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

Year 1 Annual Report

Award Number: 1946391
Jurisdiction: Arkansas
PI: Jennifer Fowler
Co-PI: Jackson Cothren
Awardee Institution: Arkansas Economic Development Commission
Award Start Date: July 1, 2020
Award End Date: June 30, 2025 (Estimated)
Report Submission Date: April 1, 2021
Reporting Period: July 1, 2020 – March 31, 2021

1. Overview 3
1.1. Mission 3
1.2. Vision 3
1.3. Project Goals 3
2. Intellectual Merit 7
3. Broader Societal Impact 8
4. Roles and responsibilities in the Project 10
5. Summary of Year 1 Major Accomplishments 12
5.1. Big Data Management 12
5.2. Model Interpretability 16
5.3. Privacy and Security 20
5.4. Education, Outreach, and Workforce Development 22
6. Summary of Significant Problems, Novel Opportunities, and Changes in Strategy 26
7. Research and Education Program 27
7.1. Coordinated CyberInfrastructure 27
7.2. Data Life Cycle and Curation 32
7.3. Social Awareness 40
7.4. Social Media and Networks 50
7.5. Learning and Prediction 54
7.6. Education 57
8. Workforce Development and Broadening Participation 60
9. Communication and Dissemination 62
10. Broadening Participation 64
11. Expenditures and Unobligated Funds 68
12. Tabular/Graphic representation of progress to date (Attachment- Stoplight Tables) 68
13. Appendices 69
13.1. DART Mentorship Program Guidelines and Materials: For Seed Grant Faculty & Postdocs 69

1. Overview

1.1. Mission

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.

As Arkansas transitions to a more diverse, data-driven economy we must create an environment for university and industry collaborations in data science that will sustain this new economy with cutting edge research and educate a workforce that enhances competitiveness in Arkansas industries. By bringing together experts from different data science sub-fields and application areas, we expect to develop both specific and comprehensive solutions that would be difficult to obtain in isolation. Collaboration with our industry partners provides a better definition of both problems and solutions in data analytics and workforce education.

1.2. Vision

The Arkansas research community - academic, government, and industry - collaborate 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 affect 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. This trust allows 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.

1.3. Project 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. Our research is organized around addressing these barriers in a multi-disciplinary but coordinated way with our research themes.

Three main barriers form the integrative research questions on which DART is focused: big data management, model interpretability, and data security and privacy. Additionally the project addresses the need for a workforce trained in a variety of data skills. Activities in each research theme contribute to those four overarching topics. The degree of interaction between themes is defined by that joint contribution and are shown as research networks below in figures 1, 2 and 3. Note that projects are integrated across themes. In these networks, colored nodes are the goals/projects in each research theme, while the grey nodes represent our faculty and their students. The numbers inside the nodes represent the estimated person-hours attributed to that particular activity. We plan to continually update these collaboration networks throughout the project.

The flow from the research activities defined below, through the integrative questions, and finally into economic development through industry partners is captured in Figure 4. Flow capacities are loosely based on the number of objectives/activities from each theme that contribute to the integrative question (these will change over time) and the-still admittedly arbitrary-importance of each barrier to economic activity in state.

Big data management: Before data streams and datasets can be used in the many kinds of learning models, they are often manually curated, or curated for a specific problem. We still rely on hosts of 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.

Security and privacy: Government agencies and private entities collect and integrate large amounts of data, process it in 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, and negatively impact diversity in our society.

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.

Figures Below.

