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

Year 2 Annual Report

Award Number: 1946391

Jurisdiction: Arkansas

PI: Jennifer Fowler

Co-PI: Jackson Cothren

Awardee Institution: Arkansas Economic Development Commission

Awardee DUNS: 619312569

Award Start Date: July 1, 2020

Award End Date: June 30, 2025 (Estimated)

Report Submission Date: April 22, 2022

Reporting Period: April 1, 2021 – March 31, 2022

Table of Contents

Overview

Vision

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

Mission

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

Goals

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

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

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

Key Accomplishments During Year 2

DART participants published 80 peer-reviewed journal articles, book chapters, and other relevant publications since the last report. 35 additional proposals were submitted by DART participants for a total request of over $21M, 12 of which were awarded for a total of about $1M from NSF and other funders. Participants also received a number of honors and awards. Dr. David Ussery (UAMS) was granted tenure as a full Professor. DART student Maryam Alimohammadi received a Gilbreth Memorial Fellowship from the Institute of Industrial and Systems Engineers (IISE) and a Doctoral Colloquium Travel Award to the Institute for Operations Research and the Management Sciences' 2021 Annual Meeting from the Forum for Women in OR/MS. Dr. Jack Cothren (UAF) was appointed for a 5-year position as Leica Geosystems Chair in Geospatial Imaging.

Dr. Fred Prior (UAMS) was promoted to Distinguished Professor, and named a member of the Editorial Board of Experimental Biology and Medicine and an AI and Machine Learning Section Editor of Cancer Imaging; he also served as a Conference Co-Organizer of the Wizardry of AI and Machine Learning in Cancer Imaging Conference.

Dr. Nitin Agarwal (UALR) received the 2021 UALR Faculty Excellence Award for Research and Creative Endeavors and was elevated to a Senior Fellow of the Institute of Electrical and Electronics Engineers, and his team won best paper awards at two International Academy, Research, and Industry Association's conferences: the 11th International Conference on Social Media Technologies, Communication, and Informatics (Barcelona, Spain) and the 7th

International Conference on Human and Social Analytics (Nice, France). They also won the Disinformation Challenge Award at the International Social Computing Conference (Washington D.C.) and were recognized as a Top 10 Team by NATO in their Innovation Hub Challenge. Dr. Agarwal has also been featured in articles and interviews in The Ritz Herald, the India West Journal, Arkansas Business, Arkansas Money and Politics, KUAR Public Radio (Little Rock), the Little Rock (AR) Daily Record, and the Magnolia (AR) Reporter as well as on the World Health Organization's African Regional Office's website and the websites for IT Arkansas and the Arkansas Center for Data Sciences.

DART participants from UAF, UAMS, and UALR were invited to make presentations at the University of Oklahoma's Industrial and Systems Engineering Graduate Seminar Series, the Wizardry of AI and Machine Learning in Cancer Imaging Conference, the Thermal Transport Café, the 2nd International Workshop on New Approaches for Multidimensional Signal Processing, the International Conference on Automation Control and Mechatronics for Industry 4.0, the Annual Meeting of the Institute for Operations Research and the Management Sciences, and the Bimonthly Seminar Series of the Korean-American Scientists and Engineers Association (Arkansas Chapter). Additional presentations were made by speakers from UALR, UCA, SAU, UAF, UAMS at the International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation, the Annual Conference of North American Fuzzy Information Processing Society, IntelliSys 2022, and the Institute of Industrial and Systems Engineers' Annual Conference.

Additionally, UAMS and UAF participants exhibited posters at the UAMS Graduate Faculty in the Interdisciplinary Biomedical Sciences Seminar Series, the NSF EPSCoR Education, Outreach and Diversity Council and the EPSCoR/IDeA Foundation's BetterPoster competition, the 13th Annual Teach the Teachers Symposium, the Annual Meeting of the Shock Society, the American Society for Microbiology World Microbe Forum, and the IDeA Central Region's Annual Meeting.

DART participants also produced a number of research products including the new public-facing DART Project website (https://dartproject.org/) which includes information on, and access to, webinar schedules, webinars, and DART FAQs. Dr. John Talburt and student Kris Anderson (UALR) produced a user guide for the Data Washing Machine, a Python program stored in BitBucket for automatically cleaning and clustering certain types of data files (primarily for identifying multiple sources of the same information).

At UAMS, seed grant recipient Dr. Kevin Phelan produced two videos for use in the data science pilot program for Arkansas middle schools: the first introduces teachers to the Common Online Data Analysis Platform (CODAP) program and how it will be used while the second introduces students to how large datasets from the NOAA "data in the classroom" website can be imported into R and visualized, plus a demonstration of the Orange3 software for data mining and image embedding. Dr. Phelan, along with Dr. Tiffany Huitt and Annice Steadman, also produced a detailed teacher guide with student worksheets to accompany this second video.

Several DART participants collaborated to develop the Transportation and Maritime Analytics Partnerships Hub (TransMAP) website in a collaboration between the Center for Advanced Spatial Technologies and Maritime Transportation Research and Education Center (MarTREC) and supported by the Maritime Administration of the U.S. Department of Transportation. The team, led by UAF in partnership with the Texas A&M Transportation Institute, makes available large-scale data and visualization tools related to maritime freight transportation on infrastructure, systems, and networks accessible to humans and machines through the Internet of Things, in order to enable improved resilience, planning, investment, and operational decisions. These tools have been adapted to support DART projects (https://oak.cast.uark.edu/explorer).

Big Data Management

Our contributions to improving the use of big data are both internal (improving the ability of DART researchers to work with big data sets) and external (the larger research questions described earlier). Internally, the CI team made signification progress developing the administrative structures, network architecture, equipment upgrades, and policies required for secure, easy, robust, and sustainable off-campus access to resources – compute notes and attached storage – that make up the Arkansas Research Platform (ARP). DART researchers now have priority access to over forty GPU and large memory nodes on Pinnacle, and normal access to HPC resources at UA and UAMS that were available before DART. Access to Open OnDemand Portals is now open to all incoming connections by authorized users on recognized research networks and two-factor authentication is available when incoming from other networks. Trainings to use ARP and its applications (programming and Globus data transfer, for example) are underway at an increasing pace.

Research into the better ways to automatically curate big data sets is well underway in the other research teams. The social awareness team mined thousands of English sentences revolving around controversial social issues, such as guns on campus, vaccine mandate etc. The team, consisting of computer and social scientists, created several datasets based on the text data that will be shared throughout the project team. The social media team developed a taxonomy to characterize the online information environment (OIE), based on a variety of social media platforms, various cyber campaigns, characteristics of platforms, and information actors involved. This will lead to more systematic research in understanding and addressing spread of mis- and dis-information and other deviant behaviors in these environments. As part of this work, they curated 9.6TB of Twitter data for studying health behavior and developed a state-of-the-art procedure to access and retrieve posts using a list of keywords.

Meanwhile, learning and prediction researchers completed development and implementation of a CNN-based architecture that allows tabular data to be represented as image pixels so to exploit the benefits of CNNs stemming from the inductive bias associated with pixel neighborhoods. They also invented autoencoder method that improved unsupervised and self-supervised deep learning methods' ability to manage and label much larger datasets. Data curation researchers developed a novel big data management framework that integrates machine learning with locality-sensitive hashing. The devised framework can transform large volume structured or un-structured data stream efficiently into data buckets based on the semantics or similarity. The data curation team generated thousands of SARS-CoV-2 genome sequences on the Arkansas Research Platform and are in the process of comparing them to the 8-million sequences in GenBank and GISAID.

Privacy and Security

Secure access to the ARP was, in many ways, the focus of the CI team in year 2. The high-performance computing centers at UAF and UAMS worked directly with the ITS departments at both universities to redefine and reconfigure their respective ScienceDMZs to support faculty, staff, and students from other campuses. Before DART, access to the Arkansas Research Platform resources outside their DMZ's was only available via the Secure Shell Protocol (SSH). DART researchers are far more likely to use the interfaces provided by the Open OnDemand portals, but these were not securely accessible. The current configuration is stable and can be used directly from all research and education networks. However, it does not yet support federated identity mechanisms. The SHARP*CCI working group – funded by a separate NSF CC* planning grant – worked with the DART CI team to develop templates for Master Information Security Plans and System Security Plans following NIST 800.171 requirements. These will allow isolated computer systems to store controlled unclassified information. Several DART researchers may use this capability in subsequent years and other researcher throughout the state will be able to use this capability.

The DC team's positive data curation prototype – also sponsored by DART industry partner SAIC - demonstrated the ability to control both access and metadata reporting for data operations in the Hadoop environment. SA researchers developed and published a suite of novel algorithms (differential privacy preserving multi-party learning, fair and robust learning under sample selection bias or attacks, uncertainty award crowdsourcing, fraud and hate detection in cyberspace, user-centric data sharing in cyberspace, and privacy-preserving analytics in health and genomics). DART researchers developed a novel statistical test based on the Min-Max ratio to handle statistical comparison applications considering privacy protection. It allows for testing hypotheses on the mean and variances of two groups of samples following normal distributions without requiring data labelling.

Model Interpretability

In Year 2, DART researchers devised a causal inference framework which, to our knowledge, is the first use of causal inference in un-structured data. This framework can explain how models produce results through causal inference, which increases interpretability of the predictions. The team is also exploring the use topological data analysis paradigms for identifying data quality faults and possible automatic solutions. Topological tools rely on persistence features, simplicial complexes, and Morse-Smale complexes, which tend to provide a better explanation, based on the morphology of the data, about why the algorithms decided to a course of action over another.

DART researchers developed and trained a deep-learning convolutional neural network to geolocate in real time unmanned aircraft using only downward-looking images from the aircraft and reference imagery embedded in the neural network. There are two major contributions of this published research. First, it demonstrated the ability to efficiently encode several terabytes in a relatively small network. Second, it included a probabilistic target function that ranks multiple locations based on the match probability like many crafted feature image matching techniques.

Social awareness examined the potential risks of deep learning models under adversarial attacks in Year 2 and designed a framework that seeks to effectively generate poisoning samples to attack both model accuracy and algorithmic fairness of fair machine learning models. Three online attacking methods, adversarial sampling, adversarial labeling, and adversarial feature modification, effectively and efficiently produce poisoning samples via sampling, labeling, or modifying a fraction of training data in order to reduce the test accuracy. The attacking methods can flexibly adjust the attack's focus and accurately quantify the impact of each candidate point to both accuracy loss and fairness violation, thus producing effective poisoning samples. We conducted experiments on two real datasets and results demonstrated the effectiveness and efficiency of our attacking framework. This framework will, we think, be invaluable not only to upcoming DART work, but to this entire research community.

Another aspect of model integrity concerns privacy preserving mechanisms in deep learning algorithms. Our researchers looked closely at the widely used differentially private stochastic gradient descent (DPSGD) method which applies gradient clipping and random noise addition in the training stage but often incurs disparate impact: the accuracy of a model trained using DPSGD tends to decrease more on minority subgroups vs. the original, non-private model. The inequality in utility loss due to differential privacy can be understood by comparing the changes in prediction accuracy with respect to each group between the private model and the non-private model. The cost of privacy with respect to each group can now, for example, be explained by how the group sample size and other factors relate to the privacy impact on group accuracy.

Data Science Educated Workforce

In year 2, we took steps to further integrate the work of the education and workforce development activities more directly with the other research activities. It now appears as one of four barriers to data science advancement and acceptance reflecting the need for a data-aware professional workforce and public (one of the four main DART topics). We have identified the first group of DART faculty who will work with the ED team to develop capstone projects based on DART research.

At the time of this report, we now have 3 institutions – the University of Arkansas, the University of Central Arkansas, and Arkansas State University- with active four-year data science programs. These institutions will serve as the ‘hubs’ in the hub-and-spoke data science educational ecosystem. The institutions in Cohort 1 are in various stages of local and state approval on their programs, including three of the state’s four historically black colleges and universities- Philander Smith College, University of Arkansas at Pine Bluff, and Shorter College. North Arkansas College, a rural community college, plans to have an associate’s degree program in place by Fall of 2022.

We implemented several surveys since the last report to get a better understanding of the institutional needs, strengths, and employer needs related to data science education and workforce. Through these surveys, we found that many interested campuses are still using extremely outdated computer software and hardware, and some do not have sufficient internet bandwidth or wi-fi to implement DASC courses. We hope to address some of those challenges through the CC* effort. Another key development is authorization from the Arkansas Division of Higher Education (ADHE) for fast-track approval of any proposed DART-DASC degree programs.

Significant Problems

Redesigning the UA research network (the ScienceDMZ) to support off-campus access by methods other than Secure Shell Protocol (SSH) proved to be a longer and more difficult effort that we anticipated when we developed the strategic plan. Delays in defining authorization and authentication mechanisms that maintained the security of UA Enterprise systems, created ripples throughout the year 2 objectives and has delayed the use of the Arkansas Research Platform across the project. In particular, the SM team requires near-real time HPC/GPU computing and large-scale and fast storage due to high volume and velocity of social media data. Furthermore, the data curation team noted a significant problem with their inability to make significant progress to migrate successful data cleansing models into a scalable process. In the original plan, the Python version of the data washing machine was to be refactored into a Hadoop map/reduce program scalable to very-large datasets. However, the DC team has only recently hired a graduate student with the background to implement the models a Hadoop Spark application was not successful. While the DWM has been implemented as a Java application which runs much faster than the Python version, it is still not a truly scalable process suitable for processing very-large datasets. We are working to solve this problem and hope to report success in Year 3.

The DART strategic plan leveraged the UAF Enterprise GitLab installation to provide an on-premises repository that researchers trusted to store and share private or embargoed code with other DART researchers. In Year 2, UAF internal IT support decided to move to a commercial cloud hosted GitLab repository. The cost to purchase and maintain an ARP-specific repository was unsustainable. While these delays in providing state-wide access to ARP resources were significant, they have not resulted in strategic changes. All DART personnel and campuses now have access and cyberinfrastructure staff are working to catch-up by accelerating their training activities. However, the delayed GitLab deployment does necessitate a change in strategy regarding a DART centralized repository. The current solution being developed involves a DART GitLab organization at gitlab.com and sharing code from private and public repositories. Git is installed on both Pinnacle and Grace and researchers can clone or copy from DART GitLab via SSH.

Research participants identified a need for additional programming support for many projects in the form of deploying code on ARP clusters (Grace and Pinnacle). While significant funds have been allocated to this (Dr. Angel at UAF, and a Postdoc at UALR), more talent is needed. These developers draw large salaries in business and finding good developers is hard given relatively small salaries at state universities. However, CI is leveraging existing resources to try to fill this gap and moving existing DART participants into these roles. Goal two of the cyberinfrastructure team faced difficulties in hiring critical system administration and retaining data visualization staff.

In addition, unexpected difficulties in equipment purchases and long lead times for delivery of computing equipment were frustrating but turned out to not delay the expansion of Pinnacle. However, purchasing networking equipment to be housed at ARE-ON with funds allocated to UAMS did add some delays to the expansion of the ScienceDMZ at UAMS. UAMS is in the process of restructuring its network and changing equipment vendors which has further delayed the purchase of network equipment. This was exacerbated at two campuses – UA and UALR – by the continued efforts to learn and implement the University of Arkansas System mandated conversion to the Workday Enterprise Resource Planning platform. This will also impact the project in year 3 (although we anticipate fewer transition issues) at the University of Arkansas for Medical Sciences.

COVID-related travel restrictions and other pandemic factors have led to decreased access to graduate student applicants and have created staffing issues. Student recruitment suffered and we observed increased student turnover (due to health reasons, or didn't want to continue online education experience, etc.). This turnover resulted in additional time invested to bring the newcomers up to speed. COVID concerns also severely limited face-to-face meetings in year 2. The project has yet to meet in person, but is looking forward to our next opportunity at the May 16-17 annual meeting in Little Rock.

Another significant challenge is mental health for faculty, staff, and students. With the combination of the pandemic, the climate crisis, the war in Ukraine, travel bans, etc. everyone is experiencing trauma, stress, and anxiety. All of this not only takes a toll on mental health, but also has slowed down the relationship-building that is so critical in a collaborative science project. Many of us have lost family members and friends, and have not been able to travel to home countries to see the friends and family that survived. Campuses are running out of counselors and mental health practitioners to help students and faculty. We've witnessed increases in lots of other issues that negatively affect home life- economic stress, domestic violence, political tension, and racial injustice. Unfortunately, DART as a project cannot do much to mitigate these larger social issues, but we are supporting each other as much as possible. To reduce Zoom burnout, we thoughtfully design every agenda for online meetings to make the most productive use of everyone's time. We began ending each meeting with sharing things to celebrate (birthdays, awards, publications, other personal achievements) so we can end each meeting on a positive note. We hired a great Science of Team Science facilitator, who is also a licensed social worker (Dr. Anne Hebeger Marino of Lean To Collaborations), to help with our DART Virtual Conference last September, and we plan to continue to work with her and others in the field.

Novel Opportunities

The award of the SHARP CCI (NSF award #2126108, start date 09/01/2021) and with it, the creation of a large working group of IT professionals across the state is the most significant new opportunity in year 2. This working group is composed of all institutions collaborating in DART and is developing a shared ARP Master Information Security Plan (MISP), resource-specific System Security Plans (SSP), policies for access and using ARP, and polices for securely and reliably adding resources (instruments, compute nodes, storage arrays, etc.) to the ARP. It has enabled DART CI to engage far more deeply and regularly with the IT staff at participating institutions and build the working relationships necessary for sustaining the ARP beyond the end date of the DART grant. It has also accelerated the process of requirements gathering from these institutions both in terms of computing and security as well as overall governance.

The extended pandemic is also creating large amounts of many types of data that DART researchers are using, and despite the significant delays in access to ARP, the DART computational infrastructure is proving useful in helping to deal with this.

The SM team has explored new collaborations with other DART teams that are using big data techniques for multimedia processing to better meet their objectives. A new collaborator was introduced who brings expertise in fusing multi-modal datasets to observe and/or predict damages after disasters. Using blocking files, pairs of generated tokens and record identification values, the team developed a method to build graph connecting records with each other. The procedure follows from the token files, transforming it to a “wide” representation in a matrix of [records x tokens] with a binary value on each entry that indicates if the token is present in a given record. The multiplication of the matrix times its transpose is a [records x records] matrix that counts how many tokens connect one record with another on each entry. This connectivity matrix can be used to compute distances using NLP models with increased efficiency and, later, create clusters of records for entity resolution. This idea is still under exploration.

One technique for topological data analysis, mapper, has shown success in reconstructing pictures that have missing areas. One opportunity for the data curation may be the use of a mapper to reconstruct records that have quality defects from the rest of the data in the data source.

Another new opportunity is the development of a distributed/virtual DASC department for the state, which will address the gap in instructors and enrolled students in budding DASC degree programs at institutions statewide. This idea was discussed in the proposal but a real solution had not been identified until Year 2, when we discovered two viable options that are now under exploration. This is further discussed in the Education section of this report.

Changes in Strategy

The project only has a few changes in strategy to report. The LP team plans to promote a centralized data repository that would be beneficial as our team moves forward with development of learning and prediction models. We have witnessed an increased emphasis on training and deployment of research cloud computing, and are seeking additional software development support to the CI team. Until the remaining CI needs are addressed, we plan to seek extramural funding for equipment through MRI/DURIP and other grant programs.

The DC team moved the development activities for Objective 2.2: "Build a POC and demo for Positive Data Control (PDC)" into the commercial arena with sponsorship from SAIC. SAIC is still evaluating the potential for the commercial development of this technology as a software product.

Personnel Changes

The project has experienced a number of personnel changes since the last report. At the central office, PI Steve Stanley retired in May of 2021 and the former EOD Jennifer Fowler was promoted to PD/PI. Brittany Hillyer, MEd was hired as the Director of Education, Outreach, & Diversity in October of 2021. She is a former educator and came to AEDC from the Arkansas Department of Education. Tom Chilton, the division director for Science & Technology (the unit that houses EPSCoR at AEDC) retired in August of 2021. AEDC's new authorized organizational representative, and PI Fowler's supervisor, is Jennifer Emerson, the Executive Vice President for Operations.

