	2 (4%)	2 (3%)	2 (3%)
Female	16 (29%)	22 (29%)	22 (30%)
Post docs	1 (1%)	3 (1%)	3 (1%)
URM-Ethnic	0 (0%)	0 (0%)	0 (0%)
Female	1 (100%)	2 (67%)	2 (67%)
Graduate Students	47 (34%)	87 (40%)	89 (41%)
URM-Ethnic	5 (13%)	9 (12%)	10 (13%)
Female	18 (38%)	26 (30%)	27 (31%)
Undergraduates	25 (18%)	41 (19%)	41 (19%)
URM-Ethnic	9 (36%)	18 (45%)	18 (45%)
Female	12 (48%)	19 (46%)	19 (48%)
Support Staff	8 (5%)	10 (5%)	10 (5%)
URM-Ethnic	0 (0%)	0 (0%)	0 (0%)
Female	4 (50%)	5 (50%)	5 (50%)
Total	137	216	218
URM-Ethnic	16 (13%)	29 (15%)	30 (15%)
Female	51 (37%)	74 (34%)	75 (35%)

* URM and Female percentages are based on the subset of participants who self-reported their ethnicity and gender

DART involved 216 in program Year 2, an increase of 79 (58%) from Year 1. The largest increase in participants in Year 2 were in the number of graduate students from 47 to 87, an increase of 85% and in the number of undergraduates from 25 to 41 an increase of 64% and faculty from 56 to 75 an increase of 34% and across all program years. Over the two program years DART has involved 218 unique individuals: one-third (34%) faculty' over two-thirds (41%) graduate students and one-fifth (19%) undergraduates.

The overall ethnic diversity among participants was 15%, ranging from 45% among the undergraduates to 13% within the graduate student participants and 3% among faculty. Female diversity was much higher with 35% over all participants to 48% among undergraduates; 31% among graduate students and 30% among faculty.

Figure 3 presents the number and percentage of participants by institution by program year and an unduplicated count across the program years.

Figure 3

Number/Percent of Participants* by Institution and Program Year

Institution	Year 1	Year 2	Unduplicated Count
Arkansas Economic Development Commission	5 (4%)	9 (4%)	9 (4%)
Arkansas State University	7 (5%)	9 (4%)	9 (4%)
Arkansas Tech University	2 (2%)	11 (5%)	11 (5%)
North Arkansas College	1 (1%)	1 (1%)	1 (1%)
Philander Smith College	3 (2%)	7 (3%)	7 (3%)
Shorter College	4 (3%)	4 (2%)	4 (2%)
Southern Arkansas University	11 (8%)	14 (7%)	14 (6%)
University of Arkansas at Fayetteville	50 (37%)	63 (29%)	63 (29%)
University of Arkansas at Little Rock	30 (22%)	67 (31%)	68 (31%)
University of Arkansas at Pine Bluff	5 (4%)	4 (2%)	5 (2%)
University of Arkansas for Medical Sciences	11 (8%)	16 (7%)	16 (7%)
University of Central Arkansas	8 (6%)	11 (5%)	11 (5%)
All Institutions	**137**	**216**	**218**

* does not include advisory board members

The institutions with the largest number of participants are at the University of Arkansas at Little Rock (31%) and the University of Arkansas-Fayetteville (29%). Participants from the University of Arkansas for Medical Sciences represent 7%; Southern Arkansas University 6%; Arkansas Tech University 5% and University of Central Arkansas 5%. The project has involved participants from many higher education institutions, including many smaller institutions, such as Arkansas Tech University (5%); North Arkansas College (1%); Philander Smith College (2%); Shorter College (2%) and University of Arkansas at Pine Bluff (2%).

Collaborations

Recognizing that for DART to be successful at improving its research competitiveness, it must maximize the collaborations between researchers within and outside the state. Figure 4 presents the number of DART collaborators by institutional type and location.

Figure 4
Number of External Collaborators by Type of Institution and Location by Program Year

Type of Institution	Year 1	Year 2	Overall
Academic-Research Institution	5	2	7
Academic-Primarily Undergraduate	3	0	3
Industry/Business	2	1	3
National Laboratories	0	0	0
Non-Profit	1	0	1
K-12 School/Provider	2	0	2
State Agency	0	1	1
Other	0	0	0
Location of Collaborator	**Year 1**	**Year 2**	**Overall**
Within Arkansas	9	2	11
Outside Arkansas, but in US	3	2	5
International	1	0	1
Total Collaborators	13	4	17

Project participants reported 17 external collaborators during the first two years of the project. More than one-third (41%) of the collaborators are at academic research institutions, while about one-third (36%) are from primarily undergraduate institutions (18%) and industry (18%). Almost two-thirds (65%) of the external collaborators reported are in Arkansas; while one-fourth (29%) are collaborators located outside Arkansas but within in the US.

While it is important to increase the number of collaborations with other researchers, it is equally important to involved collaborators from a variety of institutions. Figure 5 presents the number of institutions in which the DART collaborators work by institutional type and location.

Figure 5
Number of Collaboratoring Institutions by Type of Institution and Location by Program Year

Type of Institution	Year 1	Year 2	Overall
Academic-Research Institution	3	2	5
Academic-Primarily Undergraduate	3	0	3
Industry/Business	2	1	3
National Laboratories	0	0	0
Non Profit	1	0	1
K-12 School/Provider	2	0	2
State Agency	0	1	1
Other	0	0	0
Location of Institution	**Year 1**	**Year 2**	**Overall**
Within Arkansas	7	2	9
Outside Arkansas, but in US	3	2	5
International	1	0	1
Total Collaborating Institutions	11	4	15

DART collaborations during the project include collaborations at 15 different institutions; one-third (33%) were research institutions, while the remaining included collaborators at Primarily Undergraduate Institutions (20%); Industry/Business (20%) and other types of institutions (27%).

Material Infrastructure (Equipment, Models and Cyberinfrastructure)
As the name implies, the Research Infrastructure Improvement (RII) funding is intended to provide the resources to enhance the research capabilities of the jurisdiction and become more competitive by acquiring equipment and funding infrastructure necessary for world-class research.

Over $1.2 million have been expended or ordered during the first two years to purchase computer infrastructure at three institutions. In Year 1, project funds were used to purchase $650k of infrastructure at the University of Arkansas at Fayetteville. The equipment purchased included DELL Fiber Splitter cables, PowerEdge XE8545, Power Edge R7525, Server, NVIDIA Ampere A100 649,607.18 and a 40-port Mellanox Quantum QM8790. This equipment is part of the DART CI Plan that increases additional hardware needed to move pinnacle out from behind the firewall. Also, in Year 1 data storage servers allowing data sharing among DART researchers were purchased for $24k at the University of Arkansas at Little Rock.

In Year 2, $496k was planned to be used to upgrade the research backbone at University of Arkansas Medical Sciences to collaborate with ARE-ON and extend service to the University of Arkansas at Fayetteville.

Research (Observing, Data Collecting, Discovery, Funding Support)
Data collected, observations or field work, research conducted, and proposals submitted and awarded provide an indicator of the research outputs accomplished through the efforts of the DART researchers. Figure 8 lists the number of proposals and amount of funding requested by funding source since DART started.

Figure 8
Proposal Success Rates and Amounts Proposed, Funded and Awarded as of February 28, 2022

Funding Source	Number Proposed	Amount Proposed	Number Pending	Amount Pending	Number Funded	Amount Funded
Agency for Healthcare Research & Quality	2	$200,000	0	$0	0	$0
Arkansas Biosciences Institute	2	$300,000	0	$0	1	$150,000
Arkansas NASA EPSCoR	1	$40,000	1	$40,000	0	$0
Arkansas Research Alliance	2	$175,000	0	$0	2	$175,000
NASA	1	$99,999	1	$99,999	0	$0
National Institutes of Health (NIH)	10	$19,056,063	7	$16,242,400	2	$2,423,526
National Science Foundation (NSF)	29	$34,660,860	10	$6,725,984	11	$5,033,313
National Security Agency (NSA)	1	$18,000	0	$0	1	$18,000
Private Companies	2	$98,992	1	$47,011	1	$51,981
University Faculty Research Grants	3	$76,188	0	$0	3	$76,188
US Census Bureau	1	$104,020	0	$0	1	$104,020
US Department of Agriculture (USDA)	2	$1,095,000	1	$1,000,000	1	$95,000
US Department of Defense (DoD)	6	$12,281,115	0	$0	4	$7,671,214
US Department of Energy	2	$2,797,797	1	$1,997,797	1	$800,000
US Department of Transportation	1	$155,199	0	$0	1	$155,199
US Geological Surveys (USGS)	1	$25,000	0	$0	1	$25,000
US Office of Naval Research (ONR)	3	$4,089,066	1	$165,540	2	$3,923,526
Overall	**69**	**$75,272,299**	**22**	**$26,318,731**	**32**	**$20,701,967**

Sixty-nine proposals requesting over $75 million have been submitted by DART participants. As of March 1, 2022, 32 proposals have been funded for a total of $20.7 million, while twenty-two are still pending. The funding agencies where most of the proposals have been submitted are: NSF (42%) and NIH (14%), while the most award dollars have come from US Department of Defense (37%); NSF (24%) and US Office of Naval Research (19%).

Figure 9 presents a list of those awards over $500,000.

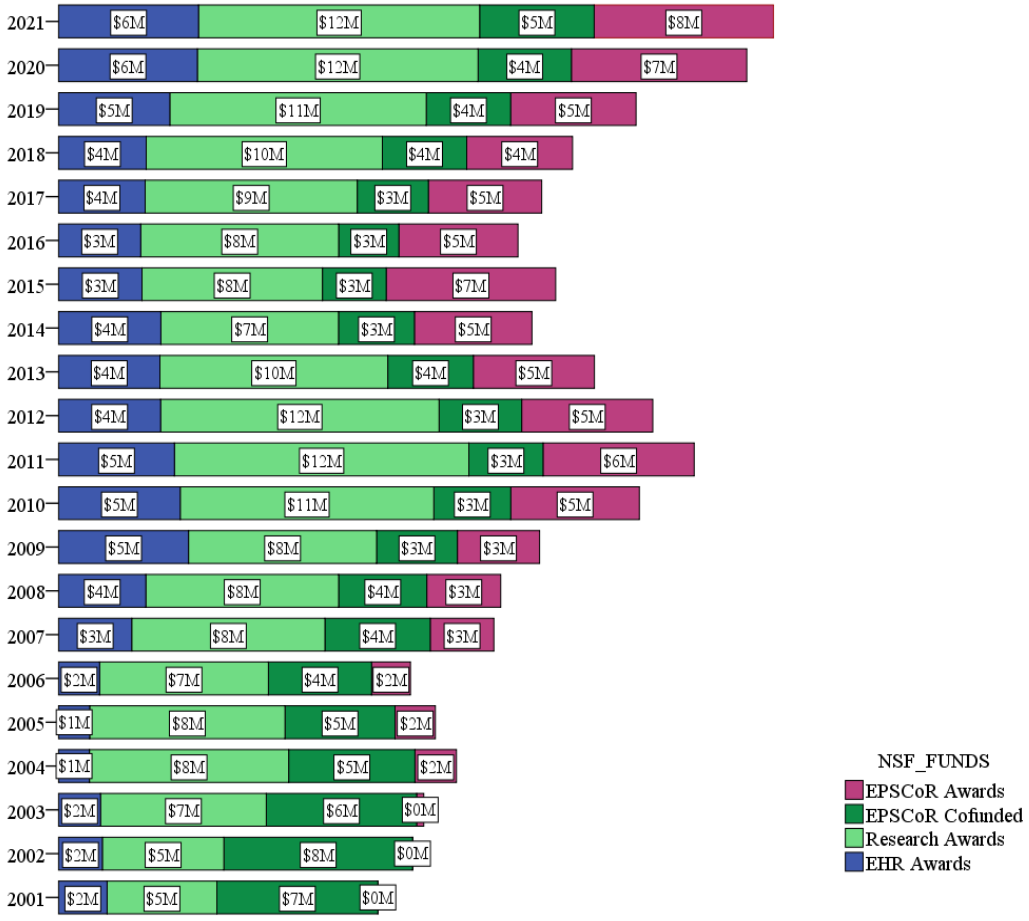
Figure 9
Awards over $500,000

Title of Award	Principal Investigator	Award Amount	Funding Source
Multi-Level Models of Covert Online Information Campaigns	Nitin Agarwal	$4,965,214	U.S. Department of Defense (DoD)
Developing Rapid Response Capabilities to Evaluate Emerging Social Cyber Threats	Nitin Agarwal	$3,773,526	U.S. Office of Naval Research (ONR)
Fusing Narrative and Social Cyber Forensics to Understand Covert Influence	Nitin Agarwal	$2,500,000	U.S. Department of Defense (DoD)
Center for studies of host response to cancer therapy	Se-Ran Jun	$2,280,000	NIH
V-INT: Automated Vulnerability Intelligence and Risk Assessment	Qinghua Li	$1,919,953	Department of Energy
RII Track-2 FEC: Artificial Intelligence on Sustainable Energy Infrastructure Network (AI SUSTEIN) and Beyond towards Industries of the Future	Haitao Liao, Xintao Wu, Xiao Liu	$1,450,003	NSF
Assessment of antibiotic resistance in fresh vegetables from farm to fork	Se-Ran Jun	$1,000,000	USDA
Photogrammetry Services, Task Order for CY2021 View Text	Jackson Cothren, Chase Rainwater	$800,000	Department of Energy
FAI: A novel paradigm for fairness-aware deep learning models on data streams	Xintao Wu	$628,789	NSF
IUCRC Phase I The University of Arkansas: Center for Infrastructure Trustworthiness in Energy Systems (CITES)	Qinghua Li	$525,000	NSF

While tracking the funding levels of DART researchers is critical to measuring the success of this award, it is equally important to track the NSF award dollars received by all researchers in the state.