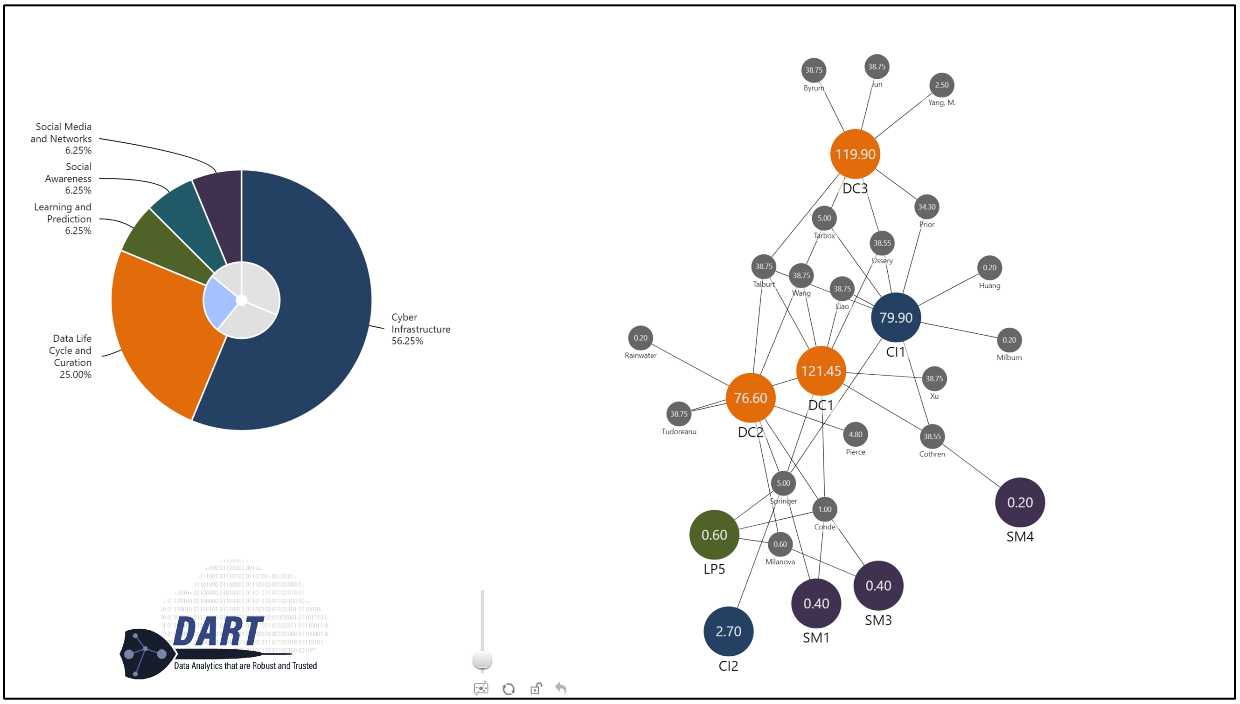


Figure 1. Big data management research network.

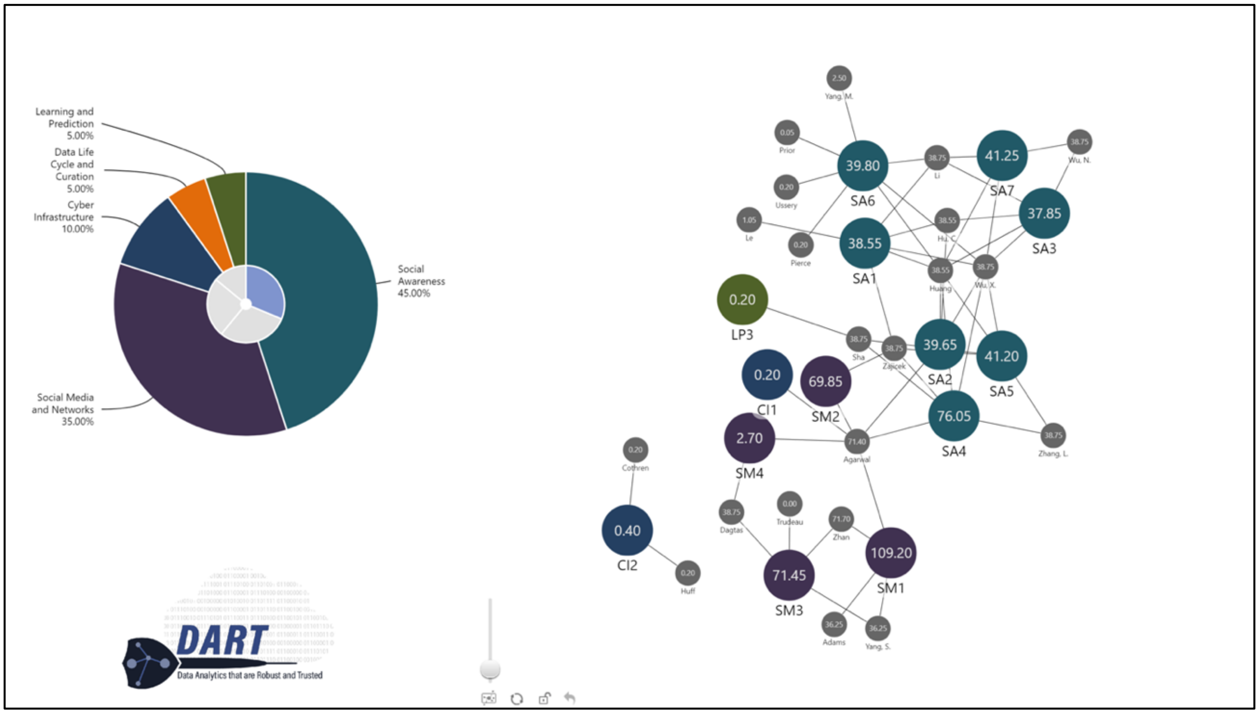
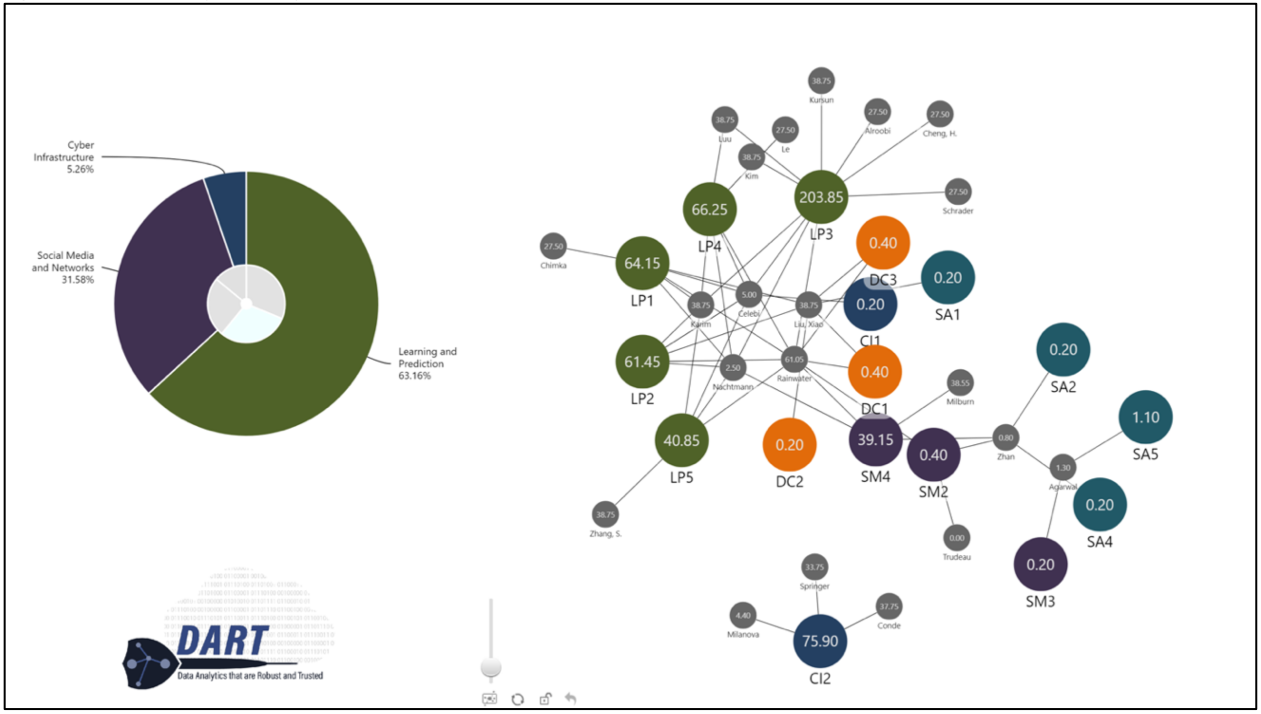