Dr. Samar Swaid (ED) took a sabbatical from Philander Smith College and was replaced in the project by Dr. Chuanlei Zhang. Dr. Olcay Kursun (LP) left the state and was replaced in the project by Dr. Sinan Kockara, Dr. Paul Schrader (LP) also left the state and was replaced by Dr. Ahmad Al-Shami.

Research & Education Program - Year 2 Accomplishments

Coordinated Cyberinfrastructure

Arkansas Research Platform and Research Computing Collaborative: In year 2 the cyberinfrastructure team engaged with an expanded number of departments and institutions to enable the Arkansas Research Platform. In doing so, we found it necessary to shift by one year all the milestones in Objective 1.a. regarding the Arkansas Research Computing Collaborative. This will allow us to integrate findings and decisions of the newly formed and funded SHARP CI working group. Despite supply change challenges through the year, we were able to put all nodes into operation in year 2 and, with network changes to the UA ScienceDMZ, make them available to all DART researchers. The impact of this upgrade cannot be overemphasized. DART funds directly increased overall ARP computing capacity (and added much needed

GPU-nodes) by 44%. But the DART purchase order also included a significant number of purchases with other funds resulting in a 73% increase in total ARP resources.

ARP Clusters	Theoretical Peak Teraflops	Comments
Trestles and Razor	214	Legacy systems
Pinnacle 1 (CASE)	178	Funds from previous Track 1 award (CASE)
Pinnacle 1 (condo)	37	Nodes purchased with research non-EPSCoR funds
Grace (UAMS)	365	Nodes purchased with UAMS funds
Pinnacle 2 (DART)	351	Nodes purchased with DART funds
Pinnacle 2 (condo)	229	Nodes purchased with other resources but on the same PO as DART funds and decreasing price per node.
Total	1,374	

Equipment Purchases. The purchase of equipment necessary to establish a 100Gbit connection to UAF has been delayed due to purchasing issues but may still reach completion in year 2. Establishing an accompanying ScienceDMZ that is compatible with the ARP plan is a topic being addressed by the SHARP CCI working group.

DART Data Sharing & Management. All activities related to Globus and GitLab have been moved from Year 1 to Year 3. The delay is partly due to the long series of meetings with the UAF IT Services group to reconfigure network architecture and create new access policies required for ARP resources at UAF to accommodate off-campus access. It is also affected by the activities of the SHARP CCI working group established early in Year 2. However, while the SHARP CCI working group develops a federated identify solution, a secure but less efficient method of provide access to ARP resources at UAF and UAMS has been established. DART participants are now able to gain access to interactive sessions (VM's, Jupyter notebooks, MATLAB, RStudio, and more) on nodes via identical Open OnDemand portals on the Pinnacle and Grace clusters. Access to storage arrays at UAF and UAMS are available in these sessions or through Globus as endpoints. The eventual goal of a federated identity mechanism (likely InCommon) is delayed until the SHARP CCI working group completes its task towards the end of Year 2. This solution will be deployed in Year 3.

The GitLab component of Activity 2 (Methods of access to GitLab and ARP) was also affected by UAF network and policy changes. The DART strategic plan leveraged the UAF Enterprise GitLab installation to provide an on-premises repository that researchers trusted to store and share private or embargoed code with other DART researchers. In Year 2, UAF ITS decided to move to a commercial cloud hosted GitLab repository. The cost to purchase and maintain an ARP-specific repository was unstainable. The current solution being developed involves a DART GitLab organization at gitlab.com and sharing code from private and public repositories. Git is installed on both Pinnacle and Grace and researchers can clone or copy from DART GitLab via SSH.

Activity 3 has also been pushed out to Year 3 to allow time to reconsider the need for a Globus Data Management contract. Globus is being reconsidered because of ScienceDMZ changes, access patterns by DART institutions, and the ability of the no-cost version of Globus to meet current needs. We will reconsider the full data management contract during year 3. Funds that were allocated to Globus licenses in years 1 and 2 were re-budgeted to address federated identify management solutions for easier, secure access to ARP. SHARP CCI planning is addressing this in year 2 and are focused on how to use best use these funds. Discussions center around InCommon membership for DART campuses but the planning is also addressing sustainability beyond the grant.

Mechanisms for storing HIPAA-related material exist at UAMS and these policies and procedures were presented as part of the SHARP CCI System Security planning. It is agreed that the NIST 800 framework for risk assessment and risk-based management form the basis of this effort. Prototype secure enclaves have been created at UAMS and experimentation using a container-based approach with Kubernetes orchestration are underway.

An ARP-wide deployment model is not yet defined although System Security Plans to host HIPAA and Controlled Unclassified Information (defined under NIST 800.171) have been developed at UAF. An SSP is being developed for the Center for Advanced Spatial Technologies (CAST) at UAF and is expected to be finished by the end of year 2. However, this SSP will only apply to a very specific resource in CAST and not to any existing ARP resources. It will nevertheless provide a completed template for more rapid and effective implementation of future SSPs.

Visualizing Complex Data. A postdoctoral fellow position was created and hired in year 2 to address some of the activities related to CI Goal 2. The postdoctoral fellow, Dr Zhao, started on August 1, 2021, and left on February 15, 2022. A new postdoctoral fellow hiring process is currently in progress. Unfortunately, the remaining DART participants faculty, Drs. Springer and Rodriguez-Conde, are unable to provide the time necessary to act as substitutes for the postdoctoral fellow's efforts in the interim. There is a high probability that activities slated for year will need to be moved (even) further out and/or reduced.

A systematic literature review on advanced visualization and immersive analytics has been conducted. Its results are internally reviewed and the report is ready for publication on the DART website.

Data-driven modeling and photogrammetry-based visualization were incorporated by Dr Zhao into virtual-reality experiences for undergraduate geoscience education. A paper was published reporting on the results of a user evaluation on different design choices of virtual field trips in the Journal of Educational Computing Research. Dr. Zhao collaborated with Prof

Rui Li, State University of New York at Albany, to examine effects of an augmented-reality display on the windshield of self-driving vehicles for drivers' spatial awareness.

Dr. Zhao teamed up with DART researchers Schubert and Cothren (UAF), and Toby Teeter, Director of the Collaborative in Bentonville, AR, to design a virtual collaborative space in both augmented and virtual reality to improve coordination between data science educators and industry partners in Arkansas. The preliminary results were presented in the Data Science for Arkansas Workshop on Dec 10, 2021. Dr. Zhao also worked with Dr. Kusum Naithani, UAF, to develop an immersive workbench to visualize and analyze environmental data collected from the rural Borneo Highlands for use in ecological research and education.

Feedback from External Advisory Board. One recommendation from the EAB was to perform thought experiments to better understand bandwidth requirements across the HPC/storage/visualization requirements in the ARP. We are addressing this recommendation by distributing (through SHARP) the Research Computing and Data Capabilities Model (RCD-CM) developed by the Campus Research Computing Consortium (CaRCC). This model is designed for institutions to assess their "support for computationally- and data-intensive research, to identify potential areas for improvement, and to understand how the broader community views Research Computing and Data support". The model addresses a wide range of roles at the university (administrator, faculty, ITS technician) and is applicable to small and large, public and private institutions. Based on responses to this survey, the CI team will consider the current and potential research which must be supported by the ARP and will develop experiments to test the ability of the ARP to support it. We anticipate that curation and analysis of gene sequences will be a major priority and poses some unique challenges. Storage of sequences at one site (UAMS for example) and processing at another (UAF), while visualizing at another (UALR) can be tested with the current architecture. There are other obvious thought and physical experiments involving processing and visualization of sUAS data collected and stored at one site, processed at another, and visualized at yet another.

The EAB also recommended more innovative CI research. One area in which we are exploring innovations is in access and analysis of geospatial data – including geotagged images, traditional spatial data, high resolution overhead imagery, and mobility data - across computing environments. One effort in this regard was a proposal to solicitation NSF 20-592 (Cyberinfrastructure for Sustained Scientific Innovation (CSSI): Elements and Framework Implementations) led by Cothren and Angel in which we collaborated with Purdue University to propose a new geospatial architecture that would create a shared platform for analyzing historical declassified satellite imagery. While this proposal was not funded, it has led to similar projects with both USGS and the National Geospatial Intelligence Agency (NGA). A similar architecture is being developed for one of the major goals in social media and networks and we will consider proposals to CSSI for this work as well.

Data Curation and Life Cycle Analysis

Automate Heterogenous Data Curation. Talbert's research group at UALR, now enhanced by researchers at UAF and UALR, released a new Python-based proof of concept application for the “data washing machine” (DWM). Version 2.21 introduces improvements made in year 2 of DART that improve the values precision and recall metrics attained in linking the eighteen base data sets. The team implemented a robotic process that enables researchers to easily perform a grid search across DWM parameters to find settings that produce the best precision and recall clustering results. Work is also underway to automatically set the DWM parameters that give the best results based on intrinsic characteristics of the input data set. All the characteristics can be generated by unsupervised processes to calculate things like the number of tokens, number of unique tokens, number of numeric tokens, mean and standard deviation of the token frequency distribution. The current approach is to find the closest match between the characteristics of a given input dataset and previously processed datasets in the repository, then use the DWM parameters for the best match to process in the new input. Going forward, the plan is to replace the search and match with a machine learning model trained on prior results. From the research, the team is developing a new unsupervised data quality assessment for data redundancy using an entropy-based “cluster quality metric”. The metric is a score from 1 to 0 where 1 represents a perfect cluster (i.e. all records have exactly the same tokens). The redundancy measure is simply the average cluster quality over all clusters created at a fixed similarity level. Other improvements focused on the zero-shot learning capabilities by incorporating natural language processing models (BERT and RoBERTa) to project records into latent vector spaces and use distance metrics to cluster references using hierarchical clustering and affinity propagation.

To further enhance the curation process embedded in the DWM, members of our theme worked with Springer in cyberinfrastructure to develop visual representations of topological data analysis of selected test data sets. The topological analysis works by embedding textual data in vector spaces allowing the textual data to be simplified and segmented using traditional topology-based tools. Visualizations of this process revealed that records representing the same entity seem to fall mostly within the same segmentation cells. This has the potential to detect and remove duplication either directly or by defining blocks that can be processed much faster with the algorithms developed in parallel by the team.

The team continued to implement the idea of collaborative and need-based data collection mechanism in decision making under uncertainty for disaster relief. The goal is to use the information from satellite images, ground-based images, and text acquired from social media sequentially to quantify damages in the critical transportation networks. To identify the geographical location related to an image acquired from social media, basic data query and collection scheme were developed for image data collection from Google Street View. A sampling strategy and general procedure for automated data retrieval are currently under development. Initial experiments on damage assessment were conducted based on a sample image dataset previously collected from the social media team. General classification on “raw” images and automatic pipeline from collection to assessment is still under development.

A method for automatic data collection from Twitter’s open API was developed. Tweets can be selected based on a wide arrange of criteria, and these can be tailored to fit specific reports on a disaster event. Basic data query and collection scheme have been developed. A general procedure for automated tweet retrieval and insights extraction using Natural Language Processing is still under development.

We studied single-cell RNA sequencing (RNA-seq) data of chronic-phase chronic myeloid leukemia stem cells to investigate the genetic variations underlying drug response. Tyrosine kinase inhibitors (TKI) were developed to target the BCR-ABL oncoprotein, inhibiting its abnormal kinase activity. TKI treatments have significantly improved CML patient outcomes. However, the patients could develop drug resistance and relapse after therapy discontinues due to intratumor heterogeneity. We applied a t-stochastic neighbor embedding(t-SNE) is non-linear dimensionality reduction approach to cluster cancer cells into different cell clusters. We found that cells with distinct responses to TKI, good vs poor were clustered together in both BCR-ABL positive and negative cells. Furthermore, we identified a set of genes that were concordantly differentially expressed in cell clusters. The putative transcription factors of these differentially expressed genes were revealed by single-cell regulatory network inference and clustering approach. This work offers new insights into TKI resistance in CML.

Year 2 has seen significant advances in the automation of data cleansing. The initial approach developed in Year 1 was to simply examine high-frequency, low-frequency token pairs as possible misspellings. Year 2 has seen this expand to include data corrections based on record-to-record comparisons with both blocks and clusters. The team has developed 7 techniques for record-to-records correction including some that can impute missing values and correct incorrectly split or joined tokens.

We developed a novel framework for scalable Entity Resolution using Natural Language Model, Locality Sensitive Hashing and Machine Learning. The preliminary experiment result shown a promising result, which achieved accuracy over 95% with a nearly linear runtime.

We developed a computational framework that integrates multi-layer genomics data to identify transcriptome and pathway dysregulations in autism spectrum disorder. Combining gene expression, protein-DNA interactions and genome-wide enhancer locations, we inferred regulatory networks differentially expressed in the disease samples as compared to control samples. This regulatory network approach centered at transcription factors provides a unique way to reveal master regulators, which position at the top of regulatory hierarchies and control the transcriptional activities of many downstream genes. The regulatory cascades approach established in the study offers a framework for revealing new disease-related genes and can be applied and extended to study other tissues and diseases.

Also, we combined breast cancer bulk and single-cell RNA sequencing data for investigating the expression alterations of survival-related genes in various immune cell types. Breast cancer was initially considered as a non-immunogenic disease. Recently, several studies have demonstrated the efficacy of immunotherapy in breast cancer treatment. Multiple survival-related genes were simultaneously differentially expressed in the CD4+ and CD8+ T cells. Our works help us to better understand the interactions of tumor and immune systems and provide novel molecular prognostic markers for survival prediction in breast cancer patients. The developed method can be applied to study other types of cancer.

To this point, all of the focus has been on unsupervised data quality assessment, cleaning, and clustering. So far, no work has started on data integration. This partly due to not having generated or acquired appropriate data sets for experimentation. We are currently exploring working with our industry partners (PiLog and SAIC) ways we could help with their data integration issues.

Explore secure and private distributed data management. All the progress on the Positive Data Control (PDC) has been done through a proof-of-concept (POC) with our industry partner SAIC. They are interested in implementing, and possibly commercializing, PDC for their operational systems and clients. The initial proof of concept for SAIC implement both downward access controls and upward metadata reporting for HDFS and Hive in the Hadoop environment. The primary control agent was Ranger with Atlas being used as the policy store. It also demonstrated attribute-based access control (ABAC) for Hive tables. The POC concluded in January 2022 and is currently on hold as SAIC evaluates the results makes a go/no-go decision on further implementation.

Harmonize multi-organizational and siloed data. We are using Amazon S3 (“Amazon Simple Storage Service”) buckets on our large (2 petabyte) object store, for storing and retrieving millions of SARS-CoV-2 genome sequences. These pipelines will be extended to other genomes. We are also developing a variety of genome visualization tools that are being used in our courses and workshops (“R-BioTools”, which is currently living in a GitHub directory managed by Hanna Ford at UA Fayetteville).

We have applied this to both viral and bacterial genomes; the SARS-CoV-2 was published in February, 2022 [https://doi.org/10.1093/femsre/fuac003], and we have two manuscripts in preparation that will be submitted in March / April, 2022, about using automated genome quality scores in helping to cluster genome species.

Dr. Se-Ran Jun and her team will develop genome databases specially for ESKAPE pathogens for genomic surveillance purpose. This activity will involve “automate data cleansing”, “automate quality control”, “automate clustering”. The automated pipeline has been tested for *Enterococcus faecium, Salmonella enterica, and Klebsiella pneumoniae.* We will extend this pipeline to ESKAPE pathogens, leading causes of hospital acquired infection.

We started collaborative efforts to apply self-supervised learning approach to modeling and integrating Omics data for diverse biomedical phenotypes. The goal is to train an Omics model that can be used for different downstream machine learning tasks for modeling, prediction and causal inference of diverse biomedical phenotypes such as cancers.

The emergence of single-cell sequencing technologies has enabled the production of high-resolution data at the individual cell level, providing opportunities to capture cell population diversity and dissect the cellular heterogeneity of complex diseases. At the same time, relatively high biological and technical noise poses new challenges for single-cell data analysis. The single-cell RNA sequencing (scRNA-seq) data often contains substantial missing values due to gene dropout events. We developed a convolutional neural network-based model to recover missing values for scRNA-seq data. The probability of dropout was computed using the gamma-normal expectation maximum algorithm. Unlike most existing approaches, our model only recovered the expression values that have a dropout probability larger than a threshold. The mean square error and Pearson correlation coefficient were used to assess the accuracy of predicted expression values. The purity and entropy were calculated to measure the homogeneity of cell clusters using imputed gene expression profiles. Across various scRNA-seq datasets, our model demonstrated robust performance and achieved comparable or better results compared to the other imputation methods. We are developing several proteogenomic pipelines, which are necessary for integrating proteomics data with genomics data.

Dr. Se-Ran Jun and her team have identified a novel pattern related to daptomycin resistance through big data analysis of genomes for the first time. This novel pattern suggests a new paradigm of daptomycin resistance dissemination. They also established a computational workflow of several machine learning approaches combined to identify biomarker for prostate cancer using metabolomics data. This work will be presented at the American Association for Cancer Research conference. Work is underway to develop a computational workflow to identify biomarker for chemotherapy-induced cardiotoxicity among breast cancer patients using metabolomics data, as well as investigation of potential new antibiotic resistance genetic markers using known markers in *Enterococcus faecium* using machine learning approach and population structure of the species.

Feedback from External Advisory Board. The only comment related to DC in the report last year was related to the novelty of our work. We have no additional updates to provide currently.

Social Awareness

Privacy-Preserving and Attack Resilient Deep Learning. We examined the potential risks of deep learning models under adversarial attacks. We designed a framework that seeks to effectively generate poisoning samples to attack both model accuracy and algorithmic fairness of fair machine learning models. We developed three online attacking methods, adversarial sampling, adversarial labeling, and adversarial feature modification. All three attacks effectively and efficiently produce poisoning samples via sampling, labeling, or modifying a fraction of training data in order to reduce the test accuracy. The attacking methods can flexibly adjust the attack's focus and accurately quantify the impact of each candidate point to both accuracy loss and fairness violation, thus producing effective poisoning samples. We conducted experiments on two real datasets and results demonstrated the effectiveness and efficiency of our attacking framework.

We studied privacy preserving mechanisms used for deep learning algorithms. The widely used mechanism is differentially private stochastic gradient descent (DPSGD) which applies gradient clipping and random noise addition in the training. However, the DPSGD mechanism may incur disparate impact, i.e., the accuracy of a model trained using DPSGD tends to decrease more on minority subgroups vs. the original, non-private model. We studied the inequality in utility loss due to differential privacy by comparing the changes in prediction accuracy with respect to. each group between the private model and the non-private model. We analyzed the cost of privacy with respect to each group and explained how the group sample size along with other factors is related to the privacy impact on group accuracy.

We examined the privacy, resilience, utility tradeoff of deep learning models and developed threat- and privacy-aware deep learning models. In particular, we developed a modified DPSGD algorithm, called DPSGD-F, to achieve differential privacy, equal costs of differential privacy, and good utility. DPSGD-F adaptively adjusts the contribution of samples in a group depending on the group clipping bias such that differential privacy has no disparate impact on group accuracy. We conducted experiments on real world datasets and evaluation results showed the effectiveness of our DPSGD-F algorithm on achieving equal costs of differential privacy with satisfactory utility.

We developed a novel adversarial adaptive defense (AAD) framework based on adaptive training such that the trained models adapt at test time to new attacks. This framework improved structures the training data into groups and each group represents one attack scenario. Different from empirical risk minimization that trains a single robust model or learns an invariant feature space, our AAD approach learns a context vector from features of each batch during training and incorporates the learned context vector into both prediction and detection models. Thus, AAD can adapt at test time to new adversarial attacks. We conducted comprehensive empirical evaluations with popular adversarial attacks and defense strategies on two real world datasets under different attack settings. Empirical results showed that AAD achieves both high prediction and detection accuracy and significantly outperforms baselines.