Figure 10 presents the dollar amount converted to 2021 dollars of NSF awards to Arkansas for the last ten years. The award dollars are distributed over the projected length of each award to smooth out the effect of a large award causing a spike in an individual year. This technique provides a better representation of when the dollars are expended and makes it easier to track trends in funding levels, now and in the future.

Figure 10
NSF Funding Awarded to Arkansas
NSF Funded Awards in Arkansas by Calendar Year: 2001 to 2021
Awarded Funding Equally Distributed Across Project Years
(Reported in 2021 CPI Adjusted Dollars)



Note: 2010-2013 were American Recovery and Reinvestment Act years;
while 2020 and 2021 saw increased funding from Covid-19 Stimulus funding

NSF funding in the state has increased from $13 million in 2001 to $30 million in 2021, in 2021 adjusted dollars. While the years 2010 to 2013 showed marked increases in award dollars to the state, this was the result of the American Recovery and Reinvestment Act. A similar increase in Federal funding occurred in 2020 and 2021 with the Covid-19 Stimulus funding. However, Arkansas is on track to maintain a high level of award dollars coming into the state.

Since 2001, Arkansas has received over $430 million dollars in NSF funding. Most (43%) of this funding has been in the form of research awards totally $184.7 million; co-funded awards (both for research and EHR) have amounted to $87.1 million (20%); EPSCoR awards (Tracks1, 2 and 3) have amounted to $83.1 million (19%) and $83.1 million (19%) has been awarded for EHR proposals.

NSF funding for Research, excluding EPSCoR co-funding and Track1, has increased from $4 million in FY 2001 to over $11 million in FY 2020, an increase of 175%. In addition, Education and Human Resource funding from NSF has more than doubled from $2 million in FY 2001 to $5 million in FY 2020.

Knowledge Generation (Professional Presentations, Publications, Patents)

Professional presentations, posters and invited talks are critical to increasing the visibility and reputations of DART researchers, in addition to disseminating their valuable research findings to their colleagues. Figure 11 presents the number and percent of presentations made by presentation type and program year.

Figure 11
Presentations by Type of Presentations and Program Year

Type of Presentation	Year 1	Year 2	Total
Invited Speaker	18	12	**30 (37%)**
Panel	1	0	**1 (1%)**
Poster	1	9	**10 (12%)**
Presentation/Talk	22	19	**41 (50%)**
Other	0	0	**0 (0%)**
Totals	**42**	**40**	**82**

DART participants reported making 82 presentations, posters, and invited talks during the first two program years. Half (50%) of the presentations were presentation/talks, while more than one-third (37%) were as invited speakers and about one-tenth (12%) were posters.

Publications

Dissemination of the knowledge and research findings by DART researchers is also an important outcome of the project. Figure 12 lists the number of publications by level of support from DART and program year.

Figure 12
Publications by Level of Support and Program Year

Level of Support	Year 1	Year 2	Totals
Partial	7	62	69
Primary	2	18	30
Totals	**9**	**80**	**89**

Researchers reported 89 publications in the first two years of the project. One-third (34%) were reported as receiving primary support from DART. Some of the journals in which DART researchers have published included: BMC Bioinformatics; Infection and Immunity; Metabolites; Microbial Genomics and International Journal of Advanced Computer Science and Applications to name a few.

Patents

There have been no reported disclosures filed since the start of the project.

External Engagement (Scientific Literacy, Public Presentations, Policymakers, Education)
Increasing the scientific literacy and understanding of scientific research at all levels of society is important for increasing the diversity of the scientific workforce. This includes the general public, undergraduates, graduate students, junior faculty, K-12 teachers and others.

Figure 16 shows the number of people who have been engaged by type of audience with these outreach efforts during the life of this project.

Figure 16
Number of People Reached through
External Engagement Efforts for All Program Years

Type of Institution	**Academic Research Institutions**		**Primarily Undergraduate Institutions**		**Minority Serving Institutions**		**K-12 Institutions**				**Total**
Type of Audience	Faculty	Students	Faculty	Students	Faculty	Students	Teachers	Students Reached Directly	Students Reached via Teacher Training	Other	
Male	44	38	18	21	10	13	8	216	0	10	378
Female	26	29	27	12	4	13	44	265	0	10	430
Underrepresented Minority	3	3	2	8	3	23	9	256	0	0	309
Total	71	67	45	33	14	26	66	577	0	181	1080
Percent	7%	6%	4%	3%	1%	2%	6%	53%	0%	17%	

Overall, an estimated 1,080 people have been involved in one or more external engagement activities/programs supported by DART during the project. About half (53%) were K-12 students reached directly. Almost hlef (46%) of the students reached directly though DART outreach are female and 44% are a member of an underrepresented minority in STEM.

Findings by Component

1. **Coordinated Data Science Cyberinfrastructure – The Arkansas Research Platform (ARP)**

Proposed: *"The proposed research will be supported by a data science cyberinfrastructure (CI) platform capable of providing secure, distributed, agile, scalable, and on-demand services. We propose to architect and build a private cloud environment, the Arkansas Research Platform ARP, and integrate it with existing high-performance computing resources. In combination, these will provide 1) libraries of pre-configured containers designed to support a variety of well-known and novel workflows in machine and statistical learning, graph theory, bioinformatics, and geoinformatics, 2) containers configured for parallel computation and distributed memory on HPC resources for analysis of very large datasets, 3) the ability for researchers to create and share new containers and share, and 4) the ability to stream data to visualization environments both proximate and distant from the computing resources to aid in analysis and meta-analysis of experiments. ARP will be managed as a unique multi-institutional resource."*

Findings:

Table of Selected Outputs of Component

Outputs	**Year 1**	**Year 2**	**Total**
Publications	0	2	2
Joint Publications*	0	3	3
Presentations	1	1	2
Joint Presentations*	0	3	3
Workshops	0	6	6
Proposals # ($)	4 ($1.1M)	2 ($3.1M)	6 ($4.2M)
Awards # ($)	1 ($100k)	3 ($500k)	4 ($600k)
Joint Proposals* # ($)	1 ($800k)	2 ($5M)	3 ($5.8M)
Joint Awards* # ($)	1 ($800k)	1 ($95k)	2 ($895k)

*Collaborative proposals/awards/publications with other DART components

Establish the Arkansas Research Platform as a shared data science resource across the jurisdiction.
Year 1
-established CI working group and held monthly meeting
-developed CI plans for UAF and UAMS
-issued purchase orders for:
  -20 nodes dual AMD 7543, 1024GB, NVMe local drive, single PCI 40GB A100GPU
  -4 nodes dual AMD 7543, 1024 GB, NVMe local drive, four SXM 40 GB A100 GPU
  -100 Gb Inifiniband connection and 10Gb Ethernet connection
  -3 Enclosed cooled racks
-provided access to ARP through Science DMZ at UAF
-implemented Gitlab with dedicated server behind Science DMZ at UAF
-established Globus Basic server at UAF with endpoints at storage arrays at UAF and UAMS
shared via a OneDrive to participants

Year 2
-all nodes became operational with network changed to the UA ScienceDMZ available to all DART researchers
-increase of 73% in ARP resources; now have 1,374 teraflops available
-established secure method of providing access to ARP resources at UAF and UAMS
-participants gain access to interactive sessions on nodes via Open OnDemand portals, Pinnacle and Grace clusters
-access to storage arrays at UAF and UAMS available through Globus as endpoints
-UAF ITS moved to commercial cloud hosted GitLab repository
-installed Git on Pinnacle and Grace - researchers can clone or copy from DART GitLab via SSH
-re-budgeted Globus licenses in year 1 and 2 to address federated identity management solutions
-created prototype secure enclaves at UAMS and experimentation using container-based approach with Kubernetes orchestration
-developed System Security Plans to host HIPAA and Controlled Unclassified information at UAF
-created and filled postdoctoral fellow position
-publish UAF CI design and configuration document on website
-host 5 software carpentry workshops and train 2 carpentry instructors
-host 2 online ARP training sessions

Visualization for complex data in diverse data-analytics application domains.
Year 1
-none reported

Year 2
-conducted systematic literature review on advanced visualization and immersive analytics
-results internally reviewed and report is ready for publication on DART website
-incorporated data-driven modeling and photogrammetry-based visualization into virtual-reality experiences for undergraduate geoscience education
-published results of a user evaluation on different design choices of virtual field trips in Journal of Educational Computing Research
-collaborated with NY State U at Albany to examine effects of augmented-reality display on windshield of self-driving vehicles for drivers' spatial awareness
-designed a virtual collaborative space in augmented and virtual reality to improve coordination between data science educators and industry partners in Arkansas
-developed an immersive workbench to visualize and analyze environmental data collected from rural Borneo Highland for use in ecological research and education
-host 1 online advanced visualization workshop

Conclusions/Recommendations
The Coordinated Data Science Infrastructure component is meeting most of its objectives. The team has established a CI working group, issued purchase orders for: 20 nodes dual AMD 7543, 1024GB, NVMe local drive, single PCI 40GB A100GP; 4 nodes dual AMD 7543, 1024 GB, NVMe local drive, four SXM 40 GB A100 GPU, 100 Gb Inifiniband connection and 10Gb Ethernet connection, installed Git on Pinnacle and Grace, developed System Security Plans to host HIPAA and Controlled Unclassified information at UAF. They have begun work on the visualization for complex data in diverse data-analytics application domains by conducting a systematic literature review on advanced visualization and immersive analytics which is ready for publication on the DART website. They have already published results of a user evaluation on different design choices of virtual field trips in the Journal of Educational Computing Research.

The team noted that the implementation of GitLab, federated identity services for ARP and Globus had been delayed because of UAF network and policy concerns regarding IT security. The enterprise version of GitLab is being replaced with a commercial cloud hosted GitLab repository which will be available to DART researchers. The development of a federated identify service for ARP is being developed by the newly NSF funded SHARP CCI award and the need for a commercial version of Globus may be replaced by the no-cost version. Workshops on using an interactive shell to access Pinnacle are helping to increase the use of these computer resources by faculty and graduate students. A one-on-one connection between IT professionals and researchers seems the best approach for increasing HPC usage. The Year 2 objectives to create containerized Haddoop-based testbed for DC and hold an advanced visualization workshop are anticipated to be accomplished by the end of the year.

The evaluator has some concern that 70% of the Year 2 milestones are reported in the stoplight tables as "will be completed by end of reporting year". This may be the result of teaching commitments during the school year and the summer may afford researchers more time to complete these milestones.

2. Data Life Cycle and Curation (DC)

Proposed: *"Our research will aim to increase the level of automation in data curation and governance. We will explore a closed-loop data analytics approach, emphasizing need- and prediction-based data collection and transmission as well as the feedback role of current decisions on future data collection. Our research will encompass data governance, data architecture, data integration, and data quality, using ARP to implement and test tools to implement these ideas."*

Findings:

Table of Selected Outputs of Component

Outputs	Year 1	Year 2	Total
Publications	5	15	20
Joint Publications*	0	2	2
Presentations	9	11	20
Joint Presentations*	0	2	2
Workshops	3	0	3
Proposals # ($)	7 ($7.3M)	11 ($3.3M)	18 ($10.6M)
Awards # ($)	1 ($400k)	5 ($500k)	6 ($900k)
Joint Proposals* # ($)	2 ($1.5M)	2 ($5.1M)	4 ($6.6M)
Joint Awards* # ($)	0	2 ($1.6M)	2($1.6M)

*Collaborative proposals/awards/publications with other DART components

DC1: Automate heterogeneous data curation

Year 1

-implemented cluster entropy metrics in the Data Washing Machine (DWM) proof-of-concept

Year 2

-progress on Python POC for DWM code in 10th release, version 2.21; improvements to code improve precision and recall values linking 19 base data sets

-implemented robotic process to perform grid search across DWM parameters to find setting to produce best precision and recall clustering results

-research developing new unsupervised data quality assessment for data redundancy based on "cluster quality metric"

-use NLP models, BERT and RoBERTa, in a Zero-Shot Learning context, to project records into latent vector spaces and measure distances or similarities

-computed metrics used to cluster references using ML algorithms

-created visual representations of topological data analysis of selected test data sets

-converted textual data converted to vector spaces and simplified and segmented using topology-based tools

-continued implementation of collaborative/need-based data collection mechanism in disaster relief decision making

-developed basic data query and collection schemes for image data collection from Google Street View

-conducted initial experiments on damage assessment on a sample image dataset previously collected from social media

-developed a method for automatic data collection from Twitter's open API

-developed basic data query and collection scheme for tweets on a disaster event

-investigated the genetic variations underlying drug response, studied single-cell RNA sequencing (rna-seq), data of chronic-phase chronic myeloid leukemia stem cells

-developed tyrosine kinase inhibitors (TKI) to target the BCR-ABL oncoprotein

-inhibiting its abnormal kinase activity

-significantly improved CML patient outcomes
-cells with distinct responses to TKI, good vs poor, were clustered together in both BCR-ABI positive and negative cells
-showed putative transcription factors of these expressed genes revealed by single-cell regulatory network inference and clustering

DC2: Explore secure and private distributed data management

Year 1

-built and compared classification models using classical and deep learning algorithms for Alzheimer's disease prediction and biomarker identification
-created and maintained GitHub repository for the DWM including proof-of-concept code
-created a separate ground-truth file to compare results of any algorithm applied to data
-setup framework and tested to extract data from public repository of all Federal contracts and awards
-conducted preliminary study for unlabeled paired samples using the Min-Max ratio test
-documented DWM POC process global (file-level) unsupervised data cleansing methods shared with team members on DWM working paper/publication and Python code on BitBucket.org
- used artificial intelligence model trained for NLP tasks to find multidimensional vector embeddings for text record in data sample
- evaluated the similarity of two text records using a Natural Language Inference tool
-created an implementation of the transitive closure Java routines in Python
-implemented a global (file-level) data cleaning routine comparing high-low frequency tokens
-developed novel algorithm, SCAN to detect clusters, hubs, and outliers in networks
-SCAN clusters vertices based on a structural similarity measure
-published SCAN in ACM SIGKDD'07 received over 821 citations (ref. Google Scholar)
-results shows a superior performance when using novel machine learning and AI models for algorithms to automate data cleaning
-DWM Refactor codes available on BitBucket.org. DWM POC is being refactored as a Python Program.