Figure 2. Privacy and security research network.

Figure 3. Model interpretability research network.



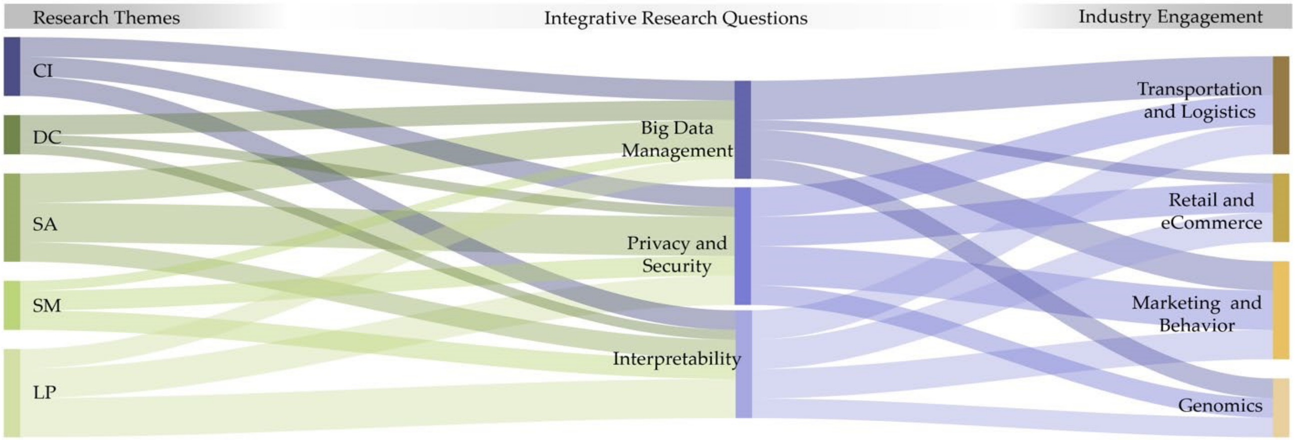


Figure 4: Sankey diagram showing 1) the contributions of each research theme to the fundamental barriers we are addressing as a project and 2) industry dependence on research for removing or mitigating the barriers. Research theme collaborations and integrations are designed to address these barriers from different research approaches and perspectives.