We also developed a framework that adopts the reweighing estimation approach for bias correction and the minimax robust estimation approach for achieving robustness on prediction accuracy. The developed framework is robust under distribution shift.

Socially Aware Crowdsourcing. We applied interval-valued labels (IVL) instead of binary-valued ones. Doing so, a worker may use a subinterval within [0, 1] to annotate an instance even when he/she is uncertain. We developed two algorithms, i.e., interval-valued majority voting (IMV) and preferred matching probability (IPMP), to derive inferences from interval-valued labels.

Our computational experiments evidence that the proposed interval-valued scheme enables the specification of uncertainties during input time. With interval specific statistic and probabilistic properties, both IMV and IPMP algorithms are able to computationally derive an inference with an above 50% probability of matching the ground truth. Moreover, the uncertainty index defined in this work quantitatively measures the overall uncertainty of collected IVLs. Furthermore, our computational experiments can produce better quality inferences with IVLs than without.

User-centric Data Sharing in Cyberspaces. We continued our exploration and development of techniques for identifying context aware sensitive information from unstructured data. We expect to document and disseminate our findings by the end of Year 2. We are researching multimodal deep learning techniques for detecting and removing sensitive information, discriminating and stigmatizing information from unstructured data. We expect to document and disseminate our findings by the end of Year 2 or the first part of Year 3.

Deep Learning for Preventing Cross Media Discrimination. We have developed a deep learning-based coded hate speech detection framework, called CODE, to detect hate speech by judging whether coded words like Google or Skittles are used in the coded hate speech or not. Findings have been disseminated to the research community through publications and presentations.

We have explored robust hate speech detection techniques by combining deep learning models with causal inference techniques, including a) causal invariant representation learning for texts under known attacking strategies on hate words, b) de-confounded hate speech detection with knowledge of hate words but no prior knowledge about attacking strategies, c) robust hate speech detection model for new hate words and attacking strategies, and d) causality-based multimodal hate speech detection for multimodal hate without using hate words. A proposal has been submitted based on the results.

We have conducted empirical analysis on several multimodal hate speech detection models. Specifically, we have evaluated the performance of the Facebook Hateful Meme Challenge baseline models on the three MMHS150K datasets which contain both image and text inputs. We trained the models using different baseline approaches including unimodal training, multimodal training with unimodal pretraining, and multimodal pretraining. We have evaluated metrics including the accuracy and the Area under the ROC Curve (AUROC). We concluded that the current multimodal training does not significantly outperform the unimodal training, indicating that there is a need to conduct further investigations.

Marketing Strategy Design with Fairness. We have conducted text mining and sentiment analysis of online reviews collected from Twitter to identify top product features and any forms of bias embedded in the advertising and marketing campaigns. We have conducted link prediction in identity network based on social network, intra-layer and inter-layer link information. Then, based on the link prediction, we can predict the number of nodes affected in the entire social network. We conducted comparison with theoretical approaches including independent cascades and liner threshold.

We quantified unfairness and analyzed its impact in the context of data-driven engineering design using the Adult Income dataset. First, we introduced a fairness-aware design concept. Subsequently, we introduced standard definitions and statistical measures of fairness to the engineering design research. Then, we used the outcomes from two supervised machine learning models, Logistic Regression and CatBoost classifiers, to conduct the disparate impact and fair-test analyses to quantify any unfairness present in the data and decision outcomes. Findings have been disseminated to the research community through publications and presentations.

Privacy-Preserving Analytics in Health and Genomics. We completed the planned Activity 1, document and disseminate the findings of literature research of privacy-preserving data analytics algorithms and software, and Activity 2, initiate investigation on mathematical optimization models. Findings have been disseminated to the research community through publications and presentations.

Cryptography-Assisted Secure and Privacy-Preserving Learning. We are developing a fully distributed learning scheme for a computation-constrained user where each distributed participant only handles a portion of the neural network (e.g., a subset of neurons) and a subset of data samples, and the complete data and model are only known to the user. Privacy is enhanced since each participant only has partial knowledge about the dataset and the trained model. Blockchain and smart contract techniques will be designed to facilitate the distributed communications and coordination among participants while preserving privacy. We expect that this work will be completed by the end of Year 2.

We are developing a privacy-preserving face recognition-based access control system. The scenario considered is face recognition-based access control to buildings, but we aim to allow the system to authenticate a user based on his/her face without revealing the face image or face features of the user to the system. We have designed a cryptography-based solution, and are currently designing a new applied cryptography construction that allows face recognition model training and testing (i.e., face-based authentication) in the ciphertext space. We expect that this work will be completed by the end of Year 2.

Feedback from External Advisory Board. The team has some updates related to the comments from the EAB last year.

Dr. Anna Zajicek is leading the effort of putting together a systematic literature review/survey paper "Big Data, Privacy, Cross-Media Discrimination: Systematic Literature Review (SLR) Research" based on the review of over 120 empirical social science studies addressing user privacy attitudes, including concerns, and behaviors in big data domains. The preliminary findings from this work were presented as a poster titled "User's Privacy Concerns, Attitudes, and Behaviors: A Systematic Review of Literature," at the virtual NSF DART Conference in September 2021. Another paper, titled "Big Data and User's Privacy Concerns, Attitudes and Behaviors at the Intersection of Age, Gender, and Race/Ethnicity: A Systematic Review of Literature has been accepted for presentation at the annual meetings of the American Sociological Association (August 2022). The social science team, including the GAs expanded their networks by connecting with Carter Buckner, a CSCE Ph.D. student, and Dr. Quinghua Li, associate professor of CSCE. We are currently collaborating on a student-led project at the intersection of public policy and user privacy concerns.

Dr. Xintao Wu has submitted for publication two survey papers on fairness. The first one, "The Causal Fairness Guide: Perspectives from Social and Formal Sciences," was recently accepted by Frontiers in Big Data – Data Mining and Management. The second one, "The Statistical Fairness Field Guide: Perspectives from Social and Formal Sciences," is under review by AI and Ethics.

We have regular meetings for SA researchers to discuss research activities within SA team and explore collaborations in DART project. We invite students to attend some of those meetings and also plan to schedule student-only meetings. We encourage SA researchers and students to collaborate more closely on SA projects. We have explored collaborations with industrial partners (e.g., Walmart Global People Analytics) on AI, data ethics and fairness. We plan to strengthen these collaborations via targeted paper and proposal submissions in coming years.

Social Media and Networks

Mining Cyber Augmentation data for collective opinions and their evolution. We have developed a prototype of the platform to be used for data collection from social science classes. We used the previous platform of cyber discourse to mine thousands of sentences, producing standard statistical datasets stored and ready for secondary use. The platform, dubbed Intelligent Cyber Argumentation System (ICAS), facilitates online discussion among students contributors and collects their opinions posted on specific social issues. ICAS is thus an experimental online deliberation platform designed to facilitate large-scale online discussion and debate while using integrated analytical models to autonomously summarize and analyze the discussions. It enables online discussions that are anonymous (users are assigned a randomly generated username at registration time), asynchronous (posts are always persistent and visible to users), and un-moderated (all posts are immediately visible to others after posting). ICAS is an issue-centric discussion platform, meaning that all of the discussions center around a topic issue. We developed a novel natural language processing algorithm to analyze discourse data collected by the platform and other existing data and will evaluate it in the remainder of year 2 and into year 3.

Socio-computational models for safer social media. We developed a taxonomy to characterize the online information environment (OIE). The taxonomy was revised based on various social media platforms, several cyber campaigns that were studied, and characteristics of platforms and information about actors. The taxonomy can be used to estimate effects of misinformation campaigns using epidemiological models such as susceptibility, number exposed, number infected, and number of skeptics (SEIZ); the theory of diffusion of innovation (DOI) addressing innovators, early adopters, early majority, late majority, and laggards; and characterizing phases of misinformation campaigns (accelerating or decelerating) using s-function characteristics. These approaches are theoretically grounded and so enhance model interpretability. The characterization approaches mentioned above have been published extensively in peer-reviewed and high-impact scientific forums.

We developed a socio-computational model to identify focal structures that leverages the theory of social network analysis and collective action using an operations research framework. The model was evaluated on data that was collected for conspiracy theory spreaders on YouTube and misinformation networks on Twitter. These datasets correspond to different application areas such as health (COVID-19), smart city infrastructure security, protests, and social movements.

Auto-annotation of multimedia data. Indexing and analysis of multimedia data on social platforms will form the basis for their subsequent use in the target applications as part of this project. In year 2, based on the characteristics we defined in year 1, we began devising indexing methods for image, video and the associated meta and text data in collaboration with the SM4 team. Once these methods are develop and tested, we'll move to integrate several indexing techniques to better learn from multimodal datasets in years 3 and 4. Learning objectives per Activity 1 were partially defined, focusing on the fundamentals of object and event detection. Effort towards Activity 1 is expected to continue into Year 2, according to the specific indexing and analysis approaches adopted. Our efforts have included recognition and verification of key landmarks and objects, content-based identification of major events such as hurricanes, floods, explosions, and more.

As part of Activity 1, 3 key applications of "smart" use of multimedia information sources have been identified, including the use of information quality aspects informing future smart data-based applications. We have focused on the reliability dimension, given the considerations and priorities for the application scenarios we are collaborating with SM4 team in disaster management scenarios. Although we have had no defined milestones in Activity 3, we have started preliminary work towards legal and ethical aspects and principles that is critical in future artificial intelligence applications involving human-generated multimedia data on social platforms.

Informing disaster response with social media. We selected key content types on social platforms (e.g., image and text) and chose Hurricane Harvey to represent a large-scale and geographically widespread event. We're also considering smaller events (like the 2019 Fort Smith, Arkansas flood) to represent a smaller scale and more localized event. For Hurricane Harvey, the team began assembling a disaster image dataset from online disaster image repositories and downloaded tweets and images from the Twitter API. The team is using content-based techniques to verify or complete the metadata associated with visual data, such as location, orientation, timestamps, and identification of major landmarks. Other data sources that will eventually be included in the disaster scenario testbeds include satellite imagery, stream gauge data, flood inundation models, and geographic topology. The team assembled some of these data from these sources for the Hurricane Harvey scenario. Year 3 efforts will focus on augmenting this initial dataset for Hurricane Harvey and adding data for a second disaster scenario.

One dimension of information quality is reliability. Several techniques have been devised to assess reliability of social data inputs. First, natural language processing (NLP) techniques devised by the data curation team allow us to use text contained in Twitter posts to clarify incomplete or uncertain information such as image location. Second, we applied content-based techniques to verify or complete the metadata associated with visual data, such as location, orientation, timestamps, and identification of major landmarks. Finally, multi-modal analysis and indexing can be used to develop or verify similar metadata to ensure the reliability and authenticity of the visual content. These efforts and their results will play a crucial role in effective, safe and secure use of social media data in disaster management that will eventually be tested in the scenarios that have been devised.

As we worked to develop techniques to automatically assess and localize disaster-related events, another part of the team addressed how to respond to those events. We selected two routing problems with application to disaster response. Orienteering problems will be used to model search and rescue applications, while Canadian Traveler problems will be used to model truckload delivery of critical supplies. We completed a literature review on disaster logistics problems and models but our planned journal article synthesizing the literature review with results of our qualitative data analysis (from another project) is not complete because the data analysis has been delayed. The qualitative analysis will conclude in year 3 and enabling its synthesis with the literature review. Activity 3, scheduled to conclude in year 3, is on track, as algorithms for two routing problems have been developed and validated. They will be tested on the identified disaster scenarios once construction of those datasets is complete. We have deployed a PostGIS database containing the data and will naje it available to the DART project working with the cyberinfrastructure team. It became clear during year 2 that before extract, transform, and load (ETL) could be developed, the team would need to identify a clear relationship between disaster event source data and the road network is necessary. This work is ongoing in year 2.

Feedback from External Advisory Board. We have no additional updates to provide currently.

Learning & Prediction

Statistical learning – Marked Temporal Point Processes (MTPP) enhancements via LSTM networks. The methodology designed to take advance of statistical learning methods to cast light into deep learning models was originally focused on the integration of MTPP and LTSM. However, we transitioned to a convolutional neural network (CNN) approach and delaying integration of MTPP until after the CNN is complete. The CNN approach is now completely implemented in Keras and we constructed a pipeline to use tabular data (civil infrastructure maintenance records in this case) to test the model. Our work on a CNN for image-based data was published in year 2 and presented at two conferences. This CNN work is serving as the foundation for the shift in methodology from MTTP/LSTM integration.

We focused testing on a public data sets of civil infrastructure datasets and decided to not curate healthcare IoT datasets. The dynamic tabular data in the civil infrastructure data is sufficient at this point and we are working with the cyberinfrastructure team to share this dataset across DART.

Deep learning – novel approaches. While the first goal focused on using statistical models to improve interpretability, in this goal we are working on a collection of novel approaches to deep learning that encourage interpretability.

Our collaboration with data curation and life cycle is an attempt to automate big data maintenance. In year 2, we developed a novel autoencoder method that improves substantially on the current state of the art. The new method addresses the problem of large-scale dataset management and labeling using unsupervised and self-supervised deep learning methods. These approaches tend to not scale well. Our method provides representation learning using small, labeled data to train a model in many specific problems. We introduce a new combination approach that operates somewhere between deterministic and probabilistic deep neural networks. That enables a powerful mechanism for reason knowledge representation in big data dataset. We are preparing a manuscript that describes the method and presents results on a tactical agility dataset. Members of our team also created a method that collects Tor network traffic data and classify websites in real time. This is important because it lays a foundation for detecting websites that disseminate illegal content. A low-dimensional feature using histogram entropy, crafted to support this work, proved to be successful in detecting a variety of malware datasets, including those active in Window OS, Android OS, and in some IoT systems. In addition, we conducted a study to acquire comparable results with small datasets in order to reduce the cost of training machine learning models on huge datasets.

Other members of this team performed a comparative analysis between a convolutional neural network (CNN), a residual neural network (RNN) and a Vision Transformer (VT) using a Ductal Carcinoma (breast cancer tissue images) dataset to assess their suitability for adoption. The VT model outperformed the CNN and RNN on different tasks achieving up to 93% accuracy. A paper on this work has been accepted for publication later this year.

We introduced a pre-input layer to a binary-decision-fusion neural network and train that layer for the out-of-distribution cases adjusting the feature values, without altering the original training of the network. A manuscript is be written and submitted for publication. Meanwhile, our investigation into high dimensionality issues in deep reinforcement learning ended with limited success. There was no clear improvement in clustering capabilities with the application of a group action to the data. In year 2, the team investigated injectivity issues in persistent homology, which happen when two distinct shapes have identical topological representation. This makes unusable as discriminating features in a neural network. We'll continue this work is year 3.

In year 1 we developed generalized solution approaches to non-analytical reasoning-assisted deep reinforcement learning (DRL). In year 2, we tested our approach on "Montezuma's Revenge," a notoriously difficult Atari game for a DRL agent. Our agent was able to perform exploration very quickly with the support of human intuitions in the form of heuristics. We have submitted a conference paper on this work.

Deep learning – efficiency and specification. Lightweight deep neural networks are required to deploy AI-based object understanding models on mobile devices. Such lightweight deep neural networks require efficient memory use, a relatively weight representation, and low-cost operators. Our research in year 2 introduced new "on-the-edge" deep learning algorithms that executed on low-cost platforms with high accuracy. We provided an algorithmic analysis of our methods which promise to eventually be competitive against large-scale deep networks while significant reducing computational time and memory consumption. We compared our methods with other current low-cost deep learning applications by evaluating their ability to label benchmark natural and medical image sets. Our lightweight model performed well enough to begin developing a manuscript for publication.

Harnessing transaction data through feature engineering. In our predictive modeling framework, we have developed new methods for incorporating time-dependent features with a new method for deriving variable importance. We have two manuscripts under revision and are expecting to resubmit by the end of April 2022. We have delivered an oral and a poster presentation at the 2021 INFORMS Annual Meeting and will also present at the 2022 IISE Annual Meeting. We are in the process of further improving our learning models, through constructing an optimization model to identify the optimal size of time windows for prediction. We are also working on representation learning for our deep learning models.

Feedback from External Advisory Board. We have no additional updates to provide currently.

Education

Education (ED) is a team represented by faculty at participating institutions as well as collaborators from other campuses across the state. The goal is to integrate the research from the other teams with industry needs to develop a statewide data science educational ecosystem, including technical certificates, associate's degrees, and bachelor's degrees in data science and data analytics at as many campuses in Arkansas as possible. Key to this effort is the cooperation with the Arkansas Department of Higher Education (ADHE) to allow fast-track approval of degree and certificate programs modeled after the University of Arkansas' bachelor of science in data science program that was implemented in 2019. Arkansas is a small but strong state, and we have undertaken what we believe to be the first effort nationwide to create consistent, modular, ubiquitous data science educational opportunities for all learners statewide.

Postsecondary Degree Programs. In Year 2, the degree programs reported in Year 1 have been expanded to include a 2+2 8-semester plan in the final review process as a "proof of concept" with North Arkansas College and UAF. UCA has developed the strategy and plan template for institutions migrating existing Concentrations or Minors to full B.S. Degrees. The team is also developing an MOU for institutions who want to opt-in to the DASC efforts to ensure full transferability statewide.

Other recent updates include that both UALR and the University of Arkansas Pulaski Technical College (community college) opted-in to implement the first two years of curriculum and serve as spokes to the established hubs. Southern Arkansas University transitioned into active participation as a 4-year hub with plans to submit a degree plan for approval in Year 3.

Middle School Coding Block. In Year 2, we continued conversations with the Coding AR Future group and 2 virtual meetings were held with educators to finalize the plan and start collecting established curricula. We plan to have a follow up meeting this fall and will report more in Year 3. We are still on track to complete the 9-week curriculum and pilot in schools by the end of the project.

Feedback from External Advisory Board. We have new activities to report on the relationship with DART and the Arkansas Center for Data Sciences. Bill Yoder, the executive director of ACDS, officially joined the DART Industry Advisory Board in early 2022, and monthly meetings have been established with the ACDS and DART ED Team. We are compiling resources and opportunities, and ACDS is positioned to help achieve related milestones for Year 2 and moving forward.

Partnerships and Collaborations

We have a number of partnerships and collaborations to report for Year 2. These collaborations and engagements helped identify challenges faced by industry and government, collaborate on solution development, and train our students for the 21st century workforce.

Government. DART researchers are working with the Arkansas Department of Health to develop methods for rapid analysis of wastewater for Covid-19 and other pathogenic outbreaks. DART is engaged with Arkansas Attorney General's office, US Department of Defense, US Senator John Boozman's office, and Australian Defense Science and Technology Organization. These collaborations/engagements helped us bridge the science, society, and policymaking through technological innovations. One new partnership of note is with the SM/LP/CI teams working with the US Department of Transportation's Maritime Administration to build a maritime shipping data hub that will provide tens of terabytes of data to DART. The data sets were chosen based on stakeholder engagements with the US Army Corps of Engineers and the US Geological Survey and will be immediately useful to the disaster response modeling.

Another new partnership is with Dr. Tina Moore of the Arkansas Department of Education. Dr. Moore was previously the math program manager at ADE, and recently was tasked with directing the Data Science and Computing Continuum to address some of the challenges facing the educational ecosystem. We are collaborating with her team on a number of events and target outreach, including K12 competencies and pathways, and building out the Arkansas Data Science & Computing Ecosystem.

Industry. We also have a number of industry collaborations. PiLog is helping DART by testing and suggesting new features for the data washing machine (DWM). This collaboration allows the DC Team to test the performance of the DWM against real-world data sets and the ability to benchmark the DWM against PiLog's fully supervised, rule-based clustering system. Year 2 saw our collaboration with industry partner SAIC rise to a new level with their support of a POC around positive data control (PDC) organization. We are also working with SAIC to help provide a mechanism for determine corporate hierarchies, branches, and subsidiaries in the USA.