Year 2

-automation of data cleansing expanded to include data corrections based on record-to-record comparisons with blocks and clusters
-developed 7 techniques for record-to-record correction, some can impute missing values and correct incorrectly split or joined tokens
-developed novel framework for scalable Entity Resolution using NLM, Locality Sensitive Hashing and Machine Learning
-preliminary experiment result shows promise, achieving accuracy over 95% with a nearly linear runtime
-developed computational framework integrating multi-layer genomics data to identify transcriptome and pathway dysregulations in autism spectrum disorder
-inferred regulatory networks differentially expressed in the disease sample as compared to control samples
-provides a way to reveal master regulators which position at top of regulatory hierarchies and control the transcriptional activities of many downstream genes
-established regulatory cascades approach offering a framework for revealing new disease-related genes
-can be applied and extended to study other tissues and diseases
-investigated the expression alterations of survival-related genes in various immune cell types when combining breast cancer bulk and single-cell RNA sequencing data
-helps to better understand the interactions of tumor and immune systems
-provides novel molecular prognostic markers for survival prediction in breast cancer patients
-developed method can be applied to study other types of cancer
-developed variety of genome visualization tools ("R-BioTools") used in courses and workshops living in a GitHub directory at UAF
-automated quality scores for biological sequence data applied to viral and bacterial genomes
-SARS-CoV-2 published February 2022, using automated genome quality scores in helping to cluster genome species
-automated pipeline tested for Enterococcus faecium, Salmonella enterica, and Klebsiella pneumoniae
-developed a convolutional neural network-based model to recover missing values for scRNA-seq data
-probability of dropout computed using gamma-normal expectation maximum algorithm
-model demonstrated robust performance, achieving comparable or better results compared to other imputation methods
-identified novel pattern related to daptomycin resistance through bid data analysis of genomes
-novel pattern suggests a new paradigm of daptomycin resistance dissemination
- established a computational workflow of several combined machine learning approaches to identify biomarkers for:
-prostate cancer using metabolomics data,
-chemotherapy-induced cardiotoxicity among breast cancer patients using metabolomics data
-investigating potential new antibiotic resistance genetic markers using known markers in Enterococcus faecium using machine learning approach and population structure of species

DC3: Harmonize multi-organizational and siloed data

Year 1

-downloaded more than 300k bacterial and archaeal genomes from the NCBI and complete set of genomes from Integrated Microbial Genomes and Microbiomes project

-built a structured, organized genome database stored on the high-performance computer at UAMS

-performed quality score analysis of the bacterial genomes in GenBank

-used a standardized pipeline for finding genes for each genome

-developed and published a program, ProdMX, to speed up genome comparison more than a million-fold compared to traditional alignment methods

-used heat maps to visualize the comparisons of more than hundred-thousand E. coli genomes

-created a matrix of distances displayed as a heat map utilizing Mash and in-house Python script

-developed methods and tools using R-BioTools for visualization of pan- and core-genomes

-released python pipeline for pan- and core-genome-based functional profile for metagenomics samples from microbial communities

-tested three commonly used sequenced-based methods for predicting an organism's taxonomy

-utilized bacterial genomes to build a machine learning approach for meta-proteomics analysis

-published a paper giving an overview of multi-omics approaches in the journal "Molecular Omics"

-developed a multi-omics data integration pipeline consisting of DNA methylation, mRNA, protein, phosphopeptides and histone post-translational modifications to understand the regulation of triple negative breast cancer subtypes MDA-MB-231 (BRCA1wt) and HCC1937 (BRCA15382insC) cell lines

Year 2

-developed integrated database for proteomics & genomics, including annotations

-published a R-Bio Tools paper for visualizing genomes

-developed ML models for known toxins

Conclusions and Recommendations

The Data Life Cycle and Curation component is meeting many of its objectives. Faculty have implemented DWM in Python POC in 10th release, version 2.21; automated data cleansing to include data corrections based on record-to-record comparisons with blocks and clusters; developed novel framework for scalable Entity Resolution using NLM for Locality Sensitive Hashing and Machine Learning achieving accuracy over 95% with a nearly linear runtime; developed computational framework integrating multi-layer genomics data to identify transcriptome and pathway dysregulations in autism spectrum disorder; investigated the expression alterations of survival-related genes in various immune cell types when combining breast cancer bulk and single-cell RNA sequencing data; established a computational workflow of several combined machine learning approaches to identify biomarkers for both
prostate cancer using metabolomics data and chemotherapy-induced cardiotoxicity among breast; downloaded more than 300k bacterial and archaeal genomes from the NCBI and complete set of genomes from Integrated Microbial Genomes and Microbiomes project; developed and published a program, ProdMX, to speed up genome comparison more than a million-fold compared to traditional alignment methods; and published a paper giving an overview of multi-omics approaches in the journal, "Molecular Omics" among other accomplishments.

It is not clear that other components in DART or others in the data science community are using any of the algorithms or programs/processes developed by this research group. Recommend that this group do more to advance the use of their algorithms, perhaps by holding a workshop for DART faculty and graduate students and track the use of their algorithms in the broader data science community. The DWM programs may also be of interest to those college students in some of the new data science programs being developed around the state.

The evaluator has some concern that half of the Year 2 milestones are reported in the stoplight tables as "will be completed by end of reporting year". Of most concern are the activities for Goal 2.3: Harmonize multi-organizational and siloed data where over 90% of the activities are indicated to be completed by the end of the reporting year. This may be the result of teaching commitments during the school year and the summer may afford researchers more time to complete these milestones.

3. Social Awareness (SA)

Proposed: *"In this research area, we focus on developing cutting-edge, socially aware data analytics to address social concerns and meet laws and regulations in national-priority applications and better enable big data analytics to promote social good and prevent social harm. In particular, we will address the following critical challenges. How do we: 1. achieve meaningful and rigorous privacy protection when mining private data or collecting sensitive data from individuals? 2. ensure non-discrimination, due process, and understandability in decision-making? 3. achieve safe adoption, and robustness of machine learning and big data analytics techniques, especially in adversarial settings? 4. incorporate social awareness in domain- or*

application-specific projects? Our research goals are to develop novel techniques to provide privacy preservation, fairness, safety,and robustness to a variety of data analytics and learning algorithms including automated data curation, social media and network analysis, and deep learning, and ensure the adoption of the developed techniques meet regulations, laws and user expectations."

Findings:

Table of Selected Outputs of Component

Outputs	Year 1	Year 2	Total
Publications	2	11	13
Joint Publications*	0	0	0
Presentations	2	12	14
Joint Presentations*	0	0	0
Workshops	0	0	0
Proposals # ($)	4 ($3.2M)	6 ($4M)	10 ($7.2M)
Awards # ($)	0	3 ($700k)	3 ($700k)
Joint Proposals* # ($)	3 ($4.6M)	1 ($800k)	4 ($5.4M)
Joint Awards* # ($)	0	2 ($1.45k)	2 (1.45k)

*Collaborative proposals/awards/publication with other DART components

SA1: Privacy-Preserving and Attack Resilient Deep Learning

Year 1

-conducted a survey of existing attacks on deep learning models broadly categorized into evasion, poisoning and model stealing
-researched representative algorithms, using threat models from four aspects: adversarial falsification, adversary's knowledge, adversarial specificity, and attack frequency
-conducted theoretical studies of the potential risks of deep learning models and privacy preserving mechanisms for deep learning
-completed literature review on definition of personal identification information (PII) from different perspective and privacy issues
-proposed frameworks to assess sensitivity of information in different contexts
-developing metric to consider sensitivity level of each PII attribute and combined sensitivity of a given set of leaked PII attributes

Year 2

-designed framework to generate poisoning samples to attack model accuracy/algorithmic fairness of fair machine learning models
-developed 3 online attacking methods: adversarial sampling, adversarial labeling, and adversarial feature modification
  -all effectively and efficiently produced poisoning samples of training data to reduce the test accuracy
  -can flexibly adjust the attack's focus and accurately quantify the impact of each candidate point to accuracy loss and fairness violation, producing effective poisoning samples
  -conducted experiments on two real datasets demonstrating the effectiveness and efficiency of our attacking framework
-studied privacy preserving mechanisms used for deep learning algorithms
  -studied the inequality in utility loss due to differential privacy
  -compared changes in prediction accuracy w.r.t. each group between private and non-private model
  -analyzed cost of privacy w.r.t. each group, explain how group sample size and other factors relate to privacy impact
  -examined the privacy, resilience, utility tradeoff of deep learning models
  -developed threat-and-privacy-aware deep learning models
  -developed a modified DPSGD algorithm called DPSGD-F to achieve and equal costs of differential privacy and good utility
  -conducted experiments on real world datasets
  -results showed the effectiveness of DPSGD-F algorithm on achieving equal costs of differential privacy with satisfactory utility
-developed a novel adversarial adaptive defense (AAD) framework based on adaptive training
  - showing trained models adapt at test time to new adversarial attack
  -framework improved structures of training data into groups and each group represents one attack scenario
  -learns a context vector from features of each batch during training
  -incorporates the learned context vector into both prediction and detection models
  -conducted evaluations with popular adversarial attacks and defense strategies on two real world datasets under different attack settings with results showing AAD achieves high prediction and detection accuracy and significantly outperforms baseline
-developed robust framework under distribution shift that adopts reweighing estimation approach for bias correction and minimax robust estimation approach for achieving robustness on prediction accuracy

SA2: Socially Aware Crowdsourcing

Year 1

-investigated interval-valued labels to enable a worker to specify both type-1 and type-2 uncertainties in his/her label without information loss

-developed algorithms to aggregate labels as an inference with a preferred probability of matching above 50% computationally
-developed strategies to pre-process collected interval-valued labels into two categories, data cleaning and normalization
-extended traditional statistic and probabilistic concepts for point-valued datasets to interval-valued including mean, variance, standard deviation, and probability density function for interval-valued labels.
-investigated two learning algorithms on deriving inferences from interval-valued labels using majority voting and preferred matching probability
-performed computational experiments to test the effectiveness of applying interval valued labels in managing uncertainty in crowdsourcing
-experiments successfully verified theoretical and algorithmic results
-paper accepted by 2021 Annual Conference of the North American Fuzzy Information Processing Society NAFIPS 2021

Year 2
-applied interval-valued labels (IVL) instead of binary-valued
-worker may use a subinterval within (0,1) to annotate an instance even when uncertain
-developed two algorithms (i.e., internal-valued majority voting (IMV) and preferred matching probability (IPMP) to derive inferences from interval-valued labels
-computational experiments evidence the proposed interval-valued scheme enables specification of uncertainties during input time.
-IMV and IPMP algorithms computationally derive an inference with above 50% probability of matching the ground truth
-uncertainty index defined in work quantitatively measures overall uncertainty of collected IVLs
-produce better quality inferences with IVLs than without

SA3: User-centric Data Sharing in Cyberspaces
Year 2
-documented and disseminated research on identification techniques for non-structured data

Year 2
-continued exploration and development of techniques for identifying context aware sensitive information from unstructured data
-researching multimodal deep learning techniques for detecting and removing sensitive information, discriminating and stigmatizing information from unstructured data