2. Intellectual Merit

The **Learning and Prediction (LP)** research theme supports the creation of novel statistical learning methods in big data environments that are equipped with capabilities for addressing heterogeneity and hidden sub-populations within big datasets. Specifically, this research team has begun to create statistical learning methods in big data environments that are equipped with capabilities for addressing heterogeneity and hidden sub-populations within big datasets. They are making contributions in mode specification and interpretation through efficient variable selection using non-parametric methods. Eventually, this theme will advance computing in big data environments for traditional statistical modeling through statistical computing performed on distributed/parallelized computing nodes. Holistically, this theme will address challenges surrounding high-dimensional, dynamic, and unstructured data sets and explore solutions in the domains of genomics, transaction scenarios in eCommerce, and supply chain logistics.

Data Life Cycle and Curation's **(DC)** goal of building a "data washing machine" that can analyze and manipulate data as well as a person (a data analyst) is challenging. A data analyst brings a tremendous amount of experience and knowledge into the process. Representing, storing, and expressing this level of knowledge and experience will stretch the current capabilities of AI technology. While a general data washing machine robot that will work for any dataset might be decades away, creating useful and scalable automated solutions for these three use-cases (data cleaning, data integration, and data tracking) is an achievable goal within the 5-year time frame of the grant.

The **Social Awareness (SA)** research theme is working to advance socially aware data analytics and sharing by 1) researching and documenting privacy breaches, security concerns, and discrimination in big data applications and understanding factors leading to those negative outcomes; 2) producing a suite of novel technologies, differential privacy preserving, attack resilient, secure multi-party computation, and crypto based mechanisms/algorithms for a variety of data acquisition and analysis tasks; 3) conducting cutting-edge research in socially aware crowdsourcing, user-centric data sharing in cyberspace, cross-media discrimination prevention via multi-modal deep learning, fairness-aware marketing strategy design, and privacy-preserving analytics in health and genomics; and 4) creating a Web portal that includes policies, regulations, practices, algorithms, tools, prototype systems, and a collection of publicly available datasets and real data from our business partners.

Social Media and Networks (SM) primarily includes 1) innovative methods, techniques, and platforms for mining argumentation data and analyzing its characteristics, such as polarization, opinion diversity, participant influence, opinion community, and opinion prediction; 2) creation of a transformative multilayered network analytic method of analyzing deviant behaviors in social media networks by modeling multi-source, supra-dyadic relations and shared affiliations among deviant groups; 3) multimodal deep learning methods to work with multimedia data from social media and other data platforms; and 4) innovative algorithms for logistics planning in disaster response using big social data analytics.

Cyberinfrastructure (CI) is engaged in building the Arkansas Research Platform (ARP), which is pushing the edge of distributed high-performance computing coupled to distributed high performance storage via high bandwidth networks at small and medium sized research computing centers. While all these commodities are generally available at larger institutions, they are often out of reach for smaller institutions. Smaller institutions that do manage to acquire small compute clusters outgrow them quickly. The goal of the ARP is to federate these scattered resources into one whole resource. As important as the lessons learned from the ARP experience is the improved access to cyber infrastructure for researchers in the state of Arkansas. This will facilitate research-particularly big data analytics- that has previously been inaccessible to researchers scattered across the smaller institutions within the state.

Education (ED) is a team represented by faculty at participating institutions as well as collaborators from two-year campuses across the state. The goal of this theme 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. A 2017 report by leading Arkansas companies identified a large gap in workforce needs that our existing computer science and information science programs were simply not meeting. 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.

3. Broader Societal Impact

DART, as a center, is integrating data science research and education across the State and creating a deep and diverse data-ready workforce. This is already paying dividends in the form of increased federal grant and industrial research funding. As the State better aligns its investments with industry strengths, more opportunities to increase educational attainment and wages will develop, improving the quality of life for all Arkansans. Each thematic research area contributes in complementary ways to this mission.

Big data analytics is a heavy consumer of computing and storage resources. The lack of access to such resources acts as a barrier to talented researchers from under-served and smaller institutions. ARP intends to flatten that playing field by giving all researchers at Arkansas institutions, regardless of size or budget, access to the compute and storage resources available at the larger institutions. Past experience has shown that having such access can greatly increase the pace of discovery by tapping intellectual resources that otherwise would be under-utilized due to a lack of access to adequate compute and storage resources. While access to the resources through ARP is crucial, it will not have a significant impact if its clients lack the technical skills to make effective use of them. Expanded data science undergraduate and graduate degree programs are necessary, but smaller, more focused training on how to build research code and tools using the platform are equally important and will translate to industry and government environments. Organizations like The Carpentries offer well developed and tested training modules on basic modern computing tools (Git, IDEs, markdown), high-level programming libraries, visualization tools, and data science libraries necessary for effective data science.

Currently data scientists are spending only 20% of their time modelling, and 80% of their time working on cleaning and preparing data. Reducing that 80% would significantly increase productivity. Even bringing this down to only half that (40% effort to clean / prepare data) would be of enormous benefit to both industry and research in Arkansas, allowing data scientists to spend more of their time (80% vs. 20%) on working with data analysis and modelling.