Other new industry collaborations include TigerGraph Corporation and Zoetis, the animal health company through industrial grant, related to a Salmonella genome project to develop salmonella vaccines. Arkansas Blue Cross Blue Shield is providing data sets and capstone curricula for DART students, and training opportunities have been provided by NVIDIA and The Carpentries. Arvest provided a mentor for a DART I-Corps participant.

Other Partners. Another partnership is with the Arkansas Center for Data Sciences (ACDS), which was the nonprofit that resulted from the 2017 Governor's report on competitiveness in data analytics. The ACDS director, Bill Yoder, joined the DART IAB in Year 2 and we collaborate with them on a number of efforts regarding the education program and making new industry connections. We also cross-post and use each other's dissemination channels to widen our news reach.

In late 2020, PI Fowler began corresponding with Weston Waldo, the new venture development program manager in the UA Office of Technology Ventures. He most recently served as the program director of Texas Tech's NSF I-Corps Southwest Node and Site programs, and several tech investment funds. He is also an I-Corps adjunct instructor and an NSF SBIR/STTR commercialization reviewer. We have been collaborating to significantly increase the number of I-Corps participants from Arkansas with EPSCoR research lineage. As a direct result of our collaboration, we've received $250,000 in I-Corps awards to DART participants, and an additional $400,000 to participants from our last Track-1, CASE, and a Track-2 participant. Our IAB members are invited to serve as mentors for these teams and we work closely with Weston to form the teams and get proposals in. We look forward to reporting even more DART / NSF TIP accomplishments in the following years.

Workforce Development

Student Training

91 graduate students, 3 postdocs, and 31 undergraduates were involved in the project in some capacity since the last report. At least 40 graduate students and 15 undergraduates annually are hired through research assistantships, and the DART Summer Undergraduate Research Experience (SURE) program funds summer research for at least 10 underrepresented minority students annually in DART Labs. All of these positions will be filled by the end of Year 2. Additional information from the SURE program can be found in the Broadening Participation section of this report.

On April 3-4, the central office co-hosted an EPSCoR Workshop on Artificial Intelligence and No-Boundary Thinking to Foster Collaborations in Research, Education and Training. This event involved faculty researchers and students from Arkansas and across the country. Many DART faculty and students attended and participated in technical sessions with national expert presenters. The AI Campus student showcase took place during the workshop and three DART students presented their work in AI.

Arkansas Summer Research Institute

The 2021 ASRI took place virtually June 14-25. Leading up to the event, presenters and student participants attended a training session that covered the various technology platforms used, as well as our code of conduct and what to expect during the event. A prep session was also held to help students install RStudio Desktop or set up accounts on RStudio Cloud for the event. A schedule document was created in Google Slides (tinyurl.com/asri2021schedule), and a program was developed with abstracts for each session, presenter bios, and other relevant information (tinyurl.com/ASRI2021packet). A digital student binder was created with all the course resources and additional materials and information (tinyurl.com/ASRI2021binder). The event is no cost for students to attend, and presenters are all paid stipends for their participation. The event is gamified, so that students receive points on a leaderboard for participating in daily synchronous and asynchronous activities.

The interactive sessions included Introduction to Data Science, Conducting a Literature Review, How to Read a Research Paper, Setting up your Machine Learning Environment, Experimental Design & Statistics, Cybersecurity Applications, Capture the Flag on the Cyber Range, Identification of Novel Breast Cancer Biomarkers with cBioPortal, Phylogenetics, Giving a compelling research presentation, Getting Started with R, Working with Data in R, Spatial Data & Mapmaking in R, Graphing & Descriptive Stats in R, Data Extraction & Reproducing Research, Intro to Artificial Intelligence, Intro to Datasets and Data Formats, Data formats for biological sequences, Creating an effective LinkedIn profile, Building your Resume and CV, Where to find data, Finding research opportunities, Equity Inclusion & Representation in Research, Visualization Fundamentals in Python, RBioTools, Individual Consultations, Getting Started with Python, MCAT and GRE prep, Machine Learning for abnormal event detection, Fundamental design and UX for Science Communication, Research Ethics, Perseverance in research, Classification in Python, Debugging in Java, and Data Ethics. Panel discussions included Data Science in Today's World, Entrepreneurship in Research, and STEM Careers. A number of faculty also presented their 'research stories' which is intended to show students the diverse backgrounds and career paths of researchers. Each day had a break with virtual office hours and ended with daily evaluations and 'reflection and prep' assignments for the next day. Students are asked to use the hashtag #ArkansasSRI on social media platforms, which can be tracked on Twitter, Facebook, and LinkedIn. Communities and groups are Facebook and LinkedIn are maintained by the central office where alumni and presenters can remain connected and share updates.

At the end of the event, each student presented a research presentation using skills learned throughout the event. The 2021 ASRI was completed by 42 undergraduates. The gender distribution was virtually equal between Female (51%) Male (49%) participants. The ethnic distribution was over one-third (37%) Asian; one-fifth (20%) White; 17% Black; 9% Multi-ethnic and 6% Other. In addition to the undergraduate student participants, the ASRI 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; 13% from University of Arkansas for Medical Science; 11% from the Arkansas School for Mathematics, Sciences, and the Arts; 11% from Arkansas Tech University and 11% representing industry. The student evaluation data showed that 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. Our external evaluator Kirk Minnick conducted the annual ASRI program evaluation resulting in a report that is available to NSF upon request.

Plans are underway for the 2022 ASRI which will take place June 6-17 virtually. We are incorporating changes to the structure of the program based on feedback from the previous two years. At the time of this report, we have over 60 applicants. We are preparing a curated ASRI data repository that will be presented to the students before the event. On the first day, each student will choose a track (beginner or intermediate) and choose a dataset that seems interesting to them. We are pulling data from UCI, NCBI, Kaggle, CRAN, and other public sources. Several domains will be represented- business, physics, environment, plant science, public health, cancer, COVID, engineering, and more. The students will then use their chosen dataset in each of the interactive technical sessions and make one slide in their presentation each day to prepare for the symposium on the last day. We look forward to reporting more success in the Year 3 report.

K20 Educator Professional Development

Since the last report, we have the following update from the partnership with EAST Initiative. 28 teachers have participated in a 3D printing workshop, 21 attended the Pi-Top workshop, and 82 attended the leadership workshop. We do plan to invite an EAST lab to give a presentation at our upcoming annual meeting in May. The central office did host a booth at EAST Conference 2022 in March with the goal of interactively teaching data science basics. Students were offered a choice of four activities to complete and earn a badge ribbon, which are highly sought after during this conference (badge ribbons are like gold to these kids, they collect as many as possible and walk around showing them off to friends). Over 2500 K12 students and educators attended the conference, and we directly engaged around 200 students and 20 educators. The activities that we designed included: write your name in binary code, test your knowledge on charts and graphs, the sorting game, or the summing game. Worksheets were provided for each activity (included as appendices in this report).

Our booth was popular- the students loved our activities. Some kids wanted to complete just one activity to get the badge ribbon, and some wanted to complete all of the activities. The sorting game was the most popular- we had a bucket full of items from the dollar store like toy cars, foam dice, artificial flowers, toy animals, hair scrunchies, etc. that all shared different attributes. For example, there were similar colors represented across the items (blue, green, red, purple), and similar materials (plastic, metal, wood). The students were asked to grab items from the bucket and work together as a team to decide how to sort the items into four bins. Then we would ask the students to describe their thought process- did they sort the items based on color, item type, item material, size? We discussed attributes and features, and then asked the students how they might sort the items in other ways. We would then ask if they knew what an algorithm was, which most kids had heard of but didn't know the definition. We described it as- imagine that you have decided the best way to sort all these items into the bins, and you wrote down a very detailed set of instructions for the next group of students to tell them how to do it correctly- that would be an algorithm.

The summing game is a demonstration of information visualization. Students were asked to compete against each other and shown a list of the numbers 1 through 9. The instructions are to take turns calling out numbers, with no repeats, until one person reaches the sum of exactly 15. It can be difficult to simultaneously keep track of your own sum and your competitor's sum, and figure out how to prevent the other person from winning. After this round, the students would be invited to play it again, but as tic-tac-toe with a magic square, which for most people is much easier to do. We would then discuss the impact of how information is presented and how that can affect our perception and understanding of information.

The educators also really enjoyed the activities we had put together, and upon request we have decided to share all the materials we developed through the Arkansas Department of Education, and make them available on the project website.

Career Development Workshops (CDW)

During Year 2, DART offered five virtual Software Carpentry workshops with 195 total attendees (not entirely from DART). We also added one certified Carpentries trainer from UAF but intend to add a trainer from another institution in the coming months. As new trainers are added the number and variety of courses will grow. Participants from all of the DART teams conducted workshops, taught special lectures, and participated as presenters or coaches during the Arkansas Summer Research Institute and AI Campus. Workshop topics included computational tools for bioinformatics, comparative genomics, R-BioTools, GitHub, machine learning, and more.

In June of 2021, DART hosted a free workshop to more than 70 participants through the NVIDIA Deep Learning Institute. The workshop covered Fundamentals of Deep Learning 1 with certified instructor Dr. Milanova, a participant in the DC team. Participants completed interactive exercises in computer vision and natural language processing, and were provided certificates at the end of the workshop.

Additional workshops offered in 2021 since the last report include (a) Communicating Science to Legislators, and (b) Distilling Your Message, both with Dr. Jory Weintraub, who recently transitioned from Duke University to a full-time position with Advancing Research Impact in Society (ARIS). The recordings for both are publicly available on the @arepscor YouTube channel. Dr. Barb Bruno presented a workshop in July 2021 on individual development plans to approximately 20 attendees.

In 2022, the central office hosted a special webinar featuring program officer Dr. Wei-Shinn Ku from the III Cluster of the CISE/IIS Division, who is also a professor in the Department of Computer Science and Software Engineering, Auburn University, Auburn, Alabama. Dr. Ku presented on a number of NSF programs, including CAREER and CISE Core Programs. Another CDW was hosted on Designing & Evaluating Better Scientific Posters with Mike Morrison where students were provided training on incorporating user experience and social science research best practices in their scientific posters. An additional workshop is planned in May to go over NSF proposal development basics including the new Biosketch, Current & Pending Support Forms, and Collaborators and Other Affiliates template.

Faculty Training

During the summer of 2021, the first DART pedagogy workshop took place. This workshop was implemented to address the fact that learning python, and learning to teach python (as an example) are different things. One of the challenges in the education efforts is the lack of trained faculty at the HBCUs and PUIs that are both skilled in the new technologies the DASC program utilizes, and also the pedagogy behind them. The pedagogy workshop uses a train-the-trainer approach, where UARK faculty who teach the first year's DASC courses taught the pedagogy to the faculty in the first cohort who will be teaching those classes on their respective campuses. This summer, those first cohort faculty will teach the pedagogy for first year's courses to the second cohort faculty, and the UARK faculty who teach the second year's DASC courses will teach that pedagogy to the first cohort. This model will provide a sustainable pipeline of instructors statewide who are teaching the DASC courses. This will also provide a platform for iterative improvement, as the instructors from the various campuses will be able to provide their feedback and unique experiences after implementing the courses.

Communication & Dissemination

Our first communication goal is to make sure all the DART participants know what's going on and are aware of opportunities, project needs and accomplishments. This includes our daily communication, monthly team meetings, monthly webinars and SSC meetings, and annual project-wide meetings. The second main goal for communication and dissemination is to educate the public about what we are doing. This includes our project website, the campus communications committee, our reporting system ER Core, technical or professional dissemination through publications and presentations, and the science journalism challenge. In Year 1 we figured out our daily communication strategy which is a blend of email, Slack, Zoom, SharePoint, and text messages. Each DART team also meets monthly. We do have two large meetings planned each year, the annual conference or all-hands meeting for all DART participants and collaborators, and then a retreat for the graduate students and faculty. Email listservs have been established for the DART SSC; DART Project-wide Faculty and Staff; and DART students. A new listserv was created during Year 2 for the project administrative leadership team (Fowler, Ma, Hillyer, Cothren, Ford, and Minnick).

Weekly Digest Email. One new communications strategy implemented in Year 2 is the DART Weekly Digest email. PI Fowler compiles a list of upcoming events, current DART or related funding opportunities, recent DART accomplishments, and other project news. The digest is emailed out every Friday for the following week.

Virtual Office Hours. Another new strategy we've implemented since the last report is hosting weekly virtual office hours every Friday from 12PM – 1PM where a member of the project leadership team is present and any DART participant can drop in to ask questions or share updates. We've also hosted a few virtual student forums to meet and network with the DART students, and are looking at hosting some virtual happy hours to stay connected and help support our participants' mental health.

Monthly Seminars. All seminars/webinars have been recorded, and those that can be shared publicly are available on the DART website. All DART faculty, staff, and students are invited to the webinar series. Recent presentations included seed grant recipients, the Education team students, and faculty from LP.

Monthly team meetings. 11 meetings per research team completed (6 teams). During Year 2 and upon feedback from the DART Virtual Retreat, we have increased participation of team liaisons attending regular meetings of other teams. Hanna Ford implemented a DART Master Calendar, which participants can import to their Outlook or Google calendars, and an RSS feed. This has helped to make sure that everyone is aware of which and when meetings are taking place. The SSC is also still meeting monthly.

All Hands & Poster Competition. Due to the prolonged pandemic, we have yet to host a project-wide face-to-face meeting. We are planning to do so May 16-17 for our Year 2 All-Hands Conference. In 2021, we hosted the all-hands conference and poster competition virtually. Over 100 people attended. During the meeting, we used Zoom and a tool called Mural to remotely collaborate, which included an ice-breaker obstacle. The facilitator also led the group through a Science of Team Science workshop which included an eye-opening jargon audit. We identified 89 terms that the participants didn’t have established definitions or understanding of, ranging from specific machine learning techniques to other important terms (like EPSCoR). Leading up to the virtual poster competition, we held some workshops with Mike Morrison and other science communication professionals to help the students develop effective virtual posters. A Better Poster-Style/Twitter Poster template was given to the students, along with lots of examples and guidance. Students recorded themselves presenting their virtual posters (3-5 slides) and the submissions were judged asynchronously before the conference. Winners were announced during the event, and all of the submissions can be viewed in our digital poster hall on the project website.

Retreat. We had been hoping that by February of 2022 we could meet in person, but it didn’t work out that way. We hosted the first DART retreat virtually, again using Zoom and Mural. About 125 people attended including 80 DART participants. We revisited the jargon audit and established definitions for some important terms, and we had a panel discussion with some members of the industry advisory board. Our external evaluator produced a report for the event that is available to NSF upon request.

Project website. A new project website was launched during Year 2, and a handful of project staff have access to log in and publish updates. The website has a lot of project information on it and is well-organized. We have a dedicated GRA position to support the website and keep it up-to-date.

AEDC Blog. Blogs are posted at https://www.arkansasedc.com/news-events/arkansas-inc-blog. Recent blog posts included a summary of the Year 1 seed grant awardees, and a post about the final outcomes from our previous Track-1. We have planned another blog announcing the seed grant program targeted to novice investigators, and two guest blogs from DART advisory board members.

Social Media. We have already reached the 5-year goals for increase in followers on Twitter and Facebook. One of the seed grant projects started their own Twitter account to highlight the K12 outreach activities taking place under the project, and that has provided a good opportunity to connect with more K12 schools, educators, and students on social media. Also, DART participants are frequently making local and regional news and we are sure to repost, share, and highlight those accomplishments on our channels. Efforts have also been made to collect social media accounts and blogs of DART faculty for cross-posting. We plan to start utilizing the YouTube channel to offer short videos about DART-relevant content.

Campus Communications Committee. The committee has been established and has met once and communicated through email since that first meeting. Another meeting is planned for the summer.

ER Core Site published & accessible. The ER Core site was implemented in August of 2020 and participants have been onboarded through March 2021. As of the time of this report, 100% of known DART participants, paid and unpaid with the exception of advisory board members, have been provided accounts and attended a mandatory training session. We have regular training sessions throughout the year where any participant can join and refresh their memory on reporting requirements.

Scientific publications. The team did publish 80 peer-reviewed articles and juried conference papers that will be included in the publication list for Year 2 and reported in NSF PAR (with the exception of one publication that PAR would not accept, despite reaching out to the HelpDesk, because it is a military journal).

Science Journalism Challenge. This activity has been postponed, but we expect to be caught up later in 2022. In the interim, we have decided to support special awards at regional science fairs for data analysis and data visualization. We look forward to reporting on this in Year 3.

Broadening Participation

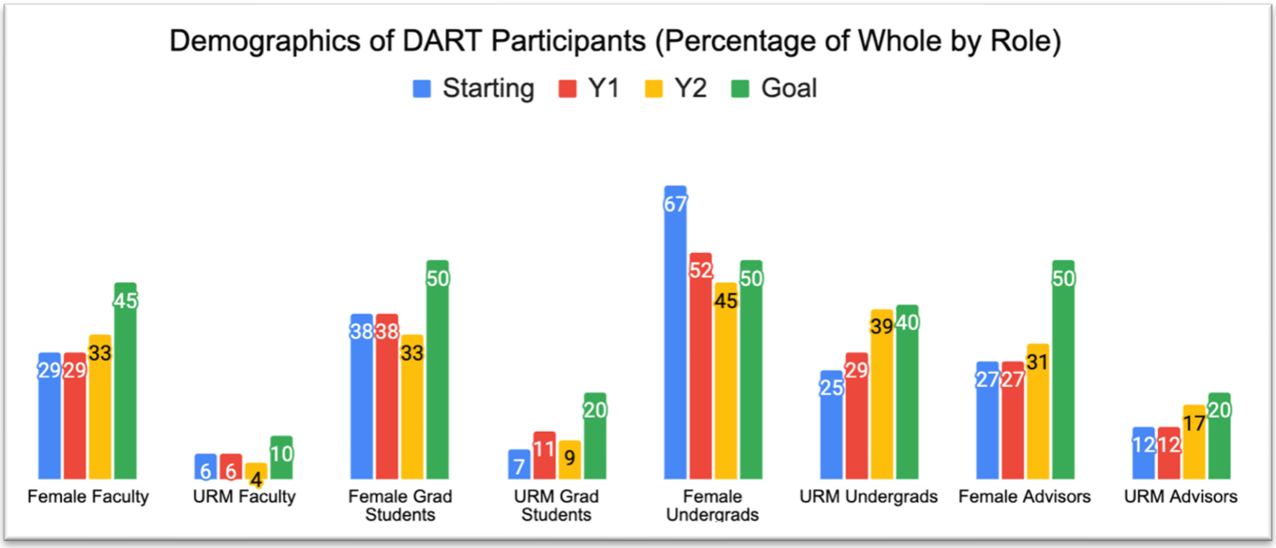
Participant Demographics. At the time of this report, DART has 205 participants and 29 confirmed advisory board members. Additional advisory board members are still being recruited. The numbers of participants and federally required demographics can be found in Table B included as an attachment to this report. Below we have described our participants in more inclusive and representative terms than Table B allows. The gender and ethnic diversity of the project will be a challenge that we actively work to improve continually. It is particularly difficult considering the disciplines involved in this project are among some of the least diverse of STEM disciplines and the lack especially of diverse faculty in those disciplines in Arkansas (computer science, data science, mathematics, etc.).

The starting diversity statistics below include participants who joined the project immediately upon award or were listed in the proposal and strategic plan. The Year 1 additional participants joined the project between October 2020 – March 2021. Participants joining since March 2021 are included as Year 2 participants. The table and chart below show the project's starting diversity and progress over the two reporting periods towards the goals outlined in the broadening participation plan*.