SA4: Deep Learning for Preventing Cross-Media Discrimination
Year 1
-conducted theoretical investigation on CNN deep learning models to identify discriminatory objects from social images
-conducted research on adopting long, short-term memory network to model text (captions, tags and discussion of social images)
-conducted research and empirical studies on multimodal hate speech detection
-detecting coded words in hate speech detection
-developed coded hate speech detection framework, CODE, to judge coded words used in the coded meaning

Year 2
-developed a deep learning-based coded hate speech detection framework called CODE
-CODE findings published and presented to research community
-proposal submitted on CODE findings
-conducted empirical analysis on multimodal hate speech detection models
-evaluated performance of Facebook Hateful Meme Challenge baseline models on 3 MMHS150K datasets, image and text inputs
-trained models using different baseline approaches, unimodal training, multimodal training with unimodal pretraining and multimodal pretraining
-evaluated metrics, accuracy and the Area under the ROC Curve (AUROC)
-evaluation shows current multimodal training does not significantly outperform unimodal training

SA5: Marketing Strategy Design with Fairness
Year 1
- performed exploratory study on fairness-aware design decision-making
-trained Logistic regression and CatBoost classifiers on the pre-processed dataset to predict individuals 'income in test data using 10-fold cross-validation approach
-conducted Disparate Impact (DI) analysis, observing gender attribute and ethnicity attribute ducted fairness tested based on the calibration scores using probability score to determine gender attribute discrimination

Year 2
-conducted link prediction in identity network based on social network, intra-layer and inter-layer link information

-can predict number of nodes affected in the entire social network
-conducted comparison with theoretical approaches, independent cascades and liner threshold
-quantified unfairness and analyzed its impact in the context of data-driven engineering design using the Adult Income dataset
-introduced standard definitions and statistical measure of fairness to the engineering design research
-used outcomes from 2 supervised machine learning models, Logistic Regression and CatBoost classifiers
-conducted disparate impact and fair-test analyses to quantify unfairness present in the data and outcomes
-findings published and presented to research community

SA6: Privacy-Preserving Analytics in Health and Genomics
Year 1
-conducted survey of existing frontier work and investigation of mathematical models related to privacy-preserving analysis
-algorithms are built based on models related to computational phenotyping, mathematical optimization and statistics models

Year 2
-documented and disseminated findings of literature research of privacy-preserving data analytics algorithms and software
-initiated investigation on mathematical optimization models
-findings published and presented to research community

SA7: Cryptography-Assisted Secure and Privacy-Preserving Learning
Year 1
-conducted survey of existing work using cryptography for privacy protected in federated learning
-analyzed each work using studied machine learning model/algorithm/method, type of dataset partition, cryptography method and whether differential privacy is provided
-designed a new cryptography-based scheme for differentially privacy federated learning
-established two goals for the new cryptography-based scheme
-reduce communication cost in training process and improve the learning accuracy while providing differential privacy
-experimental results show scheme performs better than existing work in convergence rate and learning accuracy

Year 2
-designed a cryptography-based solution
-developing a privacy-preserving face recognition-based access control system

Conclusions and Recommendations
The Social Awareness component is meeting many of its objectives. This component has researched representative algorithms, using threat models from four aspects: adversarial falsification, adversary's knowledge, adversarial specificity, and attack frequency;
developing metric to consider sensitivity level of each PII attribute and combined sensitivity of a given set of leaked PII attributes; developed a novel adversarial adaptive defense (AAD) framework based on adaptive training; investigated interval-valued labels to enable a worker to specify both type-1 and type-2 uncertainties in his/her label without information loss; developed coded hate speech detection framework, CODE, to judge coded words used in the coded meaning; performed exploratory study on fairness-aware design decision-making; conducted link prediction in identity network based on social network, intra-layer and inter-layer link information;
documented and disseminated findings of literature research of privacy-preserving data analytics algorithms and software; conducted survey of existing work using cryptography for privacy protected in federated learning and designed a cryptography-based solution.

The group may find that as they research aspects of marketing, the link predictions in identity networks may be impacted and additional variables will need to be added to their model. Similarly, the group may want to have other DART components utilize their privacy-preserving analytics in other content areas to test its applicability.

4. Social Media and Networks (SM)

Proposed: *"Our research will address these challenges in close collaboration with Wu and Sheng (SA) by exploring innovative methods and techniques of mining argumentation data in social networks and analyzing its characteristics, such as polarization, opinion diversity, participant influence, opinion community, and opinion prediction; transformative multilayered network analytic method to analyze deviant behaviors in social media networks by modeling multi-source, supra-dyadic relations, and shared affiliations among deviant groups; and innovative algorithms for logistics planning in disaster response using social media platforms. The models and insights generated from the proposed research will enhance our ability to both capitalize on the potential of social media as a force of good while mitigating its use as a weapon."*

Findings*:*

Table of Selected Outputs of Component

Outputs	Year 1	Year 2	Total
Publications	0	21	21
Joint Publications*	0	0	0
Presentations	5	16	21
Joint Presentations*	0	0	0
Workshops	0	0	0
# Proposals # ($)	9 ($30.1M)	2 ($1.2M)	11 ($31.3M)
Awards # ($)	5 ($3M)	2 ($8.7M)	7 ($11.7M)
Joint Proposals* # ($)	0	0	0
Joint Awards* # ($)	0	0	0

*Collaborative proposals/awards/publications with other DART components

SM1: Mining cyber argumentation data for collective opinions and their evolution

Year 1

-determined key featured for platform

-prepared software design guideline document for platform development

-developed and designed new social issue generator for social network data collection processing

-received IRB approval for social issue generators for Fall 2020 Data Collection

-determined baseline individual user and Social Network measure for cyber-discourse platform

- used Intelligent Cyber Argumentation System (ICAS) platform

-collected social network data to identify key variables and mine data to populate statistical compatible datasets (SPSS) from cohort of students in 2018-2020 General Sociology classes

Year 2

-mine thousands of lines of sentences from ICAS platform

-developed novel algorithm

-testing algorithm for evaluation

SM2: Socio-computational models for safer social media

Year 1

-identified social media platforms used in different cyber influence campaigns and different contexts and geographical regions

-identified characteristics and features of social media platforms

-conducted multi-taxonomy characterization of social media data

-identified data sources

-developed and published data collection methodology

-created database schema to accommodate new fields with changes in data sources or characteristics

-identified data acquisition methods to include API access and web scraping

-procured an academic data collection license (Twitter)

-submitted data access proposal and review which was accepted

-developed, tested, and deployed data collection framework with real-time dashboard to monitor progress with alerting capabilities

-identified key characteristics of classification, perceptive quality, and scalability of multimedia data on social platforms

-identified one of the three major learning objectives

-identified damage assessment and verification based on image and video data in disaster response as one of the three key applications

-published work on identifying cyber campaigns and features, and established data acquisition procedures in collaboration with practitioners and policy makers within and outside Arkansas

Year 2

-revised taxonomy to characterize OIE based on social media platforms; studied cyber campaigns and characteristics of platforms and involved information actors

-developed socio-computational model

-identified focal structures, leverages theory of social network analysis/collective action and uses operations research framework

-evaluated with data collected on YouTube conspiracy theory spreaders and Twitter misinformation networks

-datasets correspond to application areas, COVID-19, smart city infrastructure security protests, social movements

-devised indexing methods for image, video and associated meta and text data

-identified key applications of "smart" use of multimedia information sources
-use of information quality aspects as part of future smart data-based applications
-focus on reliability dimension

SM3: Auto-annotation of multimedia data
Year 1
-completed two disaster response routing problem variants using Milburn's existing qualitative interview data
-completed reviews of the CTP and OP academic literature

Year 2
-began devising indexing methods for image, video and associated meta and text data
-key applications of "smart" use of multimedia information sources have been identified with a focus on the reliability dimension

SM4: Informing disaster response with social media
Year 1
-working to identify content types on social platforms to describe transportation infrastructure status after disruptions due to a disaster
-documented the workflow surrounding how to manually transform data from individual content elements posted to social platforms into transportation infrastructure status information

Year 2
-selected Hurricane Harvey for disaster scenario and will be primary focus for initial scenarios
-represented a large-scale and geographically widespread event
-began assembling disaster image dataset from online disaster image repositories
-downloaded tweets and images from Twitter API
-uses content-based techniques to verify or complete the metadata associated with visual data, location, orientation, timestamps and identification of major landmarks
-gathered initial sample data set for Harvey and performed an analysis and evaluation of techniques
-selected 2 routing problems with application to disaster response
-literature review on disaster logistics problems and models
-deployed PostGIS database and hosted by CAST

Conclusions and Recommendations
The Social Media and Networks component is meeting most of its objectives and exceed its targets for presentations and publications. This component has determined key features and prepared the software design document for the cyber social network platform; developed, tested, and deployed data collection framework with real-time dashboard to monitor progress with alerting capabilities; revised taxonomy to characterize OIE based on social media platforms; studied cyber campaigns and characteristics of platforms and involved information actors and selected Hurricane Harvey to represented a large-scale and geographically widespread disaster scenario.

Research into the auto-annotation of multimedia data goal appears to be trailing the progress being made in the other goals/objectives. These data are becoming more prevalent every day and will be a vital part of the newly developed social media platform. The Year 2 objectives of 'obtain and index content types for at least two disaster scenarios' and 'developing GIS system to display real-time road status inputs' have not been met.

The evaluator has some concern that 70% of the Year 2 milestones are reported in the stoplight tables as "will be completed by end of reporting year". Of most concern are the activities for Goal 4.1: Mining cyber argumentation data for collective opinions and their evolution" where 100% of the activities are indicated to be completed by the end of the reporting year. This may be the result of teaching commitments during the school year and the summer may afford researchers more time to complete these milestones.

5. Learning and Prediction (LP)

Proposed: "*Research in this topical area will focus on various techniques in prediction interpretation for large-scale, deep learning using multi-source integrated data sets. In particular, we will focus on applying statistical learning techniques alongside more advanced deep learning techniques to address three major challenges. 1)Violation of fundamental statistics principles. 2)Mode specification and interpretation. 3)Computing in big data environments. We will investigate these challenges surrounding high-dimensional, dynamic and unstructured data sets and explore solutions in the domains of genomics, transaction scenarios in eCommerce, and supply chain logistics."*

Findings:

Table of Selected Outputs of Component

Outputs	Year 1	Year 2	Total
Publications	2	21	23
Joint Publications*	0	1	1
Presentations	4	20	24
Joint Presentations*	0	1	1
Workshops	2	0	2
# Proposals # ($)	3 ($2.8M)	7 ($3.3M)	10 ($6.1M)
Awards # ($)	1 ($1.2M)	1 ($50k)	2 ($1.2M)
Joint Proposals* # ($)	3 ($2.3M)	2 ($300k)	5 ($2.6M)
Joint Awards* # ($)	1 ($800k)	3 ($1.7M)	4 ($2.5M)

*Collaborative proposals/awards/publications with other DART components

LP1: Statistical Learning – Random Forests for Recurrent Event Analytics

Year 2

-transitioned MTPP and LTSM to a convolutional neural network (CNN) approach
- -implemented completely in Keras
- -established pipeline to allow maintenance based tabular data set first tested on model
- -published work on CNN for image-based data
- -work serving as foundation for shift in methodology from MTTP/LSTM integration

LP2: Statistical Learning – Marked Temporal Point Process Enhancements via Long Short-Term Memory Networks

Year 1

-replicated one of the existing methods in the literature for intelligent food-borne disease investigation based on event data
-explored the RF-SRC method and use of NHPP to model the recurrent event
-literature review identified gaps in existing methods
-curated dataset of sensor data, system attributes, and failure/repair data of 8332 oil/gas wells installed 2007-2021

Year 2

-collected civil infrastructure datasets publicly available
-replaced curated healthcare loT datasets with curated maintenance-based tabular data-curated and available
-storing datasets on shared repository via the cyberinfrastructure team

LP3: Deep Learning: Novel Approaches

Year 1

-developed library of classifiers on natural images from simple linear classifiers
-SVMS and linear autoencoders to complex deep learning approaches
-AlexNet, ResNet, and GoogleNet, and standard machine learning classifiers
-investigated efficacy of group structure on generalized neural network (GNN) architecture with smallest finite simple nonabelian group A5 action of random and clustered synthetic small size data
-applied techniques to the color channel data of three-dimensional fundamental topological structures
-resulting data studied for any significance by exploratory data analysis and statistical inferential methods
-development of generalized model of reward function in DRL addressing issues with sparse and dense feedback
-developed multiple low-cost deep learning methods
-Teacher-Student Distillation Deep Learning, Distilled ShuffleNet, Self-Knowledge Distillation Algorithms

Year 2

-investigating the injectivity issues in persistent homology
- -2 distinct shapes have identical topological representation making them unusable as discriminating features in neural network
- -studied a use-case with “Montezuma’s Revenge, Atari game for a DRL agent
- -using support of human intuitions (in from of heuristics), agent was able to perform exploration very quickly
- -submitted paper on work to a conference

LP4: Deep Learning: Efficiency and Specification

Year 1

-analyzed current issues of distillation methods for computer vision dataset
- -face recognition, action recognition, and medical imaging
- -processing a private insurance claims dataset for predicting opioid overdose