Research contributions from DART will improve the learning and prediction of data in a spectrum of applications including commerce, cybersecurity, disaster and emergency management, energy, environment, healthcare, retail, and transportation. Research outputs will generate interest in data science and help engage, encourage, and recruit a broad spectrum of learners as well as researchers. As a result, Arkansas should see a growth in research and education initiatives in data science. We expect to grow the segment of society that can benefit from Artificial Intelligence-driven solutions by eliminating economic barriers to technology access and boost Artificial Intelligence applications and efficient platforms to support Arkansas economy and workforce development.

This research will address security and privacy, practical application and acceptance of data analytics, and develop novel, integrated solutions for achieving privacy preservation, fairness, safety, and robustness in big data learning and sharing. This research will help organizations and individuals understand the uses, benefits, and risks of big data; determine whether disclosure of private information, unfair treatment, or potential risks have occurred or would occur; and assist the community in the endeavor to provide trustworthy technologies. The principles, methods, tools, datasets, and evaluation results will significantly affect the development of a socially responsible science and engineering workforce in Arkansas. Moreover, by advancing socially aware data analytics and proposing viable solutions that will assure that big data are collected and used in a safe, private, fair, and responsible way, this project will contribute to the wider acceptance and support for big data products. Finally, innovative methodologies and tools developed for socially aware learning and sharing will help U.S. companies compete and lead globally.

Societal polarization, amplified by the massive reach of "always on" social platforms, is threatening and damaging democracy around the world. The models and insights generated will enhance our ability to both capitalize on the potential of social media as a force of good and mitigate its use as a weapon. Threats to democracy are abated through new models to understand how polarization forms, methods to detect online deviant behaviors, and interventions to prevent the spread of misinformation and rise of echo chambers.

Disaster management is another direct application of this research. Extreme weather events and major natural disasters are ranked by world leaders as the biggest risks facing our planet. The research will benefit disaster response decision-making by affording new tools and technologies that extract, classify, index, and analyze diverse and semantically rich multimedia social data to boost situational awareness. SM theme engages diverse faculty and students to develop smart, explainable, and accurate data analysis techniques.

Integrating data science research across the State and creating a deep and diverse data-ready workforce will pay immediate dividends in the form of increased federal grant funding, increased industrial research funding, and increased employment in well-paying jobs. For the first time in AR EPSCoR history, we are working with community colleges and three of the state's Historically Black Colleges and Universities (HBCUs) to ensure inclusive learning pathways for a broad population of students. The HBCU roles are critical to the success of the entire education component. Together, we are working to build a statewide data science educational ecosystem to allow learners across the state opportunities to enroll in modular, scalable data science certificate and degree programs. DART also includes a data science summer institute for undergraduates, summer internships and research experiences, increased data science educational opportunities, and curriculum to include relevant data science topics and capstone projects.

Table 1. Participating Institutions

Institution Name	Institution Type	Region	Project Component
Arkansas Economic Development Commission (Awardee)	State Government	Central	Administration/Central Office
Arkansas State University	Public ARI	Northeast	Research
Philander Smith College (HBCU)	Four Year Private	Central	Education
Shorter College (HBCU)	Two Year Private	Central	Education
Southern Arkansas University	Public ARI	Southeast	Education & Research
University of Arkansas for Medical Sciences	Public ARI	Central	Research
University of Arkansas, Fayetteville (MSI)	Public ARI	Northwest	Education & Research
University of Arkansas, Little Rock	Public ARI	Central	Research
University of Arkansas, Pine Bluff (HBCU)	Public ARI	Central	Education & Research
University of Central Arkansas	Public ARI	Central	Education & Research

4. Roles and responsibilities in the Project

Science Steering Committee (SSC; also known as: Leadership Team): The Science Steering Committee is comprised of co-leads from each research theme. It provides oversight for the scientific aspects of the program and is responsible for ensuring research theme milestones and objectives are being met annually. The SSC is also responsible for participating in NSF Site Visits and annual conferences, as well as communicating progress to the external evaluation board and external evaluator via annual reports and presentations. The SSC works closely with the External Advisory Board (EAB), PI, and Co-PI to provide technical and/or scientific guidance as lead researchers on the project. Each SSC member is responsible for planned research in the theme and planning, execution, reporting, and dissemination via inter-institutional workshops.

Management Team: The Management Team is comprised of vice-provost level administrators from each campus receiving a subaward. The PI, Co-PI, and Management Team are responsible for financial decisions and other administrative duties.

Science Advisory Committee (SAC; also known as the Arkansas EPSCoR Steering Committee): The Science Advisory Committee is composed of representatives from academia, government, and the private sector. The SAC selects the topical areas for each Track-1 Project, designates the fiscal agent/proposing organization as the responsible recipient for the RII Track-1 award, and must provide support for the Track-1 Project for NSF acceptance.