Category	Starting	Y1	Y2	Goal
Female Faculty	29%	29%	33%	45%

URM Faculty	6%	6%	4%	10%
Female Grad Students	38%	38%	33%	50%
URM Grad Students	7%	11%	9%	20%
Female Undergrads	67%	52%	45%	50%
URM Undergrads	25%	29%	39%	40%
Female Advisors	27%	27%	31%	50%
URM Advisors	12%	12%	17%	20%

** As self-reported by participants*



First Generation Students. 16% of the DART undergraduate students identified as first-generation college students, and 25% of the graduate students identified as first-generation college students. 52% of the graduate students are first-generation graduate students.

Disabilities. Five participants reported having disabilities.

Veterans. Two participants identified as US Veterans.

Mentorship Program. Mentorship program templates and guidelines were developed and distributed for the categories of participants outlined in the BP plan (SURE students, graduate students, early career faculty seed grant recipients and postdocs). We do not have any additional updates at this time, other than the feedback to the implementation was generally positive and we should have post-survey and evaluation data for the Year 3 report.

DART Research Seed Grant Program. Since we have a large number of starting faculty participants and no new-hires planned, the only way we can really broaden participation in the faculty group is through our seed grant program. The first solicitation round took place during the summer of 2021, and none of the applicants were URM, though several were female. Through the seed grant awards, we were able to bring additional female faculty into the project. The second round was posted on Monday 4/18 and included adjustments to make the process easier and more enticing for novice investigators. We plan to do some targeted outreach and professional development workshops leading up to the RFP deadline to broaden participation among applicants.

DART Summer Undergraduate Research Experiences (SURE) Program. In addition to the 15+ undergraduate research assistantships that are funded through the project, DART will fund summer undergraduate research experiences (SURE), for students belonging to groups that are underrepresented in computer science, information science, and data science related fields (as defined by NSF CISE). DART faculty will apply for funds to host these students for 8 weeks, with a limit of $8,000 per award. $80,000 annually has been budgeted for this program. Funds will support student stipends, housing, student-specific supplies, and in-state travel.

We were able to fund 12 undergraduates and 1 exceptionally talented high school student to participate in the SURE program last year. Most of them presented virtual posters in our poster competition last year. Most of the SURE students had little or no prior knowledge of Python programming and machine learning. They participated in training sessions, group discussions, and also completed modules from the DataCamp classroom. Most of these students also participated in the ASRI and some were offered UGRA positions in DART labs the following semester. One student from Arkansas Tech University - Star Douangchanh - participated in both the ASRI and SURE, then won 3rd place overall in the DART undergraduate poster competition at the Virtual Conference in September. A student from Philander Smith, Francis Oledibe participated in SURE, ASRI, and continued work as a DART undergraduate research assistant. He was accepted to the Jane Street FOCUS Fellowship and recently began an internship with Facebook.

The 2022 SURE Guidelines & Application was released to the DART community in March and will be accepting applications through early May. We look forward to supporting as many students as possible for summer research, and will report further on this in Year 3.

Broadening participation mini-grants. The education and broadening participation mini-grants are small awards up to $5,000 to increase or diversify the STEM pipeline in Arkansas. Eligible applicants are schools, school districts, STEM centers, educational service co-ops, non-profits, and other community organizations. We've awarded one round so far and are in the middle of processing the second round of applicants from year 1. The first award was to partially fund a STEM Saturday virtual field trip for underserved students in the North Little Rock community with the Arkansas Regional Innovation Hub, a local makerspace and non-profit. The second award was to fund 10 elementary and middle school teachers to participate in a professional development workshop on computational thinking with the Henderson State University STEM Center. In Year 2 we awarded 10 additional mini-grants, but most of them have been delayed due to the pandemic (they are mostly events and workshops). Therefore, we did not get a chance to invite an awardee to present at our conference last September, but plan to for the May meeting. We just conducted another solicitation round and plan to make awards in the coming weeks. Three of the projects were able to take place in recent months and we are expecting reports from those awardees soon, we will have more information on the specific results in the Year 3 report.

Special Conditions

Jurisdiction-Specific PTCs

The jurisdiction's programmatic terms and conditions have been met. The cyberinfrastructure documents and broadening participation plan were submitted and approved by NSF. Both were further reported on in their respective sections in this report.

Additional EAB Feedback

In the Year 1 EAB report, the board recommended that DART continue to explore effective ways to visualize and communicate our project's collaborative research activities. Our Year 1 review activities with the EAB took place remotely. The SSC put together some video presentations which we shared with the EAB along with our Year 1 Annual Report and other project documentation, then we had a synchronous Q&A session where the board could resolve any questions they had before compiling their report. We will conduct the EAB review for Year 2 in a similar fashion. Due to the continued pandemic, increase in travel prices, and an unexpected scheduling conflict with the NSF PI meeting this year, many of the EAB cannot attend our May 16-17 Conference.

The new website has been a useful tool to store and share our project videos and updates. Now that we have more project data to report, we have also worked on a number of visualizations to show the DART collaborative activities, some examples of which have been included as appendices to this report for reference. The goal is to have a trackable DART collaboration map, with participants as nodes and collaborations as edges. We will keep working on it. We plan to continue to expand our YouTube channel and website with additional video content from DART participants.

The EAB also expressed an interest in our technology transfer and sustainability efforts. We are happy to report that some DART participants plan to file patent disclosures during 2022, and we are developing a DART Commercialization Plan in collaboration with the university innovation and intellectual property offices, as well as our colleagues at AEDC that manage the TTAG, SBIR, and other relevant incentives and grants to tech entrepreneurs. We will share it with NSF in our Year 3 report. The research seed grant program is a key component to our ability to sustain meaningful and productive research collaborations beyond the life of DART. With the first round of awards issued and the second taking place later this year, we hope to report more progress on that in Year 3 as well.

Other updates related to the Year 1 EAB report have been included in the respective team sections of this report.

Expenditures and Unobligated Funds

Our reporting Table F shows that we will reach 76% obligated expenditures. The pandemic has affected our ability to expend funds in a number of ways- delays in hiring, ordering, supply chain, and travel/event restrictions. The issues we encountered related to the ScienceDMZ and federated identity have also slowed progress on our spending. Our calculations show that these impediments have snowballed into approximately $1.8M in unobligated funds by the end of Year 2. We have considered feedback from our advisors, project leadership, and other stakeholders, and have developed some new potential strategies for these funds. We will await the recommendations from the reverse site visit panel and the year 2 feedback from the EAB before making any decisions, but below are some of the ideas. We plan to further develop these strategies and associated costs for voting by the SSC over the summer.

Technology Transfer Support. The Central Office will establish and administer a program to support the commercialization of DART research. The program would provide training to students and faculty, with a focus on broad participation. We would have to take a close look at what types of costs are allowable and would be impactful to the participants. We know of two DART patents that are being filed soon, and would like to support these efforts and raise awareness of the commercialization pathway and incentives in the state.

Broadening Participation. Some additional support to broaden participation would be useful. We could increase the amount of SURE funding and provide additional support for UGRA positions for minority students during the school year. Currently, $80,000 is budgeted annually for the SURE program, and 15 annual UGRA positions are budgeted. We also could provide some funds to support the AI Campus program, a unique training program that was founded at Arkansas State University in 2018 and has since provided experiential project-based training in artificial intelligence and machine learning to diverse participants across the South. AI Campus is administered by the Center for No-Boundary Thinking, and even a $50,000 investment would significantly impact the number of participants and the quality of the curriculum.

Cyberinfrastructure Improvements. We had initially planned a Globus subscription which is $40,000 annually for research data management services for the whole project. After assessing our ability to effectively use such a service and considering the sustainability of the subscription, we delayed the purchase and focused on building a state-wide authorization and authentication model that would allow us to fully leverage the Globus research data management Standard level solution. Based on ongoing discussion with campus IT and research communities facilitated by SHARP CI, we agreed that InCommon Federation membership is the best solution. Globus Standard supports InCommon via CILogin at no additional charge. InCommon membership for all DART campuses address additional ARP access problems beyond data transfer and will make ARP far simpler to manage state-wide. InCommon provides single sign-on access to cloud and local services and not only enable access to ARP across the state but to global resources for participants. Costs vary by institution size, but we estimate a $50,000 annual cost state-wide. Notably, the two largest universities in the project are already members, and UA Fayetteville is pursuing a contract for professional management of an expanded InCommon suite of services and would not likely need additional support. Though these services would provide immense support to the project, the sustainability of the costly subscriptions is something that would need to be addressed formally with the participating campuses. We anticipate that three or more years of InCommon membership will demonstrate the utility of the service and provide time for institutions to develop plans to continue to support the membership. Those campuses which do not find value could end their membership.

Providing consistent, adequate with straining on shared resources like ARP and other national systems is a hard problem. We need to explore more effective methods like developing a training cadre at campuses across the state. Unobligated funds would go to salary and stipends for staff at those institutions who need it most.

Additional support for software development expertise – in particular, Hadoop development, programming in MPI, and relational, non-relational database development support, and project management using CI/CD tools – is in demand across research themes. This kind of expertise is hard to find or develop among graduate students and more experienced post-docs or research staff or needed. Various Centers at participating DART universities exist and could be supported with DART funding to support the project.

Internal and External Communication & Dissemination Support. One area of need that has already been identified is dedicated staff or contracted professional services to develop an ARP website, enhance the current DART project website, provide regular maintenance and updates of the current DART project website, disseminate project results throughout the project, interface with the campus communications committee, and develop training materials and modules for ARP. These activities could utilize anywhere from $100,000 to $200,000.

Tabular Representation of Progress to Date

Included as Appendix- Stoplight Tables

Appendices

Stoplight Tables for Year 2 (Appendix 1)
DART Year 2 Participant Demographics (Appendix 2)
Example DART Collaboration Visualizations (Appendix 3)
Worksheets Developed for EAST Conference (Appendix 4)

Tabular Representation of Progress to Date – Stoplight Tables (Appendix 1)

The following tables are copied from the revised DART Strategic Plan and show the progress against Year 1 and Year 2 milestones by project team.

Cyberinfrastructure – The Arkansas Research Platform (ARP)			
Goal 1.1 Establish the Arkansas Research Platform as a shared data science resource across the jurisdiction			
Objective 1.1.a: Establish the Arkansas Research Computing Collaborative (ARCC)	**Year 2**	**Responsible**	**Status**
Activity 1: Create ARCC advisory board with regional partners (GPN)	Establish roles and responsibilities consistent with MOU for the advisory board	Cothren, Prior, Springer, Deaton	Complete by end of Yr 2
Activity 2: Establish ARCC governance, operations and staff between UA and UAMS	Document defining organizational structure, roles, and responsibilities of ARCC	Cothren, Prior	Complete by end of Yr 2
Activity 4: Create UAF CI Plan to support DART (prior to 1.1.b and 1.1.c)	Publish a design and configuration document on DART website	Cothren, Prior, Springer	Complete by end of Yr 2
Objective 1.1.b: Upgrade cluster for data science research activity and integrate with existing resources	**Year 2**	**Responsible**	**Status**
Activity 1: Specify and purchase data science cluster based on document from 1.1.a	Receive and install new data science nodes on Pinnacle	Cothren, Chaffin	Complete by end of Yr 2
Activity 2: Test and deploy hardware elements for Pinnacle expansion for DART	Install, configure, and make available data science nodes on Pinnacle	Cothren, Chaffin	Done
Activity 3: Install and configure data science cluster to work with existing resources at UA, UAMS, UALR resources	Collect testbed specifications and software/platform needs	Cothren, Chaffin, Prior, Tarbox, Springer	Complete by end of Yr 2
Objective 1.1.c: Establish a science DMZ in Little Rock (UAMS, UALR) and high-speed connection with UAMS	**Year 2**	**Responsible**	**Status**
Activity 1: Specify and purchase 100Gb switch	Issue UAMS purchase order for 100 Gb switch	Prior, Tarbox, ARCC Tech, UAMS IT	Complete by end of Yr 2
Activity 3: Establish ScienceDMZ at UAMS	Specify and acquire additional DMZ components	Prior, Tarbox, ARCC Tech, UAMS IT	Complete by end of Yr 2
Objective 1.1.d: Establish a data and code sharing environment (GitLab and Globus)	**Year 2**	**Responsible**	**Status**

Activity 4: Engage other research themes to develop research-specific training modules in e.g. Python, R, Git, HPC, Singularity	-- Host 5 software carpentry workshops -- Train 2 software carpentry instructors	All theme co-leads	Completed by end of Yr 2
Activity 5: Develop and deploy training materials for code sharing, large data transfer protocols	Host 2 online ARP-specific training sessions	Cothren, Pummill, Prior, Tarbox	Done
Objective 1.1.e: Establish necessary controls to store and manage controlled unclassified, HIPAA-related, and proprietary information at UA and UAMS (other institutions if possible)	**Year 2**	**Responsible**	**Status**
Activity 2: Setup capacity for storing and managing CUI and HIPAA data at UAF	Deploy restricted access storage	Cothren, Prior, DuRousseau	Completed by end of Yr 2
Goal 1.2 Visualization for complex data in diverse data-analytics application domains			
Objective 1.2.a: Investigate state-of-the-art visualization solutions	**Year 2**	**Responsible**	**Status**
Activity 1: Investigate/define state-of-the-art visualization.	1 presentation or report	Springer, Conde, Post-Doc	Done
Objective 1.2.b: Define domain-specific integration of visualization solutions	**Year 2**	**Responsible**	**Status**
Activity 1: Develop and deploy visualization infrastructure software	Collect research theme needs	Springer, Conde, Milanova, Post-Doc	Completed by end of Yr 2
Objective 1.2.c: Introduce/integrate visualization for shared testbeds	**Year 2**	**Responsible**	**Status**
Activity 5: Engage other research themes to develop research-specific advanced visualization training	Host 1 online advanced visualization workshops		Completed by end of Yr 2

Complete	Will be completed by end of reporting year	Behind Schedule, Moved, or Deleted milestone	Updated milestone

Data Life Cycle and Curation	
Goal 2.1 (DC1)	**Automate heterogeneous data curation - The goal is to create unsupervised, automated processes taking as input unstandardized or heterogeneously standardized entity references, then perform unsupervised data quality assessment, data cleansing, data standardization, and data integration to create information products usable for business operations, data analytics, and research**

Objective 2.1.a: Automate Reference Clustering / Automate Data Quality Assessment
Objective 2.1.b: Automate Data Cleansing
Objective 2.1.c: Automate Data Integration

Objective 2.1.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Define metrics for data quality to measure impact of unsupervised data cleansing on data standardization and reference clustering	-- Define at least one metric for completeness, standardization, and clustering quality of unstandardized reference data -- Design and implement an unsupervised algorithm for each metric	Design and implement an algorithm using ML or Graph techniques for one metric	Wang, Talburt, Cothren, Tudoreanu, Xu, Yang, Liu, Rainwater
Activity 2: Set baseline data quality for initial test datasets used in prior research and acquire additional test datasets	-- Establish baseline quality using supervised methods for existing datasets -- Compare results of unsupervised quality metrics developed in Activity 1 to supervised results	Add 5 new person and 5 new business reference datasets for testing, at least 2 real-world -- Add 5 new product reference datasets	Talburt, Tudoreanu, Xu, Haitao, Liu Rainwater, Cothren
Activity 3: Curate test datasets and make available to other researchers	Establish a repository for the reference datasets and make available to other researchers	Add new reference datasets to the repository, as needed	Talburt, Tudoreanu, Liu Rainwater, Cothren
Activity 4: Develop a framework for collaborative data collection and cleansing for knowledge discovery	Formulate a hierarchical and as-needed data collection and cleansing strategy	-- Refine the formulation by including various practical constraints and test on small-scale problems -- Formulate a collaborative data collection strategy involving multiple teams	Haitao, Talburt, Wang, Liu Rainwater, Cothren
Activity 5: Develop a need- and prediction-based feedback mechanism for future data collection and making scalable decisions	-- Formulate a framework for sequential data collection on an as-needed basis -- Refine the formulation by including various practical constraints and test on small-scale problems	-- Formulate a Bayesian framework for sequential data collection based on predictive models -- Investigate analytical approaches for using large datasets for different levels of decision making	Haitao, Xu, Liu Rainwater, Cothren

Objective 2.1.b	**Specific Milestones**		
	Year 1	Year 2	Responsible

Activity 1: Improve the unsupervised frequency-based data cleansing method used in prior POC; Explore and test alternative methods and models for unsupervised data cleansing including ML, AI, and graph approaches	**-- Document and train team on data cleansing methods developed in prior research** **-- Design and implement in Python or Java improvements to the prior frequency-based approach**	**-- Design and test an ML or Graph implementation to the prior frequency-based approach** **-- Design and test new techniques for unsupervised data cleansing in Python, Java, ML, or Graph**	Talburt, Tudoreanu, Xu, Wang, Ussery, Liu Rainwater, Cothren
Activity 2: Migrate successful data cleansing models into a scalable processes		**Behind Schedule** **Refactor and migrate prior frequency-based approach into a scalable process**	Talburt, Tudoreanu, Xu, Haitao, Wang, Cothren, Ussery, Yang, Liu Rainwater
Objective 2.1.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Improve the unsupervised frequency-based data integration method used in prior POC and explore and test alternative methods and models for unsupervised data integration including ML, AI, and graph approaches	**-- Document and train team on reference clustering method developed in prior research** **-- Design and implement in Python or Java improvements to the prior frequency-based approach**	**-- Design and test an ML or Graph implementation to the prior frequency-based approach** **-- Design and test new techniques for unsupervised data cleansing in Python, Java, ML, or Graph**	Talburt, Tudoreanu, Xu, Haitao, Wang, Ussery, Yang, Liu Rainwater, Cothren
Activity 2: Migrate successful reference clustering models into a scalable HDFS processes		**Behind Schedule** **Refactor and migrate prior** frequency-based approach into a scalable HDFS process	Talburt, Tudoreanu, Xu, Haitao, Wang, Cothren, Ussery, Liu Rainwater
Goal 2.2 (DC2)	**Explore secure and private distributed data management**		
Objective 2.2: Build a POC and demo for Positive Data Control (PDC)			
Objective 2.2	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Build a POC and demonstration code for a Positive Data Control system layer forcing all of the tools read/write operations to synchronize with the platforms metadata tool		**-- Setup a test platform with at least one processing function (e.g. Hive), metadata function (e.g. Atlas), and security function (e.g. Ranger);** **-- Build POC with a simple PDC layer where Hive user is forced to go through PDC layer for all read/write operations**	Talburt, Wang, Tudoreanu, Pierce, Liu Rainwater
Goal 2.3 (D3)	**Harmonize multi-organizational and siloed data**		

Objective 2.3.a: Standardize pipelines for genome and proteome storage, retrieval, and visualization
Objective 2.3.b: Automate quality scores for biological sequence data
Objective 2.3.c: Apply machine learning methods to systems biology

Objective 2.3.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Define and download datasets to be curated	**Build genomics database, including quality scores, gene/protein annotation**	**Extend to proteomics database - all for fast characterization of proteins (links to SwissProt)**	Ussery, Byrum, Jun, Yang, Liu Rainwater
Activity 2: Optimize data storage and retrieval	**Use Elastic Cloud Storage for fast retrieval**	**Develop integrated database for proteomics & genomics, including annotations**	Ussery, Byrum, Zhan, Liu Rainwater, Tarbox
Activity 3: Develop visualization methods	**Prototype of R-BioTools for visualizing genomes**	**Publish one (1) R-BioTools paper for visualizing genomes**	Ussery, Zhan, Liu Rainwater
Objective 2.3.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Develop pan-genome and Pan-proteome databases	**Develop architecture / structure for rapid storage/retrieval of taxa-specific pan- and core-genomes**		Jun, Byrum, Liu Rainwater
Activity 2: Develop taxonomy links to downloaded genomes/proteomes	**Compare duplicate, known type strain genomes using ANI, Mash, 16S rRNA**	**Use Mash and other methods to assign nearest neighbors in phylogenetic space.**	Ussery, Liu Rainwater
Objective 2.3.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Define training sets to be used for ML	**Identify key datasets and problems for ML**	**Develop ML models for known toxins**	Byrum, Jun, Yang, Liu Rainwater
Activity 2: Integrate multi-omic models for ML	**Integrate genomic / microbiome / taxonomy datasets (petabytes)**	**Integrate genomic, transcriptomic, proteomic, and metabolomic datasets (petabytes)**	Ussery, Byrum, Jun, Yang, Liu Rainwater
Activity 3: Benchmark ML results		**Develop Benchmarking Standards for ML of pathogens**	Ussery, Byrum, Jun, Yang, Liu Rainwater