Year 2
-introduced new Deep learning algorithms perform well in low-cost platforms with high accuracy
-introduced novel deep neural networks to productively deploy Al based object recognition on mobile devices
-provided algorithmic analysis, competitive against large-scale deep networks, significantly reducing computational time and memory consumption
-evaluated on various applications in natural images and medical images
-lightweight model can achieve high performance on large-scale challenging natural and medical image benchmarking datasets
-developed new methods in predictive modeling framework for incorporating time-dependent features with a new method for deriving variable importance
-2 manuscripts under revision, expected to be resubmitted by end of April, 2022
-delivered oral and poster presentation at 2021 INFORMS Annual Meeting
-constructing optimization model to identify optimal size of time windows for prediction
-working on representation learning for our deep learning models

LP5: Harnessing Transaction Data through Feature Engineering
Year 1
-using MIMIC data
- -identified important features related to patient outcomes under ventilators
- -created descriptive statistics to be included in the feature set, and
- -presented preliminary result at 2020 INFORMS Annual Meeting

Year 2
-developed autoencoder method
- -invention in unsupervised and self-supervised deep learning methods contributed to tackle problem of large-scale dataset management and labeling in Big Data management
- -work provides an effective solution to solve Small Data challenges in Big Data era
- -introduced new combination approach between deterministic and probabilistic deep neural network
- -allows introducing a new powerful mechanism to reason knowledge representation in big data datasets
- -created new feature based on Tor network protocol
- -collects network traffic dataset over Tor network
- -classifies websites in real time
- -results in foundation technology for detecting websites that disseminate illegal contents
- -created low-dimensional feature using histogram entropy
- -used on malware datasets, Windows, Android, and loT malware
- -conducted study to acquire comparable results with small datasets to reduce cost of training machine learning models on huge datasets
- -performed a comparative analysis between convolutional neural network (CNN), residual neural network (RNN) and Vision Transformer (VT) using Ductal Carcinoma (breast cancer tissue images) dataset
- -assess suitability for adoption
- -VT model outperformed the CNN and RNN on different tasks achieving up to 93% accuracy
- -accepted paper on work

-introduced a pre-input layer to binary-decision-fusion neural network
-train layer for out-of-distribution cases adjusting feature values, without altering original training of network

Conclusions and Recommendations
The Learning and Prediction research component is meeting many of its objectives. This component has transitioned MTPP and LTSM to a convolutional neural network (CNN) approach; curated dataset of sensor data, system attributes, and failure/repair data of over 8,000 oil/gas wells; investigated efficacy of group structure on generalized neural network (GNN) architecture with smallest finite simple nonabelian group A5 action of random and clustered synthetic small size data; introduced new Deep learning algorithms that perform well in low-cost platforms with high accuracy which significantly reduce computational time and memory consumption;
developed autoencoder method invention in unsupervised and self-supervised deep learning methods contributed to tackle problem of large-scale dataset management and labeling in Big Data management.

It is unclear to what extent the learning paradigms and algorithms are being used by other DART researchers or the data science community in general or whether the MTTP enhancements have been evaluated and assessed on real-world discrete data sets or that the MTTP/LSTM approach is scalable for implementation on different data sets.

6. Education Component

Proposed: "*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. 2. Providing resources and training for educators including $5,000 Seed Grants for project-related Education & Broadening Participation; Career Development Workshops for project participants and educators; and K12 teacher professional development on data science topics. 3. Providing educational opportunities inside and outside the classroom for students. Undergraduate and graduate research assistantships in DART labs will be funded along with intensive data science and computing summer camps for undergraduates and research-based capstone projects and internships with industry partners. 4. Ensuring broad participation to impact the pipeline of data science skilled workers through Summer Undergrad Research Experiences in DART labs for underserved students, scholarships for underserved students to the Arkansas Summer Research Institute (ASRI); and by connecting students to opportunities through the Arkansas Center for Data Sciences (ACDS)."*

Findings:

Table of Selected Outputs of Component

Outputs	Year 1	Year 2	Total
Publications	0	0	0
Joint Publications*	0	0	0
Presentations	0	0	0
Joint Presentations*	0	0	0
Workshop	0	0	0
Joint Workshops*	0	2	2
# Proposals # ($)	0	0	0
Awards # ($)	0	0	0
Joint Proposals* # ($)	0	1 ($800k)	1 ($800k)
Joint Awards* # ($)	0	0	0

*Collaborative proposals/awards/publications with other DART components

Data Science Ecosystem: Model Programs, Degrees, Pedagogy, and Curriculum

Year 1

-completed initial Middle School Coding Block Workshop with plan finalized and disseminated to stakeholders
  -5-year plan outlined for stakeholders during workshop November 2020
  -hosted approximately 50 attendees from 40 campuses and organizations in Arkansas
  -published curriculum information and plan on OneDrive site for participants at November workshop
-distributed institutional needs assessment to CS/DS faculty and IT support at all Arkansas campuses
  -approximately one-half of campuses responding
-UA Data Science program:
  -accepted its inaugural class with 45 students
  -approximately one-third are not calculus-ready,
  -one-third are "standard 8-semester plan", and
  -one third are transfer students from other majors or from other academic institutions
  -developed a suggested 6-semester plan for students who change their majors to Data Science and can be adapted for the second two years of 2+2 programs
  -representative on ABET's accreditation workgroup and is tracking requirements for readiness- program broadly distributed
-UCA developed a standalone BS in Data Science with concentrations in computer science, statistics, and business
  -program received academic approvals on campus and by UCA Board
  -submitted for review and approval by Arkansas Higher Education Coordinating Board
  -program specifics shared with cohorts at statewide meetings
-UAPB discussion on how to begin a Data Science program
-A-State developed BS in data science
  -approved by Arkansas Higher Education Coordinating Board
  -accepting students in Fall 2021
-Philander Smith College received approval for three courses:
  -Intro to Data Science using python, Machine learning and Ethics in data science
  -established partnership with IBM with 3 faculty completing Data Science training
-Shorter College finalized curriculum for two courses
  -faculty are waiting to meet with Dean and Associate Dean for approval and implementation plan
  -partnered with IBM with 4 faculty completing the data science and artificial intelligence badge programs
-completed 3 Data Science for Arkansas workshops

-attendance by post-secondary academic Arkansas institutions, Division of Higher Education (ADHE), Economic Development Commission (AEDC), and Center for Data Science (ACDS)
- shared via a OneDrive to participants, course materials, curriculum, adaptions and implemented curriculum
-identified "Opt-In" partners from the Research team

Conclusions and Recommendations
The Education research theme has completed initial Middle School Coding Block Workshop with plan finalized and disseminated to stakeholders; 5-year plan outlined for stakeholders during workshop November 2020; UA Data Science program has been established and other colleges/universities are moving forward with developing data science programs for their respective campuses. Three Data Science for Arkansas workshops have been held during the first two years involving post-secondary academic Arkansas institutions, ADHE and ACDS.

The evaluator has some concern that 88% of the Year 2 milestones are reported in the stoplight tables as "will be completed by end of reporting year". This may be the result of teaching commitments during the school year and the summer may afford researchers more time to complete these milestones.

7. Workforce Development Component
Proposed: *"Provide K20 teacher and faculty opportunities for professional development spanning multiple disciplines*

Findings:

Table of Selected Outputs of Component

Outputs	**Year 1**	**Year 2**	**Total**
Publications	0	0	0
Joint Publications*	0	0	0
Presentations	0	0	0
Joint Presentations*	0	2	2
Workshops	9	7	16
Proposals # ($)	1 ($1.8M)	0	1 ($1.8M)
Awards # ($)	0	0	0
Joint Proposals* # ($)	2 ($3.2M)	2 ($800k)	4 ($4M)
Joint Awards* # ($)	0	1 ($50k)	1 ($50k)

*Collaborative proposals/awards/publications with other DART components

Provide K20 teachers and faculty opportunities for professional development spanning multiple disciplines
Year 1
-launched March 2021 application for teachers EAST Initiative Annual Conference
-identified campuses that will submit faculty for Cohort 1 training:
-Shorter College, PSC, UAPB, North Arkansas College, and Arkansas Tech University
-identified 5+ capstone partners from the research teams
-licensed new platform, UpSquad to serve as online community with teleconferencing and telework functionality

Year 2
-28 teachers have participated in a 3D printing workshop, 21 attended the Pi-Top workshop and 82 attended a leadership workshop
-offered 5 virtual Software Carpentry workshops attended by 195 participants (not all from DART)
-hosted a free workshop for 70 participants by NVIDIA Deep Learning Institute
-provided a Communicating Science to Legislators and Distilling Your Message workshops
-20 people attended a workshop on individual development plans
-hosted a DART pedagogy workshop

Provide educational training opportunities inside and outside the classroom for students
Year 1
-provided training for graduate students and undergraduates throughout the year
-funded 10 underrepresented undergraduates for lab work during the summer

Year 2
-provided training and mentoring for 91 graduate students, 3 postdocs and 31 undergraduates
-hired 10 underrepresented undergraduate students for summer research
-42 undergraduates completed the ASRI virtually, which also involved 53 presenters and panelists from DART institutions and collaborators

Conclusions and Recommendations
The Workforce Development component has participated in two EAST conferences and hosted two support/training webinars, as well as 2 statewide workshops on data science topics. They also have hosted a virtual ASRI which was completed by 42 undergraduates and involved a significant number of DART faculty and collaborators.

It is unclear whether the objectives of 10 awards of $5k each per year for faculty training have been utilized or that 'internship opportunities for students at relevant companies have been identified'.

The evaluator has some concern that two-thirds (67%) of the Year 2 milestones are reported in the stoplight tables as "will be completed by end of reporting year". In addition, some milestones from Year 1 have been delayed, specifically related to internships for students and developing capstone courses for the data science curriculum. This may be the result of teaching commitments during the school year and the summer may afford researchers more time to complete these milestones.

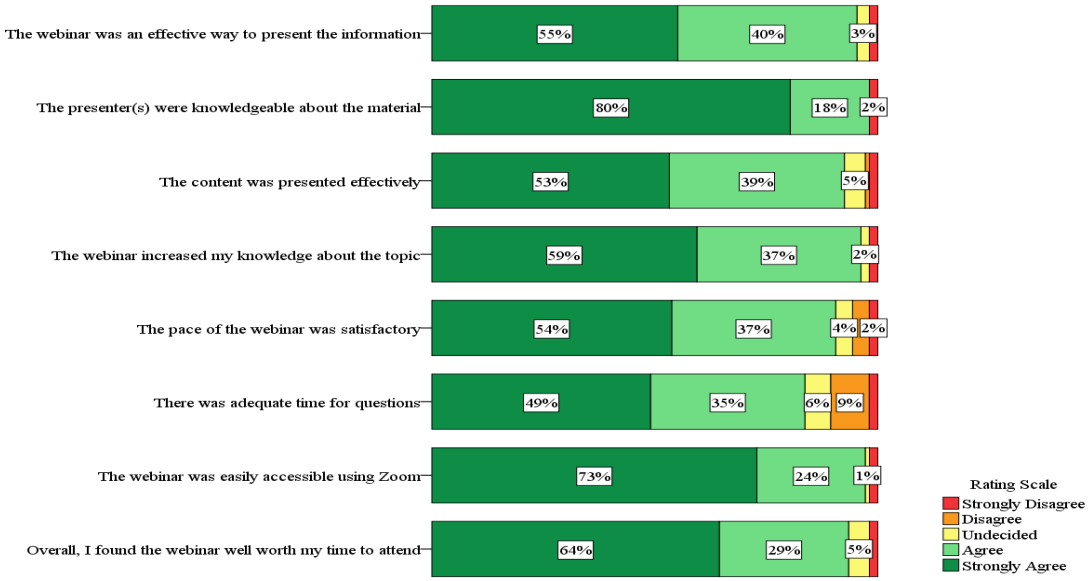
8. Communications & Dissemination Component
Proposed: "*Maintain interproject communication to accomplish milestones and relay updates."*

Findings
Year 1
-established Slack group for DART faculty who utilizes it for email for daily communication
-established DART GitLab with efforts being made to ensure cross-campus participation
-exploring an UpSquad license, a new telework and network community to be used for 2021 ASRI
-completed 11 monthly team meetings per 6 research teams
-hosted 5 DART Monthly Seminars to date with 2 additional webinars planned by end of Year 1
    -invited DART faculty, staff, and students to the webinar series
    -recorded DART webinars available on the DART website
-collected social media accounts and blogs of DART faculty for cross-posting
-established three listservs: one each for the DART SSC; DART Project Faculty and Staff; and DART students
-implemented ER Core site in August 2020 with participants onboard through March 2021
    -100% DART participants, paid and unpaid except advisory board members, provided accounts in ER Core
    - central office participating in ER Core Consortium and hired developers to make continuous improvements/upgrades to platform
    -70% of users attended at least one of the 5 webinar trainings from September to January
-published 9 peer-reviewed articles and juried conference papers included in the Scientific publication list and reported in NSF PAR
-2 statewide workshops for cohorts and waves.
-formed Science Journalism Committee

Year 2
-expanded communications with DART participants with a weekly digest email informing them of upcoming events, DART related funding opportunities and recent project accomplishments
-hosted weekly virtual office hours where a project leadership team member is available to answer participant questions
-held monthly seminars and monthly team meetings
-held a virtual retreat involving over 125 people, including 80 DART participants
-will hold first face-to-face All hands meeting before the end of project year

Monthly Webinars: Excerpt from evaluation Report
A post-seminar feedback survey was administered after each webinar. Figure 17 summarizes these results for all the webinars.