External Advisory Board (EAB): The External Advisory Board includes researchers from peer and aspirant universities or national labs who serve as technical consultants providing recommendations on research progress and strategic and long-term sustainability planning during annual site visits. The EAB will serve a critical role in the seed grant program as well as in mentoring and commercialization efforts.

Table 2. Confirmed External Advisory Board Members

Name	Organization	Title
Dr. Donald Adjeroh	West Virginia University	Professor of Computer Science Director of West Virginia-Arkansas Center for Research and Education in Smart Health
Dr. James Caverlee	Texas A&M University	Professor, Computer Science & Engineering
Dr. Carolina Cruz Neira	University of Central Florida	Agere Chair, Professor of Computer Science
James Deaton	Great Plains Network	CEO
Dr. Hoda Eldardiry	Virginia Tech	Associate Professor of Computer Science Director, Machine Learning Lab
Dr. Huan Liu	Arizona State University	Professor of Computer Science and Engineering
Dr. Michael Khonsari	Louisiana State University	Dow Chemical Endowed Chair, Professor of Mechanical Engineering
Dr. Srinivasan Parthasarathy	Ohio State University	Computer Science and Engineering Director, Data Mining Research Laboratory Co-Director, Data Analytics Program
Dr. Dirk Reiners	University of Central Florida	Professor of Computer Science
Dr. Weisong Shi	Wayne State University	Professor of Computer Science Associate Dean for Research and Graduate Studies
Dr. Jason Leigh	University of Hawaii at Manoa	Professor, Information and Computer Sciences

Industry Advisory Board (IAB): The IAB members serve as an intermediary between academia and industry. The IAB includes representatives from Arkansas industry sectors who will be impacted by DART research. Ex-officio members include the project Co-PI (to communicate scientific results), the project PI (to serve as the liaison to government and policy organizations), and members from related organizations like the Arkansas Center for Data Science (ACDS) and the Arkansas Research Alliance (ARA). The IAB will meet annually in conjunction with the Annual All-Hands Meeting, and they will meet quarterly to review results and recommend new areas of research and collaboration based on industry needs. One member of the IAB (rotating annually) will serve as a member of the EAB during site visits.

Table 3. Confirmed Industry Advisory Board Members

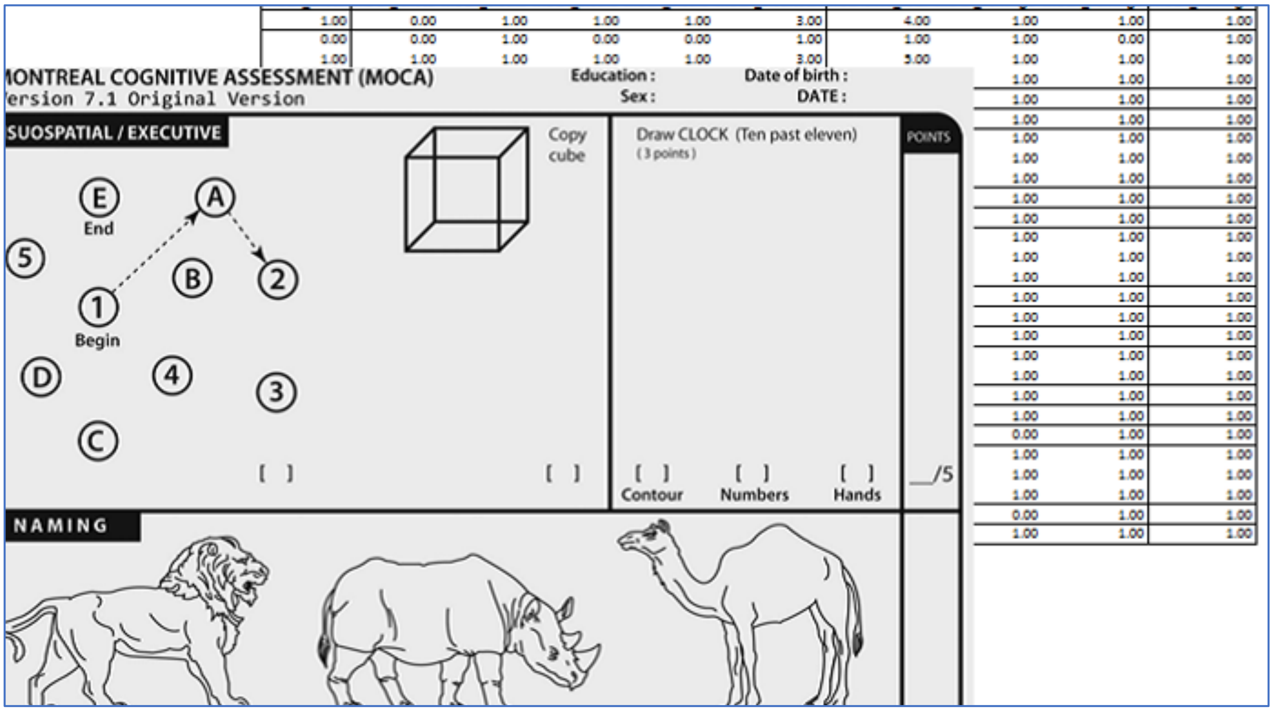
Name	Organization	Title
Dr. Salomon de Jager	PiLog Group	President and CEO
Dr. Justin Magruder	Science Applications International Corporation	Chief Data Officer
Dr. Vikram Manikonda	Intelligent Automation, Inc	President and CEO
Kash Mehdi	Informatica	Data Governance and Privacy Domain Expert
Dr. Adewale Obadimu	LinkedIn	Sr. Network Scientist
Ben Moyer	Hytrol	Director of Engineering
Josh Stanley	Cartwheel Startup Studio	Managing Partner
David Hauser	GENPACT	Chief Science Officer