Complete	Will be completed by end of reporting year	Behind Schedule, Moved, or Deleted milestone	Updated milestone

Social Awareness			
Goal 3.1 (SA1)	**Privacy-Preserving and Attack Resilient Deep Learning**		
Objective 3.1.a: Identify potential vulnerabilities of deep learning algorithms Objective 3.1.b: Develop a universal threat- and privacy-aware deep learning framework Objective 3.1.c: Conduct comprehensive evaluations of the proposed framework and models			
Objective 3.1.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Research existing attacks including model inversion attacks and data poisoning attacks and capture mechanisms behind the threat models	**Document literature research of attack models and mechanisms behind attacks.**		Xintao Wu, Qinghua Li
Activity 2: Study the potential risks due to correlations among input data features, parameters, output, target victims, and latent feature space in deep learning algorithms	**Initiate theoretical investigation on the risks of deep learning algorithms**	**Disseminate the findings of both theoretical and empirical studies on risks of deep learning algorithms**	Xintao Wu, Qinghua Li
Activity 3: Study the sensitivity and impact of input data features, parameters, and the objective functions on the model output and identify appropriate differential privacy preserving mechanisms for different computational components in a variety of deep learning models	**Initiate theoretical investigation of privacy preserving mechanisms.**	**Disseminate the findings of both theoretical and empirical studies on privacy preserving mechanisms used for deep learning algorithms**	Xintao Wu
Objective 3.1.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Investigate the tradeoff of achieving privacy, resilience to adversarial attacks, and utility		**Research the tradeoff of privacy, resilience, and utility.**	Xintao Wu
Activity 2: Study the mechanisms of redistributing injected noise across input data features, model parameters, and coefficients of objective functions based on their vulnerability and impact on the model output		**Examine the noise redistribution mechanism.**	Xintao Wu
Activity 3: Develop and implant threat- and privacy-aware deep learning models		**Design algorithms of threat- and privacy-aware deep learning models**	Xintao Wu, Qinghua Li
Goal 3.2 (SA2)	**Socially Aware Crowdsourcing**		

Objective 3.2.a: Improve crowdsourcing data quality with considerations of uncertainty Objective 3.2.b: Enhance available inference and learning models with novel algorithms for improved effectiveness and efficiency Objective 3.2.c: Verify and validate the robustness and trustworthiness of information from crowdsourcing data			
Objective 3.2.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Allow uncertain labels in crowdsourcing data collection	**Selected the approaches through literature review**	**Implemented and tested**	Chenyi Hu, Ningning Wu, Xintao Wu
Activity 2: Aggregate raw labels after label collection	**Computational schemes are identified**	**Implemented and tested**	Chenyi Hu, Ningning Wu, Xintao Wu
Activity 3: Filter out possible noises to further improve data quality	**Identified possible sources of noises**	**Filtering algorithms designed**	Chenyi Hu, Ningning Wu, Xintao Wu
Objective 3.2.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Build theoretic foundations	**Specified mathematical requirements**		Chenyi Hu, Xintao Wu
Activity 2: Develop learning models and inference algorithms		**Algorithms designed to meet specification**	Chenyi Hu, Xintao Wu
Activity 3: Test and apply these learning models and algorithms		**Testing dataset selected**	Chenyi Hu, Ningning Wu
Goal 3.3 (SA3)	**User-centric Data Sharing in Cyberspaces**		
Objective 3.3.a: Investigate on personal identifying information and their privacy issues Objective 3.3.b: Investigate appropriate multimodal deep learning techniques to identify discriminative and stigmatizing information Objective 3.3.c: Develop a user-centric privacy monitoring and protection framework			
Objective 3.3.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Research state-of-art entity identification techniques for non-structure data	**Document and disseminate the findings on personal identifying information and their privacy issues**		Ningning Wu, Qinhua Li, Chenyi Hu
Activity 2: Investigate appropriate techniques for identifying context-aware sensitive information	**Identify and disseminate the findings on the sensitivity of information in different context**	**Identify appropriate techniques for identifying context aware sensitive information**	Ningning Wu, Qinhua Li, Chenyi Hu

Activity 3: Develop appropriate text analysis techniques to identify sensitive information from unstructured data	**Research appropriate text analysis techniques to identify sensitive information from unstructured data**	**Develop appropriate techniques for identifying sensitive information from unstructured data**	Ningning Wu, Qinhua Li, Chenyi Hu
Objective 3.3.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Research state-of-art multimodal deep learning techniques for identifying private sensitive information	**Study and document state of art multimodal deep learning techniques**	**Document and disseminate the findings on the determination of appropriate multimodal techniques for detecting sensitive information**	Ningning Wu, Xintao Wu, Qinghua Li, Chenyi Hu
Activity 2: Investigate appropriate techniques for identifying discriminating and stigmatizing information		**Document and disseminate the findings of state-of-art techniques for identifying discriminating information**	Ningning Wu, Xintao Wu, Qinghua Li, Chenyi Hu
Activity 3: Develop appropriate deep learning text analysis techniques to accurately remove discriminating and stigmatizing information		**-- Design deep learning techniques for removing discriminating and stigmatizing information** **-- Implement deep learning techniques for removing discriminating and stigmatizing information**	Ningning Wu, Xintao Wu, Qinghua Li, Chenyi Hu
Goal 3.4 (SA4)	**Deep Learning for Preventing Cross-Media Discrimination**		
Objective 3.4.a: Explore deep learning-based techniques to detect cross-media discrimination Objective 3.4.b: Design generative adversarial models to remove cross-media discrimination Objective 3.4.c: Develop a joint multi-modal deep learning framework to detect and prevent cross-media discrimination. Test and evaluate the proposed techniques and models with large-scale social media data			
Objective 3.4.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Use deep convolutional neural networks (CNN) to recognize discrimination-sensitive objects from images	**Initiate theoretical investigation on using CNN to recognize discriminatory objects**	Updated Milestone: Complete exploration and comparison of different multimodal hateful image-text detection models	Lu Zhang, Xintao Wu, Zhenghui Sha
Activity 2: Adopt long short-term memory (LSTM) network to model the text	**Initiate theoretical investigation on using LSTM to model discriminatory text**	**Complete design and implementation of the LSTM-based model**	Lu Zhang, Xintao Wu, Zhenghui Sha

Activity 3: Utilize bilinear model to capture the implicit relationship between the detected discrimination-related objects and the text		**Initiate the theoretical investigation on the implicit relationship between the detected discrimination-related objects and the text**	Lu Zhang, Xintao Wu, Zhenghui Sha
Objective 3.4.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 2: Test and evaluate the proposed techniques and models from available data sources in social networks like Facebook, Instagram, and Foursquare		**-- Complete social media data collection** **-- Compete evaluation of CNN and LSTM models**	Lu Zhang, Xintao Wu, Zhenghui Sha, Anna Zajicek
Goal 3.5 (SA5)	**Marketing Strategy Design with Fairness**		
Objective 3.5.a: Text mining and sentiment analysis of user-generated data from social media and consumer shopping records to extract customer-desired product features Objective 3.5.b: Network-based modeling of customer preference incorporating marketing parameters Objective 3.5.c: Design of marketing strategies with fairness consideration and validate the approach			
Objective 3.5.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Data collection from social media data, e.g., Amazon and Facebook, using vacuum product as the application context	**Document and disseminate the findings of literature research and evaluation of the target products for case study**	**Document and disseminate the findings of data collection from social media including both review data and production information**	Zhenghui Sha, Lu Zhang
Activity 2: Data collection and analysis of consumer panel data from Nielsen datasets at the Kilts Center for Marketing	**Document and disseminate the findings of processing the data from Nielsen datasets and extract the information needed (e.g., demographics, product market segment, etc.)**	**Document and disseminate the findings of the data analysis obtained from Nielsen dataset to complement the data from social media**	Zhenghui Sha
Activity 3: Collection and analysis of marketing cases and/or ads with unfairness and exclusions		**Document and disseminate the findings of the analysis of unfair marketing cases and extract the features/forms of biases**	Zhenghui Sha
Activity 4: Text mining and sentiment analysis of the collected data for the development of metrics of fairness in marketing, and the identification of customer-descried product features		**Document the text mining and sentiment analysis for the data collected from social media**	Zhenghui Sha, Lu Zhang

Objective 3.5.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Quantification and rating of bias and unfairness of marketing strategies and relate it to customer-desired product features	**Document and disseminate the findings of researching the existing methods for fairness quantification and adverting parameterization**	**Define and document the quantification methods for the features identified from the data collected**	Zhenghui Sha, Lu Zhang, Xintao Wu
Activity 2: Network-based approach for choice modeling by incorporating customer preferences and perception to marketing (e.g., price) (un)fairness		**Define and document the network-based model for choice prediction and demand forecasting**	Zhenghui Sha

Objective 3.5.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Parameterize marketing strategies and incentive design for improved advertisement with fairness consideration		**Document and disseminate the findings of the literature study on computational marketing and computational advertising**	Zhenghu Sha, Xintao Wu, Lu Zhang

Goal 3.6 (SA6)	**Privacy-Preserving Analytics in Health and Genomics**

Objective 3.6.a: Design and develop machine learning algorithms and software, and advanced security and privacy technologies, for privacy-preserving data analytics
Objective 3.6.b: Train, test and validate the models and algorithms with publicly available data and some controlled genomics and health data; develop innovative frameworks and practical privacy-preserving techniques
Objective 3.6.c: Test the algorithms and technologies to work with a wide range of data types and high-dimensional heterogeneous data sources; Develop and deploy bioinformatics workflows into the private cloud environment, the Arkansas Research Platform ARP

Objective 3.6.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Design and develop machine learning and deep learning algorithms and software for privacy-preserving data analytics; Data and Infrastructure request and preparation	**Document and disseminate the findings of literature research of privacy-preserving data analytics alights and software**		Huang, Li, M. Yang, Ussery, CI ARP team
Activity 2: Develop privacy-preserving analytics algorithms, which will be based on high-dimensional tensor mathematical optimization	**Initiate investigation on mathematical optimization models**		Huang, Li, M. Yang,

model and combinatorial models			
Activity 3: The optimization models will be incorporated with machine learning and deep convolutional neural network models		**Document and disseminate findings on theoretical investigation of privacy preserving algorithms**	Huang, Li, M. Yang, CI ARP team
Goal 3.7 (SA7)	**Cryptography-Assisted Secure and Privacy-Preserving Learning**		
Objective 3.7.a: Develop privacy-preserving federated learning methods through combining cryptography techniques and privacy models Objective 3.7.b: Explore how to protect the privacy of classification input data from the server hosting machine learning models Objective 3.7.c: Assess/Protect the trustworthiness of training data and machine learning models			
Objective 3.7.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Research the hybrid use of existing cryptography techniques and differential privacy in federated machine learning	**A survey of existing cryptography techniques and their applications in differentially private federated learning**		Qinghua Li; Xiuzhen Huang; Ningning Wu
Activity 2: Develop new applied cryptography techniques to use in combination with differential privacy for federated machine learning	**Design of preliminary new cryptography techniques used for differentially private federated learning**	**Design of blockchain-based private distributed learning**	Qinghua Li
Activity 3: Develop unified security models for theoretical analysis of hybrid solutions		Preliminary united security models for analysis of hybrid solutions	Qinghua Li; Xiuzhen Huang
Objective 3.7.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Develop methods for building/perturbing the model so that it can respond to encrypted or perturbed classification input		**Updated milestone:** Methods for building/perturbing model to support perturbed classification input	Qinghua Li

Complete	Will be completed by end of reporting year	Behind Schedule, Moved, or Deleted milestone	Updated milestone

Social Media and Networks

Goal 4.1 (SM1)	**Mining cyber argumentation data for collective opinions and their evolution**		
Objective 4.1.a: Develop a cyber discourse social network platform **Objective 4.1.b:** Collect data using the developed cyber discourse social network platform **Objective 4.1.c:** Develop natural language processing algorithms to analyze discourse data collected by the platform as well as existing data			
Objective 4.1.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Design the "hot button" questionnaire items	**Develop questionnaire for collecting discourse data**		Adams, Yang, Zhan
Activity 2: Develop Individual and network question measures and submit IRB consent form	**The question measures are determined and IRB protocol is approved**		Adams, Yang, Zhan
Objective 4.1.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Develop advanced natural language processing algorithms	**The advanced natural language processing algorithms are developed**		Zhan, Yang, Adams
Activity 2: Test the natural language processing algorithms using the existing data		**The algorithms are tested using existing data**	Yang, Zhan, Adams
Goal 4.2 (SM2)	**Socio-computational models for safer social media**		
Objective 4.2.a: Characterize online information environment (OIE) **Objective 4.2.b:** Develop socio-computational models to identify key actors and key groups of actors **Objective 4.2.c:** Study tactics, techniques, and procedures (TTPs) of deviant cyber campaigns **Objective 4.2.d:** Develop socio-computational models to measure power of a cyber campaign			
Objective 4.2.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Study social media spaces and cyber campaigns to identify characteristics and features	**-- Social media platforms C30dentified -- Cyber campaigns identified -- Characteristics and features identified**		Agarwal, Trudeau
Activity 2: Create a taxonomy of dimensions to characterize social media spaces	**Taxonomy developed**		Agarwal (Zhan)
Activity 3: Revisit and adjust taxonomy as social media space evolves		**Revised taxonomy developed and published based on new social media, campaigns, features, and characteristics**	Agarwal (Zhan, Milburn)
Objective 4.2.b	**Specific Milestones**		
	Year 1	Year 2	Responsible

Activity 1: Review cyber campaigns and social media data	**-- Data sources identified -- Data acquisition procedures established Database setup**	**-- Data reviewed and modifications incorporated -- Data collected and shared with DART teams**	Agarwal (Zhan, Dagtas, Milburn)
Activity 2: Identify behavioral traits for key actors and key groups by leveraging OIE characterization		**Key actors and key groups identified empirically**	Agarwal, Trudeau
Activity 3: Develop computational model(s) for key actor and key group discovery		**Model(s) developed**	Agarwal
Goal 4.3 (SM3)	**Auto-annotation of multimedia data**		
Objective 4.3.a: Develop multimedia indexing methods for social media data **Objective 4.3.b:** Design and implement deep learning methods for multimedia data **Objective 4.3.c:** Build Integrated smart applications based on unstructured multimedia data			
Objective 4.3.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Define priorities and characteristics for multimedia data on social platforms	**Key characteristics defined**		Dagtas, Trudeau
Objective 4.3.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Define learning objectives for social data from multimodal sources	**Identify and define three major learning objectives document**		Dagtas, GS
Objective 4.3.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Define key applications for the implementation and testing of the indexing and retrieval mechanisms	**Three key applications defined**		Dagtas, GS
Goal 4.4 (SM4)	**Informing disaster response with social media**		
Objective 4.4.a: Extract and index content describing transportation infrastructure status from social platforms **Objective 4.4.b:** Fuse data from social platforms describing transportation infrastructure status with other data sources **Objective 4.4.c:** Assess credibility of data inputs from Objectives 4.4.a and 4.4.b **Objective 4.4.d:** Develop routing algorithms that use inputs from Objectives 4.4.a-4.4.c to support routing for disaster response			
Objective 4.4.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Study social platforms to identify types of content that describe transportation infrastructure status (integrates with Objective 4.3.a)	**Identify and define social platform content types of interest (e.g., image, video, text, etc.)**		Milburn, Dagtas

Activity 2: Develop and implement extraction techniques for identified types of social platform content (integrates with Objective 4.3.a)		**Develop social platform extraction techniques for content types of interest and pilot test on at least two disaster scenarios**	Dagtas
Activity 3: Develop and implement indexing techniques for extracted social platform content (integrates with Objective 4.3.b)		**Develop and implement indexing techniques for extracted social platform content and pilot test on at least two disaster scenarios**	Dagtas
Objective 4.4.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Identify other data sources that contain real-time information regarding transportation infrastructure status	**Identify and define content types of interest (e.g., satellite imagery, traffic cameras) from sources other than social platforms**		Milburn, Cothren
Activity 2: Obtain and index transportation infrastructure data from other data sources		**Obtain and index identified content types for at least two disaster scenarios**	Liao, Milburn, Cothren
Objective 4.4.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Develop and implement machine learning classifiers to detect quality of information	**Obtain testing data from social platforms**	**Develop machine learning classifiers to detect false or low-quality information**	Zhan, Liao
Objective 4.4.d	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Identify critical routing problems with application in disaster response	**Select at least two disaster response routing problem variants using Milburn's existing qualitative interview data**		Milburn
Activity 2: Develop models of identified disaster response routing problems and assess state of the literature	**Conduct literature review for identified routing problem variants and publish journal article synthesizing review with qualitative data from 4.4.c.1**		Milburn
Activity 3: Develop and implement routing algorithms for identified routing problem variants		**For at least two routing problem variants, develop, validate and test at least one solution algorithm each on randomly generated test networks**	Milburn

Activity 4: Implement GIS testbed capable of displaying and analyzing real-time road status and routing algorithm outputs	**Define GIS system requirements**	**Develop GIS system to display real-time road status inputs**	Milburn, Cothren
Activity 5: Demonstrate models and solution approaches via pilot study of one or more disaster scenarios		**Select one or more disaster scenarios for pilot**	Milburn, Cothren, Dagtas, DC lead, LP lead

Learning and Prediction

Goal 5.1 (LP1)	**Statistical Learning – Random Forests for Recurrent Event Analytics**		
Objective 5.1.a: Create the Random Forests for Recurrent Event Analytics, which integrates the RF algorithm with classical statistical methods allows dynamic feature information to be incorporated into a tree-based method **Objective 5.1.b**: Create the Gradient Boosting method for Recurrent Event Analytics, which integrates the boost trees with classical statistical methods allows dynamic feature information **Objective 5.1.c:** Perform comparison study between the methodologies above and identify future research directions			
Objective 5.1.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Establish a preliminary model, and complete the theoretical investigation	**Complete the preliminary theoretical investigation on the proposed modeling approach**		Liu, Chimka
Activity 2: Complete the coding and numerical examples; write, submit, revise paper		**Complete the numerical studies, and submit a research paper**	Liu, Chimka
Goal 5.2 (LP2)	**Statistical Learning – Marked Temporal Point Process Enhancements via Long Short-Term Memory Networks**		
Objective 5.2.a: Develop methodology integrating the marked temporal point process (MTPP) with long short-term memory networks (LSTM) **Objective 5.2.b:** Create scalable implementation of MTTP/LSTM approach applicable to real-world data analysis scenario **Objective 5.2.c:** Evaluate and assess MTTP/LSTM approach on real-world discrete data sets			
Objective 5.2.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Formally define approach integrating intensity function of MTTP into LTSM	**Submit conference paper with initial model**		Rainwater, Liu
Activity 2: Establish proof-of-concept implementation of MTTP/LTSM approach	**Present conference paper with preliminary results of implementation**	**Submit journal article with conceptual findings and initial implementation of approach**	Rainwater