Figure 17
Please rate the webinar, including its organization, content and potential impact.
(Number responding=107)



Survey respondents 'Strongly Agreed' or 'Agreed' that the webinars were 'an effective way to present the information' (95%); 'presenter(s) were knowledgeable' (98%); 'increased my knowledge' (92%) and 'overall, well worth my time to attend' (93%).

Annual Retreat: Excerpt from evaluation report
DART held a virtual retreat in February 2022 attended by over 125 people attended for part or all of the two-day meeting; including faculty (51%); graduate students (17%); administrators (6%); undergraduate students (6%); industry professionals (6%); post-docs (2%); partners (2%) and not reports/others (10%). Over one-third (36%) of the attendees were from the University of Arkansas-Fayetteville; 12% from the University of Arkansas-Little Rock; 8% from the University of Arkansas Medical School; 6% from Arkansas State; 6% from industry; 5% from University of Central Arkansas; 5% from Southern Arkansas University; 5% from AEDC and the remaining from a variety of institutions.

A post meeting survey was taken by 39 of the 115 participants for a return rate of approximately 34%. Two-thirds or more of respondents rated the following 'Very Good' or 'Excellent': 'Retreat Agenda' (82%); 'Round Robin Breakouts' (81%) and 'Research Team Breakout Session using Mural' (75%); 'Pre-retreat Communication' (72%); 'Promoting Inter-Institutional Collaboration' (72%); 'Jargon Audit Discussion' (71%) and 'Promoting Inter-Disciplinary Collaboration (70%).
The following are selected responses from the open response survey questions.
What did you find MOST useful about the retreat?

- The breakouts where I got to meet members of other teams.
- Learn better about the program and build potential research collaboration.
- The time for research teams to meet and explore their individual contribution to the project. Very open discussions and contributions by junior faculty to discussions.
- Get updated about the progress of the project.
- The break-out sessions wherein I was able to make new contacts for potential collaborations!
- The opportunity to learn more about the cross-section of DART project in a small group setting (round-robin). Also the mural with the questions to think through our contributions and needs (big picture -global perspective). I want to thank the DART Team and leadership --Jennifer, Hanna, and Jack -- for how you put this conference together and the thoughtfulness in its implementation. Also, many thanks for all the background, behind-the-scenes work!! You are a DREAM TEAM!

Based on the feedback surveys, the retreat was a success at involving a large percentage of the DART participants who remained engaged during most of the two days. Most participants were appreciative of the opportunity to have the time in the breakouts to "meet members of other teams" and "learn better about the program and build potential research collaboration".

Participants also expressed an appreciation for "the research teams to meet and explore individual contributions to the project. Very open discussions and contribution by junior faculty to discussions".

Educate the public about DART accomplishments.
Year 1
-published basic project website presence for the project at https://dart.cast.uark.edu/
-details about each research theme and the faculty/graduate students,
-relevant documents, Strategic Plan, Arkansas S&T Plan and Annual Reports, and identified improvements
-published 2 AEDC blogs with 2 additional blogs end of year 1 at https://www.arkansasedc.com/news-events/arkansas-inc-blog
-hired communications intern at central office to assist in content generation and publicity of DART
-increased social media following on Facebook and Twitter by 7%
-formed Campus Communications Committee
-made initial contact with communications offices at most of the participating campuses
-host first meeting with Campus Communications Committee summer 2021

Year 2
-project website became live
-meet 5 year goals for social media followers on Twitter and Facebook
-met with campus communications committees

Conclusions and Recommendations
The Communication and Dissemination component established Slack group for DART faculty who utilizes it for email for daily communication; completed monthly team meetings for each research team; hosted 12 monthly webinars during the first two years; published a website; AEDC blogs; increased social media presence on Facebook and Twitter; formed Campus Communications Committee; held an annual retreat and annual all hands meetings.

Evaluations of the retreat revealed that the DART participants remained engaged during the two days. Most participants were appreciative of the opportunity to have the time in the breakouts to "meet members of other teams" and "learn better about the program and build potential research collaboration". Participants also expressed an appreciation for "the research teams to meet and explore individual contributions to the project. The webinar evaluations revealed that survey respondents 'Strongly Agreed' or 'Agreed' that the webinars were 'an effective way to present the information' (95%); 'presenter(s) were knowledgeable' (98%); 'increased my knowledge' (92%) and 'overall, well worth my time to attend' (93%).

9. Broadening Participation Component

Proposed: "*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"*

Findings:

Mentorship of students and early career faculty
Year 1
-filled 15 undergraduate (UG) student research assistantship positions
-filled 40 graduate (GA) student research assistantship positions

Year 2
-supported 40 graduate and 15 undergraduate students

Research Seed Grants Program
Year 1
-first round of seed grants offered in October 2020 with two awards of $5,000 issued to:
-Arkansas Regional Innovation Hub for a virtual field trip project entitled "STEM Saturdays", and
-Henderson State University STEM Center for a project entitled, "ConneCTED: Developing Communities of Practice with an Emphasis on Computational Thinking and Engineering Design"

Year 2
-second round of seed grants awarded over $752k to 16 projects
-ATU: "Development of Interdisciplinary Research Collaborative to Provide Datasets in Support of Education Research in Data Science" (ED)
-UARK: "Toward Fair and reliable consumer acceptability prediction rom from appearance" (LP)
-UAMS: "Geospatial Data Science in Public health: Interinstitutional educational collaboration to enhance data science curriculum in Arkansas" (ED)
-UARK: "Interpretable Multimodal Fusion Networks for Fault Detection and Diagnostics of Two-Phase Cooling Under Transient Heat Loads" (LP)
-UAMS: "Piloting Bid Data Science in Arkansas Middle School Classrooms" (ED)
-UARK: "Machine Learning-based emulation and prediction in ensembles in disordered photocatalytic composites" (LP)
-A-State: "AgAdapt: An evolutionarily-informed algorithm for genomic prediction of crop performance in novel environments" (LP)
-UALR: "Machine Learning for Predicting Refugee Counts" (LP)
-UCA: "Crying out Data Science in the center of Arkansas: Invitation to High School students to the World of Data Science" (ED)
-PSC: "Generating big Radiogenomic Data of Cancer Using Deepfake Learning Approach"

Summer Undergraduate Research Experienced (SURE) Program
Year 1
-supported 12 underrepresented undergraduate and 1 high school students in DART labs

Year 2
-intend to fund an additional 15 undergraduate students in the summer of 2022

Broadening Participation Seed Mini-Grants
Year 1
-none reported

Year 2
-Awarded 7 mini-grants a total of $15.5k
-UCA: "Integrating Data Science to Rethink Mathematics and Science"
-Harding: "Robots, Rocketry and Programming summer Camp"
-Ozarks Unlimited: "K-5 STEM Integration: Designing Authentic & Meaningful STEM"
-Northwest Arkansas ESC: "K-5 STEM Integration: Designing Authentic & Meaningful STEM"
-Northcentral Arkansas ESC: "K-5 STEM Integration: Designing Authentic & Meaningful STEM"
-Dawson ESC: "K-5 STEM Integration: Designing Authentic & Meaningful STEM"
-Southside School District: "Southside Summer STEM Institute"

Arkansas Summer Research Institute (ASRI)
Year 1
-none reported

Year 2
-40 of the 100 student applicants recruited for attendance at Arkansas Summer Research Institute, ASRI

ASRI: Excerpt from evaluation report
The Year 2 ASRI was attended by 42 undergraduates in June 2021. The ethnic distribution was over one-third (37%) Asian; one-fifth (20%) White; 17% Black; 9% Multi-ethnic and 6% Other. Over three-fourths (81%) of participants reported attending an institution in Arkansas. One-fourth (26%) of the students are attending an institution in the University of Arkansas system; UA: Fayetteville (17%); UA: Little Rock (7%); UA: Pine Bluff (2%) and Philipps Community College (2%). In addition to the undergraduate student participants, the ARSI involved 53 presenters and panelists, which included graduate students, faculty, staff and entrepreneurs. About one-fourth (23%) were from the University of Arkansas at Fayetteville.

A daily feedback survey was completed by students at the end of each day. More than three-fourths of the students rated their ASRI experiences prior to this first day as 'Excellent' or 'Very Good' with 'Registration' (86%) and 'Orientation' (80%).

The percentage of students rating sessions during the first three days of week one as 'Excellent' or 'Very Good' ranged from less than two-thirds (63%) for the session 'Setting up Your Machine Learning Environment' on Day 2 to more than three-

fourths (87%) for the sessions 'Giving a Compelling presentation' and 'Dr. Krakowiak's Research Story' both on Day 3. The percentage of students rating sessions during the last two days of week one as 'Excellent' or 'Very Good' ranged from two-thirds (69%) for the session 'Spatial Data and Mapmaking in R-ADV' on Day 4 to more than ninety percent for the sessions 'Dr. Ussery's Research Story' (97%) on Day 5; 'Panel: Entrepreneurship in Research' (94%) on Day 4 and 'Getting Started with R' (93%).

The percentage of students rating sessions during the first three days of week two as 'Excellent' or 'Very Good' ranged from less than three-fourths (72%) for the sessions 'R Bio Tools Part 1', 'Getting Started with Python' and 'Capture the Flag on the Cyber Range' to more than ninety percent for the sessions 'Individual Consultations' (93%) and 'Panel: STEM Careers' (91%). The percentage of students rating sessions during the last two days of the ASRI as 'Excellent' or 'Very Good' ranged from three-fourths (75%) for the sessions 'Getting Started with Python' to more than ninety percent for the sessions 'Presentations' (93%) and 'Classification in Python -Adv' (91%). The percentage of students rating as 'Excellent' their overall daily ASRI experience ranged from half (55%) for Week 2: Tuesday to more than three-fourths (83%) for 'Week 2: Friday. Overall daily ratings with a combined 'Excellent' and 'Very Good' all exceeded 90% for the two weeks.

The following are selected responses from the final day of the Institute.
"What is the #1 BEST thing that has come out of the ASRI for you? This could be an internship opportunity, an interview to join a professor's lab, a new skill you plan to use in the future, an interest in a new research subject, etc."

- This research experience has given me more insight on the world of research and sparked an interest in computer/data science for me!
- Connections and getting to know more about different topics I might enjoy.
- I have met so many people, including professor and student. I especially enjoy hearing dramatic story from Dr. K.
- Definitely confidence. I've always been a singer, and for a long time, that's all that I felt I was good at. Everyone with ASRI has been so encouraging and I am looking forward to next year!
- I plan to advance my learning in programs and use them in the future. (python and Rstudio I had never heard of them before this experience). I never thought I would consider being a doctor but I'm really thinking about it now. I love macro research on human behavior and development. I am grateful for the speakers and their comments and encouragement to advance.

As evidenced by these findings, the leaders of the Arkansas Summer Research Institute provided a two-week training program for undergraduates that was engaging, educational, inspirational, comprehensive, rewarding and fun for the participants. Faced with the challenge of Covid-19 the leaders did a lot of planning and hard work to pivot from a face-to-face Institute to one that was totally virtual. They did not try to replicate the face-to-face training but instead thoughtfully designed a wholly new program that maximized the strength of a variety of virtual tools available to them and the students.

Conclusions and Recommendations

The Broadening Participation component awarded 2 broadening participation mini-grants of $5k each in Year 1 and 7 mini-grants a total of $15.5k in Year 2; hosted the ASRI for 42 undergraduates in June 2021-26% who are URM; mentored 15 undergraduates and 40 graduate students each academic year and supported 10 URM students each year in the SURE program.

The evaluation of the ASRI found that the two-week training program for undergraduates that was engaging, educational, inspirational, comprehensive, rewarding and fun for the participants. Faced with the challenge of Covid-19 the leaders did a lot of planning and hard work to pivot from a face-to-face Institute to one that was totally virtual. They did not try to replicate the face-to-face training but instead thoughtfully designed a wholly new program that maximized the strength of a variety of virtual tools available to them and the students. The BP group has implemented mentoring protocols and instruments that should help faculty better support students, especially those who are URM, in achieving their academic and career goals.

This group needs the support of the research themes to achieve its diversity goals of: Faculty (45% female, 10% URM); Graduate Students (50% female, 20% URM); Undergraduate Students (50% female, 40% URM) and Advisory Boards (50% female, 20% URM). According to the self-reported participant data the DART participation data show: Faculty (30% female; 3% URM); Graduate students (31% female; 13% URM) and Undergraduate students (48% female; 45% URM).

While the ASRI and SURE programs, along with the new mentoring programs, will help the project achieve its diversity goals for undergraduate and graduate students. However, it is unclear how DART will achieve its faculty diversity goals. The project may want to include the benefit of broadening participation in research teams as a focus in its mentor training. This could have the benefit of expanding the impact of mentoring for both students and faculty recruitment.