5. Summary of Year 1 Major Accomplishments

5.1. Big Data Management

The three most time-consuming big data preparation processes are data cleaning, data integration, and data tracking (data governance). The idea of a “data washing machine” is prominent in our efforts to improve big data management. People are accustomed to throwing their dirty laundry into the washer along with some soap, setting the dials for the type of clothes, and letting the washer operate automatically. A data washing machine would work in a similar manner on dirty data - simply ‘throw in dirty data’, push a button, and out comes ‘clean’ or curated data. Building such a machine implies a ML/AI approach and the DC theme is taking steps.

The concept of a data washing machine guides researchers in the data lifecycle and curation theme as we work to remove most of the human effort and time that's required to take random information and turn it into a well-organized, well-understood data repository. We want to automate this to the greatest extent possible so that we can clean and integrate data before we attempt to build learning models from it. We want not only to track where it is, but maintain the provenance of its transition from being acquired to the various models and predictions made from it. One activity related to this question is being undertaken by Prior and Eubank (CI1 and DC4). We are exploring the quality of data used to track the progression of Parkinson's disease - a transition from normal cognitive behavior to dementia through what is called mild cognitive impairment. This is typically assessed via a battery of tests, such as the Montreal Cognitive Assessment, or MOCA. This is a complex form where people are required to make drawings and connect the dots and answer a collection of questions. In short, it is a multimedia document that is typically scored by a human (figure 5).



1.00	0.00	1.00	1.00	1.00	3.00	4.00	1.00	1.00	1.00
0.00	0.00	1.00	0.00	0.00	1.00	1.00	1.00	0.00	1.00
1.00	1.00	1.00	1.00	1.00	3.00	5.00	1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							0.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							1.00	1.00	1.00
							0.00	1.00	1.00
							1.00	1.00	1.00

Figure 5. Montreal Cognitive Assessment example

The goal is to try to improve on the rather subjective standards for scoring these tests and more reliably assess mild cognitive impairment. While we will apply standard and novel machine learning techniques, we are first faced with this daunting problem of not just scoring this test, but an entire battery of such tests-all of which are multimedia types of events-as well as trying to correlate that information with imaging data, particularly MRI and fMRI data, as well as positron emission tomography (PET) data and electrical electroencephalography or EEG signals. As we acquire this data from electronic health records and research repositories, we have massive data quality issues because these pieces of information are stored as PDF files that are largely unstructured and non-uniformly incomplete. One goal of the research is to improve the quality of this information - this large, multimedia, highly complex dataset.

No big data management research is possible without a shared platform on which developers can build and deploy applications for research and testing. This is the focus of CI1 with assistance primarily from DC research theme members who are providing applications for the platform. Together we have defined and made progress toward implementing an infrastructure for managing and analyzing big data which we call the Arkansas Research Platform, or ARP. Our research program requires a state-of-the-art cyberinfrastructure that combines high-performance computing and cloud computing in both private and public clouds with petabyte scale storage and high bandwidth networking. Our approach is to take the high-performance and cloud computing resources at UA-Fayetteville and UAMS in Little Rock, link them together through an extended ScienceDMZ that links the two institutions to form one relatively large computing resource that is made accessible to other members of the program throughout the state. In year 2 we will add our colleagues at UA-Little Rock, who are providing advanced data visualization (CI2) resources to ARP. In year 1 we have worked to implement best practices and tools for software development and code management included a shared GitLab repository that will be linked directly to HPC resources in year 2.

For example, we want to enable a DART research team (SM4) that is studying how to leverage mobility data, social media data, and overhead imagery to better re-route emergency vehicles during and after a disaster to 1) develop code for data pipelines that stream data from multiple application interfaces (Planet, Twitter, and others), 2) apply a sequence of operations to extract, transform, and load the data into databases at UA-Fayetteville and/or UAMS, and 3) make the transformed data available to other researchers in DART. The team will store all code supporting the work in a project-wide git repository so that other researchers in this theme and in others can clone or branch it to their projects, containerize it with its dependency environment, and demonstrate it using a collection of Jupyter Notebooks in a Jupyter Lab environment. In addition to extract/transform/load (ETL) pipelines, machine learning models developed for this data will be created and managed by other projects in a similar way to make it easy for others to see progress, replicate on different data, learn from the work, and use components as necessary.