Activity 3: Perform benchmark of MTTP/LSTM tests on small simulated data sets		**Publish white paper and GitHub repository with benchmark tests/results**	Rainwater, Liu
Objective 5.2.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Assess data collected from Activities 1 and 2 of Objective 5.2c to define methodology computation performance requirements		**Produce system requirements document for V2 implementation**	Rainwater, Liu
Activity 2: Create version 2 implementation of approach using lessons learned from Activity 2 of Objective 5.2a		**Submitted conference paper**	Rainwater, Liu
Objective 5.2.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Acquire healthcare IoT datasets	**Publish curated data to GitHub**		Rainwater
Activity 2: Acquire civil infrastructure datasets	**Publish curated data to GitHub**		Rainwater
Activity 3: Establish baseline performance of predictions made by existing approaches applied to datasets from Objective 5.2c		**Publish white paper and GitHub repository with benchmark predictions**	Rainwater, Liu
Goal 5.3 (LP3)	**Deep Learning – Novel Approaches**		
Objective 5.3.a: Extract explanatory features from Deep Network **Objective 5.3.b:** Address high dimensionality issues in Deep Reinforcement Learning (DRL) using algebraic and topological methods **Objective 5.3.c:** Designing a novel rewarding model, and addressing interpretability issues in DRL			
Objective 5.3.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Development of novel self-supervised and flow-based deep learning approaches	**Development of the first unsupervised convolutional area and the first flow-based deep learning approach**	**Adaptation of the novel methods for application to the tactical agility dataset**	Celebi, Kursun, Luu, Kim, Karim
Activity 2: Developing a library of classifiers for benchmarking	**Development of linear dimensionality reduction methods**	**Development of the standard autoencoder method**	Celebi, Kursun, Luu, Kim, Karim
Activity 3: Application of the developed methods on real-world datasets	**Application of the developed methods with applications on natural images and textures, and classification of a malware dataset**	**Comparisons of the developed methods/libraries on the tactical agility dataset and malware dataset**	Celebi, Kursun, Luu, Kim, Karim
Objective 5.3.b	**Specific Milestones**		

	Year 1	Year 2	Responsible
Activity 1: Investigate group theoretical and topological properties of generalized neural network architectures	**Investigate group theoretical approaches to generalized NN architecture design within the context of interpretability for Objectives 5.3a (Activity 1) and 5.3c (Activity 1)**	**Exploration of internal topologies in generalized NN structures using TDA and PH to identify architectures which address high dimensionality and enhance the developments in Objectives 5.3**	Schrader, Cheng, Luu, Kim, Kursun, Karim
Objective 5.3.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Design an improved reward process for DRL	**Development of a generalized model of reward function in DRL addressing the issues with both sparse and dense feedback**	**Development of use cases of the developed reward model**	Rami, Karim
Goal 5.4 (LP4)	**Deep Learning – Efficiency and Specification**		
Objective 5.4.a: Create Novel Deep Learning Networks Executable with Reduced Computational Resources and Assess Performance **Objective 5.4.b:** Address Low-cost Deep Learning Algorithmic Analysis and Challenges **Objective 5.4.c**: Explore Low-cost Deep Learning Applications in Natural Images and Medical Images			
Objective 5.4.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Develop and demonstrate new low-cost deep neural network algorithms.	**-- Develop Teacher - Student Distillation Deep Learning Algorithms** **-- Develop Light-weight Deep Learning Algorithms**	**- Develop Deep Network Compression Algorithms** **-- Develop Deep Network Pruning Algorithms**	Khoa Luu, Ngan Le
Objective 5.4.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Mathematically analyze the proposed deep learning methods	**Develop analytic approaches to the proposed methods in Activities 1.1**	**Develop analytic approaches to the proposed methods in Activities 1.2**	Khoa Luu, Ngan Le
Objective 5.4.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: The developed deep learning algorithms will be optimized and implemented in two applications, including natural images and medical imaging.	**Develop Low-cost Deep Learning Approaches in Image Classification**	**Develop Low-cost Deep Learning Approaches in MRI Segmentation**	Khoa Luu, Ngan Le
Goal 5.5 (LP5)	**Harnessing Transaction Data through Feature Engineering**		
Objective 5.5.a: Design advanced feature engineering techniques for high-dimensional temporal data **Objective 5.5.b:** Create an improved prediction and decision-making framework incorporating feature			

engineering with health transaction data
Objective 5.5.c: Employ and validate the new framework for prediction and decision making with business transaction data

Objective 5.5.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Extract and process APCD data	**Obtain and prepare cleaned data for research**		Zhang, Nachtmann
Activity 2: Extract and engineer features from the high-dimensional temporal data	**Acquire features that are highly representative**		Zhang, Nachtmann
Activity 3: Explore and test automation of feature engineering in transaction data		**Achieve automotive feature engineering**	Zhang, Nachtmann
Objective 5.5.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Develop deep learning prediction models and algorithms with feature engineering	**Complete selection and testing of deep learning models**	**Improve the predictive models**	Zhang, Nachtmann
Activity 2: Incorporate representation learning in prediction with engineered features		**Implement and test autoencoders**	Zhang, Nachtmann

Complete	Will be completed by end of reporting year	Behind Schedule, Moved, or Deleted milestone	Updated milestone

Education

Goal 6.1 (ED1)	**Developing a combination of model programs, degrees, pedagogy, and curriculum including a 9-week middle school coding block; a technical certificate, certificate of proficiency, and associate of science in data science; and a Bachelor of Science in data science with minors or concentrations.**

Objective 6.1.a: Create a full 9-week curriculum for the middle school coding block to help struggling K12 teachers meet state coding requirement and provide rich training to K12 students
Objective 6.1.b: Create a set of postsecondary programs of core courses with options for electives for a consistent set of Data Science Undergraduate degrees (e.g., Assoc. Degrees, 2+2 and "2, then 2"), Concentrations, and certificates

Objective 6.1.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Hold a two-day workshop to include K20 computer science educators to outline the curriculum and pedagogy and establish the project timeline, roles, and deliverables	**Workshop completed, plan finalized and disseminated to stakeholders**		Fowler

Activity 2: Develop curriculum		**Curriculum 50% complete**	Fowler

Objective 6.1.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Create the 5-year Plan to meet the Objective	**Plan disseminated to stakeholders**	**Review 5-yr plan & update as needed**	Addison, Schubert
Activity 2: Identify the level of involvement and timing by academic institutions within the State	**Cohorts identified, all collaborators assigned**	**Begin Cohort 1**	Addison, Schubert
Activity 3: Review UA-Fayetteville and UCA Data Science Programs with the Teams	**1 meeting complete**		Addison, Schubert
Activity 4: Convene workshops annually of engaged academic and government institutions to establish baseline	**3 Workshops completed**	**3 Workshops completed**	Addison, Schubert
Activity 5: Define Data Science Objectives and Outcomes base for defined degrees and certificates	**Info disseminated to stakeholders**		Addison, Schubert
Activity 6: Define Data Science Courses Objectives, Learning Outcomes, and applicability to the defined degrees and certificates		**Info disseminated to stakeholders**	Addison, Schubert
Activity 7: Dissemination of developed program details with collaborating institutions, government, and industry partners	**Info disseminated to stakeholders**	**Updated Info disseminated to stakeholders**	Addison, Schubert
Activity 8: Ensure defined programs are in line with appropriate accrediting bodies	**Identify "Wave 1" of accreditation candidates**		Addison, Schubert
Activity 9: Prepare and submit program proposals of each type at each level for appropriate approval	**Begin "Cohort 1" Proposal Preparation**	**Submit "Cohort 1" Proposals**	Addison, Schubert
Activity 14: Create and maintain clearing house for course materials	**Create shared resources with UAF UCA existing materials and establish cataloging methodology**	**Add Cohort 1 developed materials**	Addison, Schubert
Activity 15: Connect students, courses, problems, data, etc., with the Research Themes	**Identify "Opt-In" Research Theme Researchers & Collaboration Types & Timing**	**Group 1 Collaboration**	Addison, Schubert

Workforce Development and Broadening Participation			
Goal 7.1 (WD1)	**Provide K20 teacher and faculty opportunities for professional development spanning multiple disciplines**		
Objective 7.3.a: Summer Undergraduate Research Experiences for underserved students: Fund summer undergraduate research experiences (URE), for underserved students **Objective 7.3.b:** Scholarships for underserved students to the Arkansas Summer Research Institute (ASRI) **Objective 7.3.c:** Connecting students to opportunities through the Arkansas Center for Data Sciences (ACDS)			
Objective 7.1.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Host one training session annually, issue technology kits to teacher participants		**1 Training session complete, kits issued**	Fowler
Activity 2: Host two support/training webinars annually		**Two webinars completed**	Fowler
Activity 4: Participate (booth or breakout) in EAST Initiative annual conference	**1 Conference completed**	**1 Conference completed**	Fowler
Activity 5: Invite participating teachers to annual All-Hands		**1 Teacher presentation at All Hands complete**	Fowler
Objective 7.1.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Fund seed mini-grants annually at $5,000 each	**10 seed grants awarded**	**10 seed grants awarded**	Fowler
Activity 2: Recipients attend Annual All-Hands	**Canceled- 1 awardee presentation at All Hands complete**	**1 awardee presentation at All Hands complete**	Fowler
Objective 7.1.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Identify faculty to be trained and assign to 5 cohorts	**Cohort 1 established**	**Cohort 2 established**	
Activity 2: Fund faculty annually at $5000 each for training	**10 faculty trained**	**10 faculty trained**	Fowler
Objective 7.1.d	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Host annual workshops on a variety of grantsmanship and entrepreneurship topics	**3 Workshops completed**	**3 Workshops completed**	Fowler
Goal 7.2 (WD2)	**Provide educational training opportunities inside and outside the classroom for students**		
Objective 7.2.a: Student Support at Participating Institutions: Support undergraduate research assistants during the fall and spring semesters at each participating primarily undergraduate institution, and graduate research assistantships at the academic research institutions **Objective 7.2.b:** Summer Internships: Facilitate industry internships for student participants at companies			

in relevant sectors and research centers
Objective 7.2.c: Connect with other research thrusts to develop relevant research-based capstone projects
Objective 7.2.d: ASRI-intensive data science and computing summer camps for undergraduates

Objective 7.2.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Provide undergraduate research assistantships annually	**15 UG supported**	**15 UG supported**	Fowler
Activity 2: Provide graduate research assistantships annually	**40 GA supported**	**40 GA supported**	Fowler
Objective 7.2.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Identify internship opportunities for students at relevant companies		**5 internships placed**	Schubert, Addison
Activity 2: Follow up with hosting companies for feedback and evaluation		**Develop intern and hosting company feedback and evaluation methodologies and instruments**	Schubert, Addison
Objective 7.2.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Develop capstone projects annually	Delayed- **5+ capstones identified**	**5+ capstones identified**	Schubert, Addison
Activity 2: Disseminate projects to all collaborating institutions	Delayed- **3+ capstones published**	**3+ capstones published**	Fowler
Activity 3: Invite capstone students to present at annual All-Hands		**1 student presentation at All-Hands complete**	Fowler
Objective 7.2.d	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Host ASRI Annually and invite all DART undergrads	**1 ASRI Complete**	**1 ASRI Complete**	Fowler
Activity 2: Evaluate and revise programming based on student and presenter feedback		**Evaluation report disseminated to stakeholders**	Fowler
Goal 7.3 (WD3)	**Ensuring broad participation to impact the pipeline of data science skilled workers**		

Objective 7.3.a: Summer Undergraduate Research Experiences for underserved students: Fund summer undergraduate research experiences (URE), for underserved students
Objective 7.3.b: Scholarships for underserved students to the Arkansas Summer Research Institute (ASRI)

Objective 7.3.c: Connecting students to opportunities through the Arkansas Center for Data Sciences (ACDS)			
Objective 7.3.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Provide summer UREs to URM students annually	**10 UG supported- 13 were funded**	**10 UG supported**	Fowler
Activity 2: Students participate in annual All-Hands meeting poster competition		**1 poster competition complete**	Fowler
Objective 7.3.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Provide scholarships/recruit students annually	**20+ scholarships provided**	**20+ scholarships provided**	Fowler
Objective 7.3.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Co-host statewide workshops on DS topics	**1+ workshop completed**	**1+ workshop completed**	Schubert, Addison
Activity 2: Collaborate on DS apprenticeship programs- recruiting partners and developing curriculum	**Develop apprentice and hosting company feedback and evaluation methodologies and instruments**	Year 1 feedback and evaluation; iterative improvement for Year 2	Schubert, Addison

Complete	Will be completed by end of reporting year	Behind Schedule, Moved, or Deleted milestone	Updated milestone

Communication & Dissemination			
Goal 8.1	**Maintain interproject communication to accomplish milestones and relay updates**		
Objective 8.1.a: Day to Day Communication: Daily project-related communication will take place mostly via email and GitLab. If during the first year the project faces challenges with these two platforms, other platforms like Slack will be explored **Objective 8.1.b:** Monthly Webinars & Component Meetings: Monthly webinar meetings will be held to share updates, events, and other news between faculty, students, administrators, and industry partners. Components will also meet monthly to manage project milestones and activities. **Objective 8.1.c:** Face-to-Face Meetings: Two project-wide face-to-face meetings per year will be hosted. The Annual All-hands Meeting and Poster Competition will be attended by all project faculty, students, industry partners, administrative committee members, evaluators, and external advisory board members. The Annual Retreat will be for faculty and graduate student participants. These meetings will facilitate team building and foster a sense of collaboration among the group.			
Objective 8.1.a	**Specific Milestones**		
	Year 1	Year 2	Responsible

Activity 1: Establish platform to maintain daily communication	**Platform established and participants onboarded**		Fowler, Cothren
Objective 8.1.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Host DART topical webinars monthly	**11 webinars complete**	**11 webinars complete**	Fowler, Cothren
Activity 2: Host monthly component team meetings	**11 meetings per component (6 components)**	**11 meetings per component (6 components)**	Co-Leads
Objective 8.1.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Host annual All-Hands meeting & poster competition	**1 All Hands & Poster Competition complete**	**1 All Hands & Poster Competition complete**	Fowler
Activity 2: Host annual retreat for faculty and grad students		**1 Retreat completed**	Fowler
Goal 8.2	**Educate the public about DART accomplishments**		

Objective 8.2.a: Maintain public-facing communication outlets to inform public about DART
Objective 8.2.b: Campus Communications: A project-wide communications team composed of communications staff from each participating institution, including AEDC, will be created. The communications team will use uniform citations and branding for all project-related releases. AEDC will issue press releases and blog posts related to overall project success, special events, and seed grant opportunities. The communications team will work together to release other pertinent information like new grant awards, patents, publications, and other highlights from each campus.
Objective 8.2.c: Project Data: Project data will be submitted by participants to the project's internal reporting system, ER Core. Mandatory NSF reporting data will be collected, as well as additional information like startup companies and other major accomplishments.
Objective 8.2.d: Technical dissemination channels: Project faculty will submit journal articles to scientific publications associated with data science and computing. Funds for travel stipends will be reserved to send students and faculty to national meetings related to data science and computer science research and education. Impacts and significant findings of research activities will be presented at these meetings. Relevant meetings include national meetings for professional societies and industry meetings.
Objective 8.2.e: Science Journalism Challenge: Beginning in year 2, the project will host a statewide science reporting challenge. Journalism students (undergraduate and graduate level) from Arkansas institutions may apply to win first, second, or third place awards. Applicants will be paired with a project participant student and faculty member to develop a story about relevant research. Submissions will be evaluated by a committee of Arkansas public relations professionals, including representatives from major advertising agencies and news outlets. The first-place award will include publication in a major Arkansas news outlet. Award recipients will also be invited to present their stories at the annual all-hands meeting.

Objective 8.2.a	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Establish project website	**Project website published**		Ford
Activity 2: Maintain project website and refresh content at least quarterly		**Content posted**	Ford

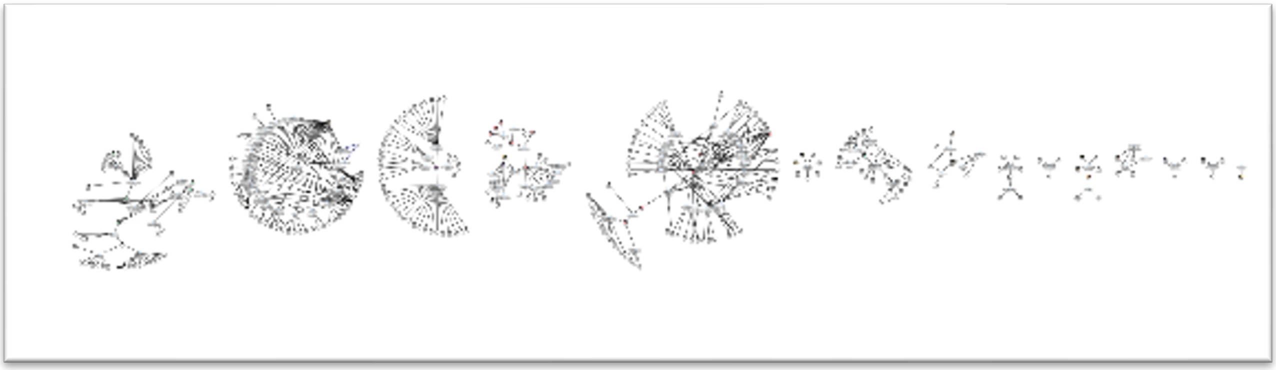
Activity 3: Publish quarterly blog posts about DART on AEDC blog	**4 blogs published**	**4 blogs published**	Fowler
Activity 4: Maintain @arepscor Facebook, Twitter, and YouTube channels and refresh DART content frequently	**Following increased by 10%**	**Following increased by 10%**	Fowler
Objective 8.2.b	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Establish listserv and group of communications reps from each participating campus	**Committee formed, first meeting complete**		Fowler
Activity 2: Hold annual check-in meetings to ensure proper citation of project and related messaging and disseminate project updates		**1 meeting complete**	Fowler
Objective 8.2.c	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Establish DART ER Core Site	**ER Core Site published & accessible**		Fowler
Activity 2: Maintain DART ER Core site and provide annual training to participants	**Participants onboarded, 3 training webinars complete**	**3 training webinars complete**	Fowler
Objective 8.2.d	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Presenting at national conferences / professional societies		**2 presentations complete**	Co-Leads
Activity 2: Publications		**1 publication complete**	Co-Leads
Activity 3: State-wide Workshops for Cohorts and Waves	**2 Workshops complete**	2 Workshops complete	Schubert
Objective 8.2.e	**Specific Milestones**		
	Year 1	Year 2	Responsible
Activity 1: Create committee responsible for implementing SJC		**Committee formed; first meeting complete**	Fowler
Activity 2: Develop plan for SJC		**Plan disseminated to stakeholders**	Fowler

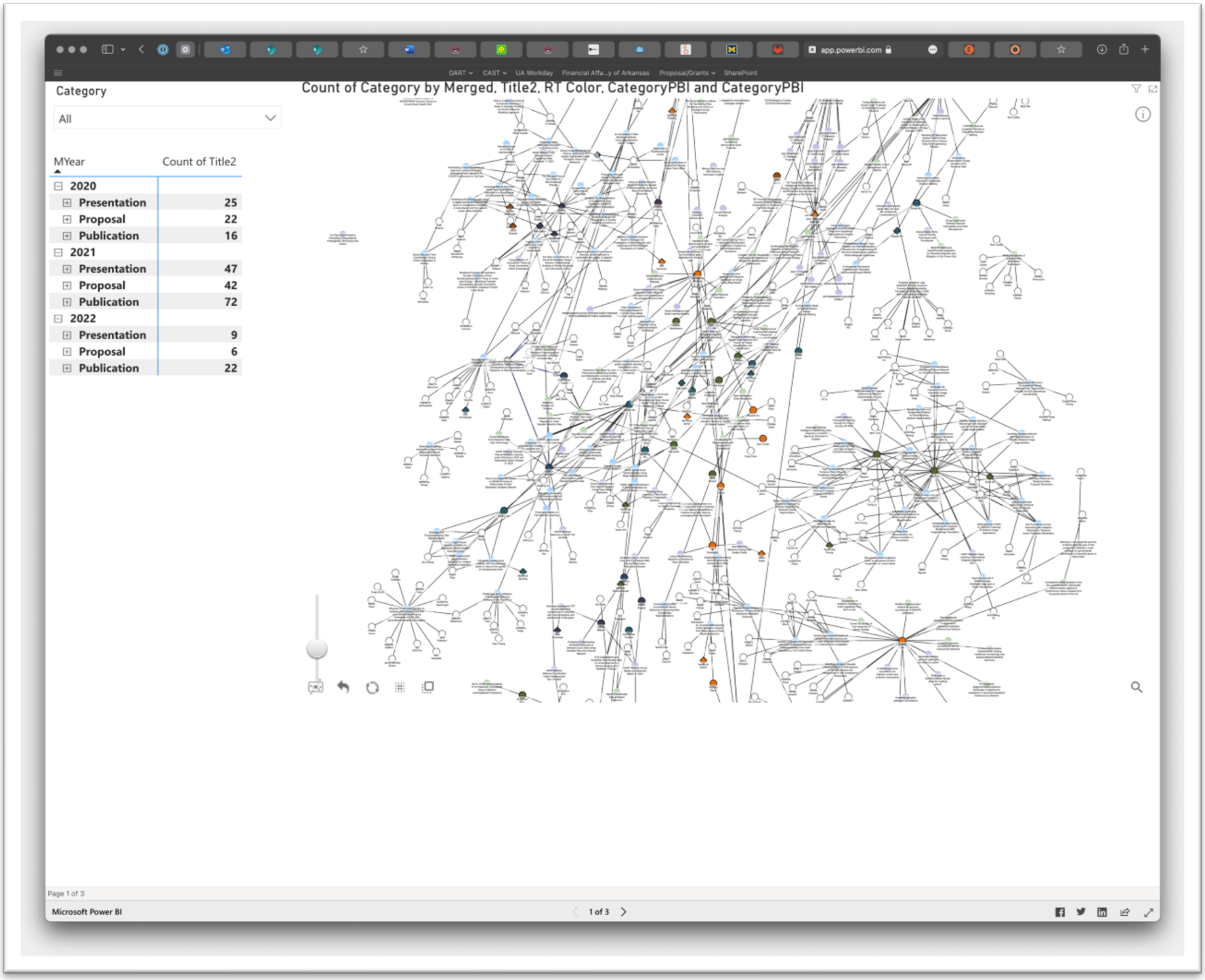
DART Year 2 Participant Demographics (Appendix 2)

The table below lists all Year 2 DART participants by self-identified race, ethnicity, and gender.