Strategic Plan

A virtual strategic planning meeting was held in August 2020. The final strategic plan was submitted to NSF in December 2020. The plan provides metrics with baselines and 5-year targets to determine whether the project is on track. The strategic plan metrics for the entire project can be found in the Appendix of this report. A summary of the strategic plan metrics is presented in Figure 23.

Figure 23
Strategic Plan Metric Summary Table

Strategic Priority Area	Metric Status					
	#	Met/ Exceeded Target	On Track	Delayed	Not Met	Not Available*
Cyberinfrastructure	10	7 (70%)	1 (10%)	2 (20%)	0	0
Research	13	3 (23%)	10 (77%)	0	0	0
Education & Workforce Development	17	0	17 (100%)	0	0	9
Communication	11	2 (18%)	9 (82%)	0	0	2
Broadening Participation	13	1 (8%)	12 (92%)	0	0	3
Overall	64	13 (20%)	49 (77%)	2 (3%)	0	14

* Status not available due to activity having not occurred.

Overall, DART 'Met or Exceeded Targets' in 13 (20%) with 77% of metric ‘On Track’ of the strategic plan metrics. A limited number of the metrics were 'not available' for assessment because of issues with timing or delays caused by Covid-19.

The Cyberinfrastructure Component ’Met/Exceeded Target’ in 70% of its objectives and 10% ‘On Track’. The delay caused by UAF change in computer security policies resulted in 20% of metrics deemed ‘Limited Progress’.

The Research Component ‘Met/Exceeded’ one fifth (23%) of their metrics and are ‘On Track’ with three-fourths (77%) of others. Publications, presentations and algorithm development are ahead of schedule.

The Education and Workforce Development Components are ‘On Track’ with 100% of their metrics that could be measured at this time. There was the largest number of metrics not available to measure because of delays caused by Covid and timing in the project.

The Communication Component met two out of eleven (18%) of its objectives, with 9 (82%) ‘On Track’. The project website has been developed and the ER Core reporting system has been implemented.

The Broadening Participation component is ‘On Track’ with 92% of its objectives and has ‘Met/Exceeded’ its objective of developing/modifying templates to implement a mentoring program.

Summary

DART has involved 216 direct participants in the project so far. Over half (59%) of the project participants are students with over one-third (40%) 'Graduate students' and one fifth (19%) 'Undergraduates'. Over one-third (35%) are 'Faculty participants' with Post Docs representing 1% of the total number of participants. The remaining participants were 'Support staff' (4%).

DART involved 216 in program Year 2, an increase of 79 (58%) from Year 1. The largest increase in participants in Year 2 were in the number of graduate students from 47 to 87, an increase of 85% and in the number of undergraduates from 25 to 41 an increase of 64% and faculty from 56 to 75 an increase of 34% and across all program years. Over the two program years DART has involved 218 unique individuals: one-third (34%) faculty' over two-thirds (41%) graduate students and one-fifth (19%) undergraduates.

The overall ethnic diversity among participants was 15%, ranging from 45% among the undergraduates to 13% within the graduate student participants and 3% among faculty. Female diversity was much higher with 35% over all participants to 48% among undergraduates; 31% among graduate students and 30% among faculty.

The institutions with the largest number of participants are at the University of Arkansas at Little Rock (31%) and the University of Arkansas-Fayetteville (29%). Participants from the University of Arkansas for Medical Sciences represent 7%; Southern Arkansas University 6%; Arkansas Tech University 5% and University of Central Arkansas 5%. The project has involved participants from many higher education institutions, including many smaller institutions, such as Arkansas Tech University (5%); North Arkansas College (1%); Philander Smith College (2%); Shorter College (2%) and University of Arkansas at Pine Bluff (2%).

Project participants reported 17 external collaborators during the first two years of the project. More than one-third (41%) of the collaborators are at academic research institutions, while about one-third (36%) are from primarily undergraduate institutions (18%) and industry (18%). Almost two-thirds (65%) of the external collaborators reported are in Arkansas; while one-fourth (29%) are collaborators located outside Arkansas but within in the US.

Over $1.2 million have been expended or ordered during the first two years to purchase computer infrastructure at three institutions. In Year 1, project funds were used to purchase $650k of infrastructure at the University of Arkansas at Fayetteville. The equipment purchased included DELL Fiber Splitter cables, PowerEdge XE8545, Power Edge R7525, Server, NVIDIA Ampere A100 649,607.18 and a 40-port Mellanox Quantum QM8790. This equipment is part of the DART CI Plan that increases additional hardware needed to move pinnacle out from behind the firewall. Also, in Year 1 data storage servers allowing data sharing among DART researchers were purchased for $24k at the University of Arkansas at Little Rock. In Year 2, $496k was planned to be used to upgrade the research backbone at University of Arkansas Medical Sciences to collaborate with ARE-ON and extend service to the University of Arkansas at Fayetteville.

Sixty-nine proposals requesting over $75 million have been submitted by DART participants. As of March 1, 2022, 32 proposals have been funded for a total of $20.7 million, while twenty-two are still pending. The funding agencies where most of the proposals have been submitted are: NSF (42%) and NIH (14%), while the most award dollars have come from US Department of Defense (37%); NSF (24%) and US Office of Naval Research (19%).

NSF funding in the state has increased from $13 million in 2001 to $30 million in 2021, in 2021 adjusted dollars. While the years 2010 to 2013 showed marked increases in award dollars to the state, this was the result of the American Recovery and Reinvestment Act. A similar increase in Federal funding occurred in

2020 and 2021 with the Covid-19 Stimulus funding. However, Arkansas is on track to maintain a high level of award dollars coming into the state.

Since 2001, Arkansas has received over $430 million dollars in NSF funding. Most (43%) of this funding has been in the form of research awards totally $184.7 million; co-funded awards (both for research and EHR) have amounted to $87.1 million (20%); EPSCoR awards (Tracks1, 2 and 3) have amounted to $83.1 million (19%) and $83.1 million (19%) has been awarded for EHR proposals.

NSF funding for Research, excluding EPSCoR co-funding and Track1, has increased from $4 million in FY 2001 to over $11 million in FY 2020, an increase of 175%. In addition, Education and Human Resource funding from NSF has more than doubled from $2 million in FY 2001 to $5 million in FY 2020.

DART participants reported making 82 presentations, posters, and invited talks during the first two program years. Half (50%) of the presentations were presentation/talks, while more than one-third (37%) were as invited speakers and about one-tenth (12%) were posters.

Researchers reported 89 publications in the first two years of the project. One-third (34%) were reported as receiving primary support from DART. Some of the journals in which DART researchers have published included: BMC Bioinformatics; Infection and Immunity; Metabolites; Microbial Genomics and International Journal of Advanced Computer Science and Applications to name a few.

Overall, an estimated 1,080 people have been involved in one or more external engagement activities/programs supported by DART during the project. About half (53%) were K-12 students reached directly. Almost hlef (46%) of the students reached directly though DART outreach are female and 44% are a member of an underrepresented minority in STEM.

The Coordinated Data Science Infrastructure component is meeting most of its objectives.

The team has established a CI working group, issued purchase orders for: 20 nodes dual AMD 7543, 1024GB, NVMe local drive, single PCI 40GB A100GP; 4 nodes dual AMD 7543, 1024 GB, NVMe local drive, four SXM 40 GB A100 GPU, 100 Gb Inifiniband connection and 10Gb Ethernet connection, installed Git on Pinnacle and Grace, developed System Security Plans to host HIPAA and Controlled Unclassified information at UAF. They have begun work on the visualization for complex data in diverse data-analytics application domains by conducting a systematic literature review on advanced visualization and immersive analytics which is ready for publication on the DART website. They have already published results of a user evaluation on different design choices of virtual field trips in the Journal of Educational Computing Research.

The team noted that the implementation of GitLab, federated identity services for ARP and Globus had been delayed because of UAF network and policy concerns regarding IT security. The enterprise version of GitLab is being replaced with a commercial cloud hosted GitLab repository which will be available to DART researchers. The development of a federated identify service for ARP is being developed by the newly NSF funded SHARP CCI award and the need for a commercial version of Globus may be replaced by the no-cost version. Workshops on using an interactive shell to access Pinnacle are helping to increase the use of these computer resources by faculty and graduate students. A one-on-one connection between IT professionals and researchers seems the best approach for increasing HPC usage. The Year 2 objectives to create containerized Haddoop-based testbed for DC and hold an advanced visualization workshop are anticipated to be accomplished by the end of the year.

The evaluator has some concern that 70% of the Year 2 milestones are reported in the stoplight tables as "will be completed by end of reporting year". This may be the result of teaching commitments during the school year and the summer may afford researchers more time to complete these milestones.

The Data Life Cycle and Curation component is meeting many of its objectives.

Faculty have implemented DWM in Python POC in 10th release, version 2.21; automated data cleansing to include data corrections based on record-to-record comparisons with blocks and clusters; developed novel framework for scalable Entity Resolution using NLM for Locality Sensitive Hashing and Machine Learning achieving accuracy over 95% with a nearly linear runtime; developed computational framework integrating multi-layer genomics data to identify transcriptome and pathway dysregulations in autism spectrum disorder; investigated the expression alterations of survival-related genes in various immune

cell types when combining breast cancer bulk and single-cell RNA sequencing data; established a computational workflow of several combined machine learning approaches to identify biomarkers for both
prostate cancer using metabolomics data and chemotherapy-induced cardiotoxicity among breast; downloaded more than 300k bacterial and archaeal genomes from the NCBI and complete set of genomes from Integrated Microbial Genomes and Microbiomes project; developed and published a program, ProdMX, to speed up genome comparison more than a million-fold compared to traditional alignment methods; and published a paper giving an overview of multi-omics approaches in the journal, "Molecular Omics" among other accomplishments.

It is not clear that other components in DART or others in the data science community are using any of the algorithms or programs/processes developed by this research group. Recommend that this group do more to advance the use of their algorithms, perhaps by holding a workshop for DART faculty and graduate students and track the use of their algorithms in the broader data science community. The DWM programs may also be of interest to those college students in some of the new data science programs being developed around the state.

The evaluator has some concern that half of the Year 2 milestones are reported in the stoplight tables as "will be completed by end of reporting year". Of most concern are the activities for Goal 2.3: Harmonize multi-organizational and siloed data where over 90% of the activities are indicated to be completed by the end of the reporting year. This may be the result of teaching commitments during the school year and the summer may afford researchers more time to complete these milestones.

The Social Awareness component is meeting many of its objectives.

This component has researched representative algorithms, using threat models from four aspects: adversarial falsification, adversary's knowledge, adversarial specificity, and attack frequency;
developing metric to consider sensitivity level of each PII attribute and combined sensitivity of a given set of leaked PII attributes;
developed a novel adversarial adaptive defense (AAD) framework based on adaptive training; investigated interval-valued labels to enable a worker to specify both type-1 and type-2 uncertainties in his/her label without information loss; developed coded hate speech detection framework, CODE, to judge coded words used in the coded meaning; performed exploratory study on fairness-aware design decision-making; conducted link prediction in identity network based on social network, intra-layer and inter-layer link information; documented and disseminated findings of literature research of privacy-preserving data analytics algorithms and software; conducted survey of existing work using cryptography for privacy protected in federated learning and designed a cryptography-based solution.

The group may find that as they research aspects of marketing, the link predictions in identity networks may be impacted and additional variables will need to be added to their model. Similarly, the group may want to have other DART components utilize their privacy-preserving analytics in other content areas to test its applicability.

The Social Media and Networks component is meeting most of its objectives and exceed its targets for presentations and publications.

This component has determined key features and prepared the software design document for the cyber social network platform; developed, tested, and deployed data collection framework with real-time dashboard to monitor progress with alerting capabilities; revised taxonomy to characterize OIE based on social media platforms; studied cyber campaigns and characteristics of platforms and involved information actors and selected Hurricane Harvey to represented a large-scale and geographically widespread disaster scenario.

Research into the auto-annotation of multimedia data goal appears to be trailing the progress being made in the other goals/objectives. These data are becoming more prevalent every day and will be a vital part of the newly developed social media platform. The Year 2 objectives of 'obtain and index content types for at least two disaster scenarios' and 'developing GIS system to display real-time road status inputs' have not been met.

The evaluator has some concern that 70% of the Year 2 milestones are reported in the stoplight tables as "will be completed by end of reporting year". Of most concern are the activities for Goal 4.1: Mining cyber argumentation data for collective opinions and their evolution" where 100% of the activities are indicated to be completed by the end of the reporting year. This may be the result of teaching commitments during the school year and the summer may afford researchers more time to complete these milestones.

The Learning and Prediction research component is meeting many of its objectives.

This component has transitioned MTPP and LTSM to a convolutional neural network (CNN) approach; curated dataset of sensor data, system attributes, and failure/repair data of over 8,000 oil/gas wells; investigated efficacy of group structure on generalized neural network (GNN) architecture with smallest finite simple nonabelian group A5 action of random and clustered

synthetic small size data; introduced new Deep learning algorithms that perform well in low-cost platforms with high accuracy which significantly reduce computational time and memory consumption;
developed autoencoder method invention in unsupervised and self-supervised deep learning methods contributed to tackle problem of large-scale dataset management and labeling in Big Data management.

It is unclear to what extent the learning paradigms and algorithms are being used by other DART researchers or the data science community in general or whether the MTTP enhancements have been evaluated and assessed on real-world discrete data sets or that the MTTP/LSTM approach is scalable for implementation on different data sets.

The Education research theme is meeting most of its objectives.

The Education research theme has completed initial Middle School Coding Block Workshop with plan finalized and disseminated to stakeholders; 5-year plan outlined for stakeholders during workshop November 2020; UA Data Science program has been established and other colleges/universities are moving forward with developing data science programs for their respective campuses. Three Data Science for Arkansas workshops have been held during the first two years involving post-secondary academic Arkansas institutions, ADHE and ACDS.

The evaluator has some concern that 88% of the Year 2 milestones are reported in the stoplight tables as "will be completed by end of reporting year". This may be the result of teaching commitments during the school year and the summer may afford researchers more time to complete these milestones.

The Workforce Development component is 'On Track' with most of its objectives.

The Workforce Development component has participated in two EAST conferences and hosted two support/training webinars, as well as 2 statewide workshops on data science topics. They also have hosted a virtual ASRI which was completed by 42 undergraduates and involved a significant number of DART faculty and collaborators. It is unclear whether the objectives of 10 awards of $5k each per year for faculty training have been utilized or that 'internship opportunities for students at relevant companies have been identified'.

The evaluator has some concern that two-thirds (67%) of the Year 2 milestones are reported in the stoplight tables as "will be completed by end of reporting year". In addition, some milestones from Year 1 have been delayed, specifically related to internships for students and developing capstone courses for the data science curriculum. This may be the result of teaching commitments during the school year and the summer may afford researchers more time to complete these milestones.

The Communication and Dissemination component has met many of its objectives so far.

The Communication and Dissemination component established Slack group for DART faculty who utilizes it for email for daily communication; completed monthly team meetings for each research team; hosted 12 monthly webinars during the first two years; published a website; AEDC blogs; increased social media presence on Facebook and Twitter; formed Campus Communications Committee; held an annual retreat and annual all hands meetings.

Evaluations of the retreat revealed that the DART participants remained engaged during the two days. Most participants were appreciative of the opportunity to have the time in the breakouts to "meet members of other teams" and "learn better about the program and build potential research collaboration". Participants also expressed an appreciation for "the research teams to meet and explore individual contributions to the project. The webinar evaluations revealed that survey respondents 'Strongly Agreed' or 'Agreed' that the webinars were 'an effective way to present the information' (95%); 'presenter(s) were knowledgeable' (98%); 'increased my knowledge' (92%) and 'overall, well worth my time to attend' (93%).

The Broadening Participation component has met many of its objectives.

The Broadening Participation component awarded 2 broadening participation mini-grants of $5k each in Year 1 and 7 mini-grants a total of $15.5k in Year 2; hosted the ASRI for 42 undergraduates in June 2021-26% who are URM; mentored 15 undergraduates and 40 graduate students each academic year and supported 10 URM students each year in the SURE program.

The evaluation of the ASRI found that the two-week training program for undergraduates that was engaging, educational, inspirational, comprehensive, rewarding and fun for the participants. Faced with the challenge of Covid-19 the leaders did a lot of planning and hard work to pivot from a face-to-face Institute to one that was totally virtual. They did not try to replicate the face-to-face training but instead thoughtfully designed a wholly new program that maximized the strength of a variety of virtual tools available to them and the students. The BP group has implemented mentoring protocols and instruments that should help faculty better support students, especially those who are URM, in achieving their academic and career goals.

This group needs the support of the research themes to achieve its diversity goals of: Faculty (45% female, 10% URM);

Graduate Students (50% female, 20% URM); Undergraduate Students (50% female, 40% URM) and Advisory Boards (50% female, 20% URM). According to the self-reported participant data the DART participation data show: Faculty (30% female; 3% URM); Graduate students (31% female; 13% URM) and Undergraduate students (48% female; 45% URM).

While the ASRI and SURE programs, along with the new mentoring programs, will help the project achieve its diversity goals for undergraduate and graduate students. However, it is unclear how DART will achieve its faculty diversity goals. The project may want to include the benefit of broadening participation in research teams as a focus in its mentor training. This could have the benefit of expanding the impact of mentoring for both students and faculty recruitment.

Overall, DART 'Met or Exceeded Targets' in 13 (20%) with 77% of metric ‘On Track’ of the strategic plan metrics. A limited number of the metrics were 'not available' for assessment because of issues with timing or delays caused by Covid-19. The Cyberinfrastructure Component ’Met/Exceeded Target’ in 70% of its objectives and 10% ‘On Track’. The delay caused by UAF change in computer security policies resulted in 20% of metrics deemed ‘Limited Progress’. The Research Component ‘Met/Exceeded’ one fifth (23%) of their metrics and are ‘On Track’ with three-fourths (77%) of others. Publications, presentations and algorithm development are ahead of schedule. The Education and Workforce Development Components are ‘On Track’ with 100% of their metrics that could be measured at this time. There was the largest number of metrics not available to measure because of delays caused by Covid and timing in the project. The Communication Component met two out of eleven (18%) of its objectives, with 9(82%) ‘On Track’. The project website has been developed and the ER Core reporting system has been implemented. The Broadening Participation component is ‘On Track’ with 92% of its objectives and has ‘Met/Exceeded’ its objective of developing/modifying templates to implement a mentoring program.

APPENDIX

Appendix A
Strategic Plan Metrics: DART Wide Research Thrusts
Strategies, Outputs, Targets, Progress and Status

Strategies	Outputs	Targets	Year 1	Year 2	Current Status
Computer Infrastructure					
Hardware and software infrastructure improvements	Install, configure, and make available data science nodes on Pinnacle Portal	1	0	1	Met Target
	Science DMZ at UA and UAMS/UALR	1	0	1	Met Track
	100GB connection between ScienceDMZs	1	0	1	Met Track
	Establish dedicated DART Gitlab repository	1	0	0	Delayed
	Setup Globus data management services to point at DART storage arrays	1	1	0	Met Target
Documentations and user guides	Create technical management document defining organizational structure, roles, and responsibilities of ARCC	1	1	0	Met Target
	Amend existing MOU for ARCC expansion	1	0	1	Met Target
	UAF and UAMS will create CI Plans to support DART (1 x UAMS)	2	1	1	Met Target
	Create and publish document outlining GitLab user guidelines and minimum standard for code repository	1	0	0	Delayed
Instructors trained in software carpentry	CI=8	8	0	2	On Track
Workshops (Online)	CI=30; DC=0; SA=0; SM=0; LP=0	30	NA	5/195	On Track (5/195)
Research Themes					
Journal publications	CI=0; DC=28; SA=7; SM=9; LP=10	54	9	80	Exceeded Target (89)
Develop software modules/prototypes	CI=0; DC=0; SA=0; SM=3; LP=0	3	0	1	On Track (1)
Conference Workshops	CI=0; DC=5; SA=0; SM=0; LP=1; WD=15	21	1	3	On Track (4)
Conference presentations, seminars, papers and posters	CI=0; DC=25; SA=16; SM=5; LP=14	60	20	65	Exceeded Target (85)
Invention disclosures/patents	CI=0; DC=2; SA=0; SM=0; LP=0	2	0	0	On Track
Data Sets and algorithms/mathematical formulation	CI=0; DC=20; SA=4; SM=12; LP=1	37	2	5	On Track (7)
Applications and Platforms	CI=5; DC=0; SA=0; SM=2; LP=0; WD=1	8	0	1	On Track (1)
Workshops, demonstrations/trainings	CI=4; DC=4; SA=0; SM=0; LP=2; ED=16; WD=19	45	14	15	On Track (29)
Training and Webinars (Online)	CI=0; DC=0; SA=0; SM=0; LP=0; WD=8	8	1	5	On Track (6)
Proposals Submitted	CI=0; DC=0; SA=7; SM=0; LP=0/1	8	28	28	Exceeded Target (56)

Appendix B

Education and Workforce Development

Strategies, Outputs, Targets, Progress and Status

Strategy	Outputs	Target	Year 1	Year 2	Current Status
Education & Workforce Development					
Middle school curriculum developed	Posted at ADE and Co-ops	12	0	2	On Track (2)
	Piloted in schools	6	0	1	On Track (1)
Undergraduate Data Science degree program	Certification by ADE	1	0	0	On Track
	Evaluation report	2	0	0	On Track
	Implemented on campuses	3	1	1	On Track (2)
Collaborate with DART faculty to develop capstone projects	Develop capstone projects	25	NA	NA	NA
	Publish capstone projects	15	NA	NA	NA
	Student presentations at AHM	4	NA	NA	NA
Collaborate with ACDS to co-host workshops on data science topics	# held/# attending	5/100	1/##	1/##	On Track (2/##)
Annual workshops for faculty on grantsmanship and entrepreneurship topics	# held/# attended	15/100	1/##	1/##	On Track (2/##)
Seed grants for K-12 teachers	Number awarded	50	0	7	On Track (7)
	Awardee presentations at AHM	15	NA	NA	NA
Training grants for higher ed faculty	Number awarded	50	NA	NA	NA
Present (booth or breakout) at annual EAST conference	Number completed	5	1	1	On Track
Training Webinars	Number held/#attending	8/100	1/##	1/##	On Track (2)
Training sessions for K-12 teachers	Number held/# teachers attending	4/200	1/##	1/##	On Track (2/##)
Sharing platform for K-12 teachers	Established/# using	1/50	NA	NA	NA
Undergraduate research assistantships during school year	# assistantships/# unique students	75/125	25	41	On Track (41/41)
	% students supported URM/female	35%/25%	36%/48%	45%/46%	On Track
Undergraduate summer research experience	# URE/# unique students	50/35	10	10	On Track
	% students supported URM/female	50%/50%	NA	26%/51%	On Track
Graduate research assistants	# assistantships/#students	200/100	47	87	On Track (89)
	% students supported URM/female	50%/50%	13%/38%	12%/30%	On Track (13%/31%)
Industry internships	# internships/# students	20/10	NA	NA	NA
	% students supported URM/female	25%/35%	NA	NA	NA
	Evaluation form	1	NA	NA	NA
Master Theses	CI=0; DC=0; SA=7; SM=0; LP=3	10	1	NA	On Track (1)
PhD Dissertations	CI=0; DC=7; SA=0; SM=0; LP=2	9	1	NA	On Track (1)

Appendix C

Communication Component

Strategies, Outputs, Targets, Progress and Status

Communications					
All Hands Meeting	# held/# attending per meeting	5/100	1/110 (virtual)	1/125	On Track
	K-12 teacher presentations	4	NA	NA	On Track
	Poster competitions/# participants per competition	5/25	1/10	1/20	On Track
Annual retreat for faculty and grad students	# held/# attending per retreat	5/25	Covid	1/125 (virtual)	On Track
Monthly DART topical webinars	#held/#attending per webinar	55/20	5/25	6/20	On Track (11/22)
Monthly DART component team meetings	#held/#attending per meeting	330/10	88/##	96/##	On Track (184/##)
Project website	Published	1	0	1	Met Target
Quarterly Blogs posts about project	# posts/#views	20/200	4/##	5/##	On Target (9/##)
Maintain Facebook, Twitter and YouTube channels	Increase Following 10% a year	50%	7%	50%	Met Target (50%)
ER Core reporting system	Published and accessible	1	1	0	Met Target
	Annual webinar training	15	3	3	On Track
Science Journalism Challenge	Host/# attending per year	4/20	NA	NA	NA
	Invite SJC winners to AHM	20	NA	NA	NA

Appendix D
Broadening Participation
Strategies, Outputs, Targets, Progress and Status

Strategy	Outputs	Target	Year 1	Year 2	Current Status
Mentorship of students and early career faculty	Updated versions of University of Hawaii IDP templates	3	NA	3	Met Target
	Provide training to mentors on how to use IDP and to be a mentor	50	NA	NA	NA
Research Seed Grants Program	Develop and widely distribute RFP for research seed grants	5	1	1	On Track
	In conjunction with EAB and IAB select best RFPs and award funding	50	2	16	On Track (18)
	Research seed grant presentations to DART community	25	1	2	On Track (3)
	Early career faculty complete IDP templates pre and post	25	NA	NA	NA
	Connect early career faculty with senior faculty	25	2	16	On Track (18)
Summer Undergraduate Research Experienced (SURE) Program .	Develop and widely distribute SURE RFP for DART faculty to host students	50	10	10	On Track (20)
	Award maximum of $8k per faculty award	25	NA	NA	NA
	Recruit URM students	50	10	10	On Track
	Students complete IDP template at beginning and end of experience	50	NA	0	On Track
	Provide rewarding experience for SURE students	80%	NA	NA	On Track
Host ASRI	# held/# attending per ASRI	5/35	NA	1/42	On Track
	% students supported URM/female	50%/50%	NA	26%/51%	On Track
	Evaluation report	4	NA	1	On Track
	Scholarships for underserved	100	NA	Virtual	On Track