Race and Ethnicity (Self-Identified by Participants)	Gender				Grand Total
	Demigendered	Female	Male	Prefer not to say	
African			1		**1**
Asian		27	58		**85**
Asian, Caucasian			1		**1**
Asian, Hispanic, Latina / Latino		1			**1**
Black or African American		6	7		**13**
Black or African American, Native American, White or European American		1			**1**
Caucasian	1	11	18		**30**
Caucasian, Native American			1		**1**
Caucasian, White or European American		4	5		**9**
Hispanic			2		**2**
Hispanic, Latina / Latino		1	1		**2**
Hispanic, Latina / Latino, Native American, White or European American		1			**1**
Latina / Latino			2		**2**
Middle Eastern or North African		3	5		**8**
Middle Eastern or North African, White or European American		2			**2**
Prefer Not to Say		1	3	2	**6**
White or European American		13	27		**40**
Grand Total	**1**	**71**	**131**	**2**	**205**

Example DART Collaboration Visualizations (Appendix 3)





Category

All

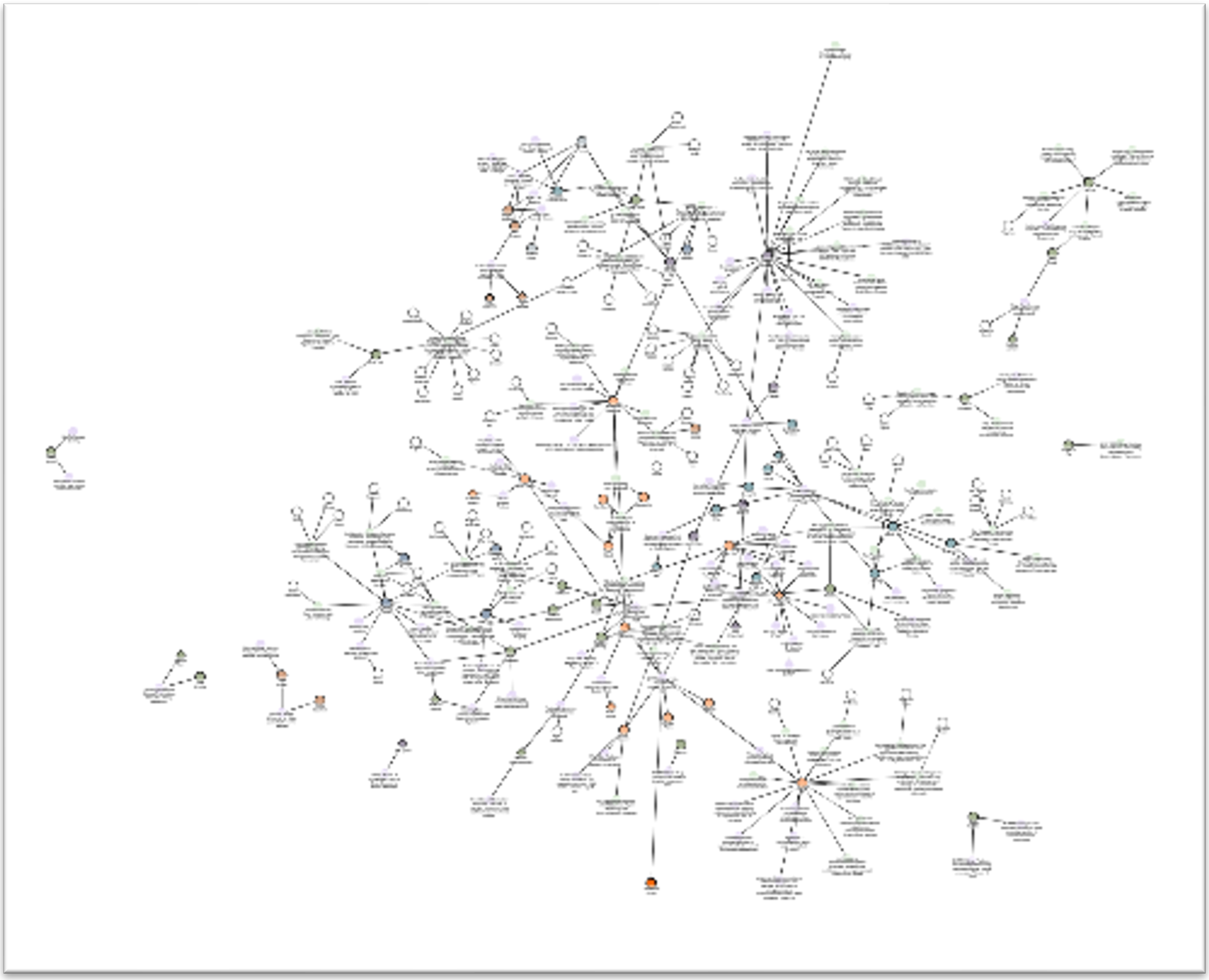
MYear	Count of Title2
2020	
Presentation	25
Proposal	22
Publication	16
2021	
Presentation	47
Proposal	42
Publication	72
2022	
Presentation	9
Proposal	6
Publication	22

Count of Category by Merged, Title2, RT Color, CategoryPBI and CategoryPBI

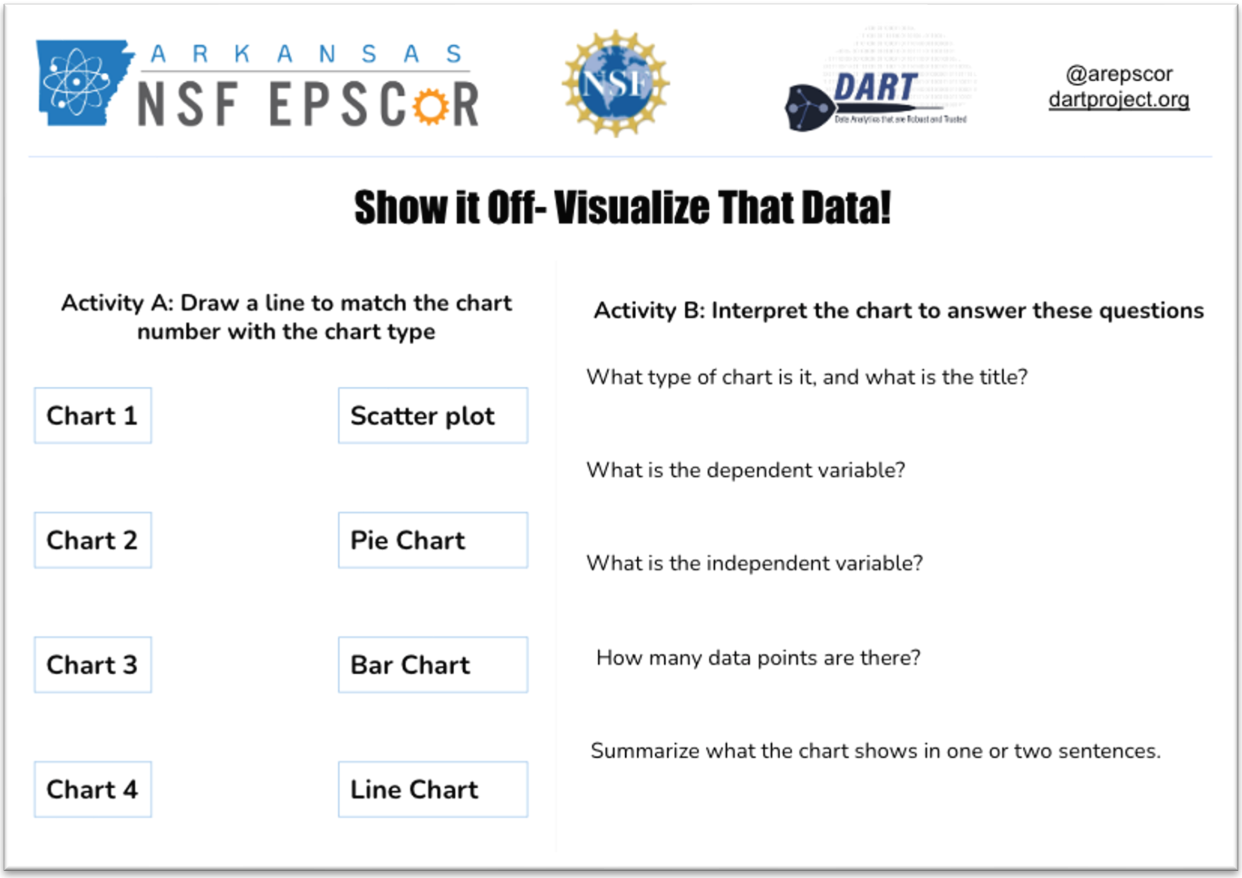
Page 1 of 3

Microsoft Power BI

1 of 3



Worksheets Developed for EAST Conference (Appendix 4)



@arepscor
dartproject.org

Show it Off- Visualize That Data!

Activity A: Draw a line to match the chart number with the chart type

Chart 1	Scatter plot
Chart 2	Pie Chart
Chart 3	Bar Chart
Chart 4	Line Chart

Activity B: Interpret the chart to answer these questions

What type of chart is it, and what is the title?

What is the dependent variable?

What is the independent variable?

How many data points are there?

Summarize what the chart shows in one or two sentences.

Why is information visualization important? It helps us think about and understand data! It provides insight into data which can help with decision making or information discovery. Data visualization does not require computers or technology, but computers and technology can often improve visualization. Visualization styles can range from basic or practical to artsy and beautiful.

Here are some things to consider when it's time to visualize your data. First think about the type of data, the size or amount of data, and the complexity of the data. Is it a simple time series with a few data points, or is it a large and complex dataset with several types of data? Next, think about the information you want to convey, or information that you need to extract from the data. What is the story you want to tell? Finally, you need to think about who you are presenting the data to. Who is your audience? Will they understand the terms and components you selected for the visualization?

Data visualization tips:

- Use color, shading, and patterns carefully and intentionally- remember that your viewers could be color blind or have other visual impairments.

- Avoid using familiar colors in unfamiliar ways. If you use red to represent increases and green to indicate decreases, it might confuse the viewer who is used to red meaning negative and green meaning positive.
- Use consistent scales. If your chart or graph is meant to show the difference between data points, your scale must remain consistent. If your scale is inconsistent, it can cause significant confusion for the viewer.
- Qualitative data tends to be better suited to bar graphs and pie charts, while quantitative data might be best represented in formats like charts and histograms.
- There are LOTS of resources online to help you in your visualization journey! A simple internet search will provide many quality results.

Letter	Binary	Letter	Binary	Letter	Binary	Letter	Binary
A	1000001	H	1001000	O	1001111	V	1010110
B	1000010	I	1001001	P	1010000	W	1010111
C	1000011	J	1001010	Q	1010001	X	1011000
D	1000100	K	1001011	R	1010010	Y	1011001
E	1000101	L	1001100	S	1010011	Z	1011010
F	1000110	M	1001101	T	1010100		
G	1000111	N	1001110	U	1010101		

Use the key above to write your name in binary code:

@arepscor
dartproject.org

WHAT DOES IT ALL MEAN???

Your Guide to Common Data Science & Computer Science Terms

Data is plural- the singular unit of data is datum. Data are facts and statistics that are collected together for a reason, it could be for others to reference or for analysis and problem-solving. Data can be organized into tables, lists, charts, and graphs.

A variable is something that you measure- some examples are temperature, location, quantity, quality, color, size, or weight.

An independent variable is not changed by other variables you are trying to measure in an experiment. For example, if you want to measure housing prices in different zip codes, the independent variable would be the zip code- it doesn't change.

A dependent variable is affected by other factors- in our housing prices example, the price of the house would be dependent on a number of things, the size, age, condition, and location of the house.

The x-axis on a standard graph is usually horizontal and represents the independent variable.

The y-axis on a standard graph is usually vertical and represents the dependent variable.

Features are what data scientists call the columns in a dataset. A feature is a unique measurable property or characteristic. Features can be qualitative or quantitative.

Observations are what data scientists call the rows in a dataset. Each observation represents the measurement of a corresponding feature.

Quantitative data is measuring the quantity, or amount of something. It can be counted or compared on a numeric scale. For example, how many students attended EASTCon22?

Qualitative data is data describing attributes or qualities, something you can see or feel. The properties can be categorized and may be assigned numeric values to help organize it. For example, what was your favorite part of EASTCon22?

A bit is a binary digit- it is the smallest unit of information. A bit can only be one of two possible values and is commonly represented as 0 or 1, but could also be represented as on or off, yes or no, true or false.

A byte is a string of 8 bits, which is the number of bits it takes to encode a single letter or number in a computer. A byte is the smallest unit of memory in a computer.

Nominal data can be organized into categories that do not have a natural order. For example, colors (blue, green, yellow, red) or types of fruit (oranges, apples, bananas, pears).

Ordinal data can be organized into categories that do have a natural order. For example, school grade level- 5th grade, 6th grade, 7th grade, 8th, 9th, 10th, and so on. Another example is a Likert scale- a point scale used by researchers to take surveys and get people's opinion on a subject matter. How satisfied are you with this definition? You could choose very dissatisfied (1), slightly dissatisfied (2), indifferent (3), satisfied (4), or very satisfied (5).

Numerical data refers to the data that is in the form of numbers that can't be described with language or descriptive form.

Data science is solving problems with data. It involves collecting data, organizing it, analyzing it, and developing solutions.

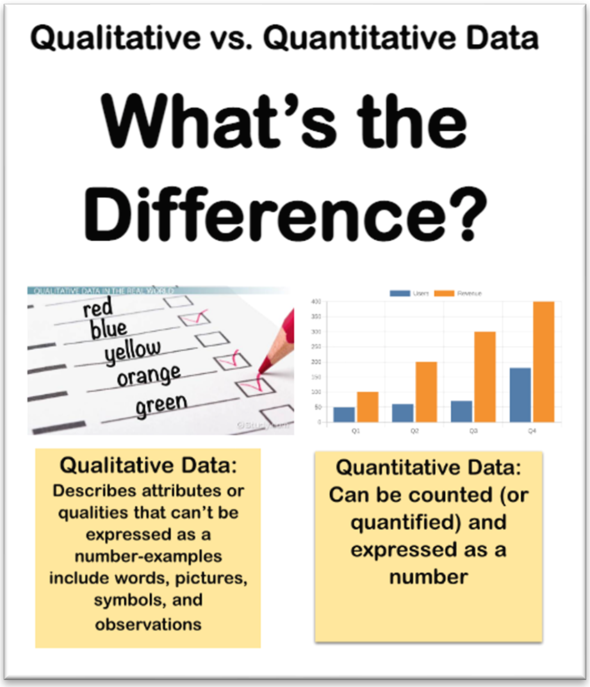
Computer engineering is the science of building computers- the physical electronic parts, and how they all work together.

Computer science is solving problems with computers- understanding how they work and teaching them to do things for us, like writing software or programming algorithms.

Algorithms are complicated steps of instructions describing how to complete a task, such as solving a problem. Use the Sort it Out game as an example. When you examine the objects together, you will think and make a decision about how the objects should be sorted based on their attributes. You could sort them by color, material, or object type. Then you move the objects into the appropriate basket based on their attributes. If you wrote down the sorting instructions for someone else, you have written an algorithm. You could also teach a computer how to sort them.

Artificial intelligence is the intelligence of machines, as opposed to natural intelligence of animals and humans. Intelligence means the ability to learn and apply knowledge. The most common application of artificial intelligence is called machine learning, and many people use the terms synonymously or interchangeably.

Machine learning is the study and design of computer algorithms that can improve themselves automatically over several rounds of analysis, also known as epochs or iterations, by the use of data. Machine learning algorithms build a model based on sample data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so.

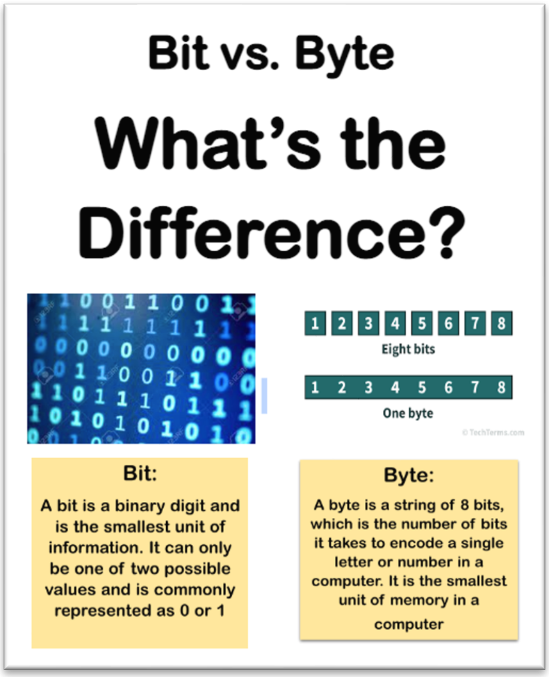

Qualitative vs. Quantitative Data

What's the Difference?

QUALITATIVE DATA IN THE REAL WORLD
red
blue
yellow
orange
green
Users
Revenue
400
350
300
250
200
150
100
50
0
Q1
Q2
Q3
Q4

Qualitative Data:
Describes attributes or qualities that can't be expressed as a number-examples include words, pictures, symbols, and observations

Quantitative Data:
Can be counted (or quantified) and expressed as a number


Bit vs. Byte

What's the Difference?

1 2 3 4 5 6 7 8
Eight bits
1 2 3 4 5 6 7 8
One byte

Bit:
A bit is a binary digit and is the smallest unit of information. It can only be one of two possible values and is commonly represented as 0 or 1

Byte:
A byte is a string of 8 bits, which is the number of bits it takes to encode a single letter or number in a computer. It is the smallest unit of memory in a computer