Progress in year 1 towards implementing this capability DART-wide centered around two elements: expanding the number of GPU compute nodes at UA-Fayetteville and re-architecting the ScienceDMZ on both UA-Fayetteville and UAMS campuses. At the end of year 1, this capability – representing a nearly complete transition from SSH shell access and batch programs run on clusters at two campuses – is not yet widely available. However, the technical framework exists and will be introduced in year 2 in accordance with the planning activities described below.

Significant improvements to upgrade Pinnacle are underway and despite COVID-based supply chain issues are on track for year 1 completion (details can be found in a subsequent section of this report). However, we learned that working across campuses can be an extremely complicated process that involves often conflicting responsibilities and priorities, multiple organizations, and people who have not traditionally worked together (especially across research computing and enterprise IT organizations). Nevertheless, we made extensive progress that will pay dividends in sustainability and buy-in by engaged enterprise IT staff and resource hungry research staff:

We organized a series of weekly meetings including IT networking staff, research computing staff, and the state's Internet 2 provider (AREON). These meetings allowed DART CI team members to describe the research computing environment we required, the IT staff to voice network security concerns, and all attendees to establish a trusted working relationship. This proved invaluable for future work. This began with UA-Fayetteville and UAMS – the two primary resource providers – but will expand with SHARP CCI to all campuses (item 3).

AREON, along with ITS staff and DART researchers at UA-Fayetteville and UAMS engaged with the TrustedCI (NSF #1920430) and EPOC (NSF #1826994) teams at Indiana University to develop a coordinated ScienceDMZ at UA-Fayetteville and UAMS that could be accessed or even joined by other DART institutions in the state. The document will be published in year 2 of the DART project.

UA-Fayetteville applied for, and was awarded a CC* planning grant. The Shared Arkansas Research Plan for Community Cyber Infrastructure (SHARP CCI, NSF #2126108) project involves all DART universities and during year 2 will work with the DART CI team to

a. define the current state of the RCI on each campus and determine the basic equipment and staffing capabilities needed to access the ARP and meet ongoing and future research requirements, including:
    - on-premise ability to utilize high-speed connectivity (10 Gbps) to the ARE-ON backbone
    - participation in InCommon Federation for identity and access management solutions
    - IPV4 vs. IPV6 capability
    - access to the ARP science DMZ, including DTN's, distributed file systems, cluster login nodes and other computing resources,
    - use of HPC clusters and high-speed storage systems at UAF and UAMS,
    - use of OSG, GPN, XSEDE, Internet2, ESnet, and elastic Cloud research services,
    - participation in the Arkansas Cyber Team (ACT) collaboration to support managed RCI systems and offer user training and research support services.
    - develop a Master Information Security Policy & Procedures (MISPP) manual in accordance with the Trusted CI Cybersecurity Framework to govern the access and use of the Arkansas Research Platform (ARP)

b. complete the ARP Master Information Security Policy & Procedures (MISPP) documentation and integrate it with the campus RCI Plans into a Regional RCI Plan for the State.

The activities surrounding SHARP CCI will push some year 1 milestones, notably the formalization and final acceptance of a governance and management structure for the Arkansas Research Platform and the use of InCommon as the authorization/authentication method, into year 2 and beyond. However, they create a much stronger and more robust foundation for ARP and will sustain the platform beyond the end of DART. We cannot over emphasize the support of the UA-Fayetteville and UAMS CIO's who dedicated significant staff time and reorganized around research computing to support DART and the significant improvements in policy and networking required to support the Arkansas Research Platform.

5.2. Model Interpretability

Machine learning has great potential for improving products, processes, and research, but computers usually do not explain their predictions, which is a barrier to the adoption of machine learning. Hence, the problem of un-interpretability arises. Model un-interpretability is prevalent in data mining, machine learning, and AI based systems. Recently, we have witnessed an explosion of emerging applications of AI and machine learning: detecting hate speech from text and images; detecting deepfakes; assessing coordination among bot accounts on social media that amplify misinformation; recommendation algorithms that predict our shopping behaviors; books and articles to read; videos to watch; autonomous driving; traffic management and smart city initiatives. In the DART project, interpretability is at the core of our models.

During year 1, we made several advances to maintain accuracy while enhancing explainability and transparency. Model accuracy and model interpretability need not be a trade-off. Model interpretability could in fact increase a model's accuracy because by being able to interpret a model's outcome, we will be able to explain why the model predicted wrongly in the first place. Recent research in formal methods have shown that it is possible to achieve model interpretability by converting black box models that are completely un-interpretable to white box models that are totally interpretable. In the DART project, we have adopted algebraic transformations to explain machine learning models. By coupling algebraic transformation with diagrammatic visualizations, we can further enhance model interpretability.

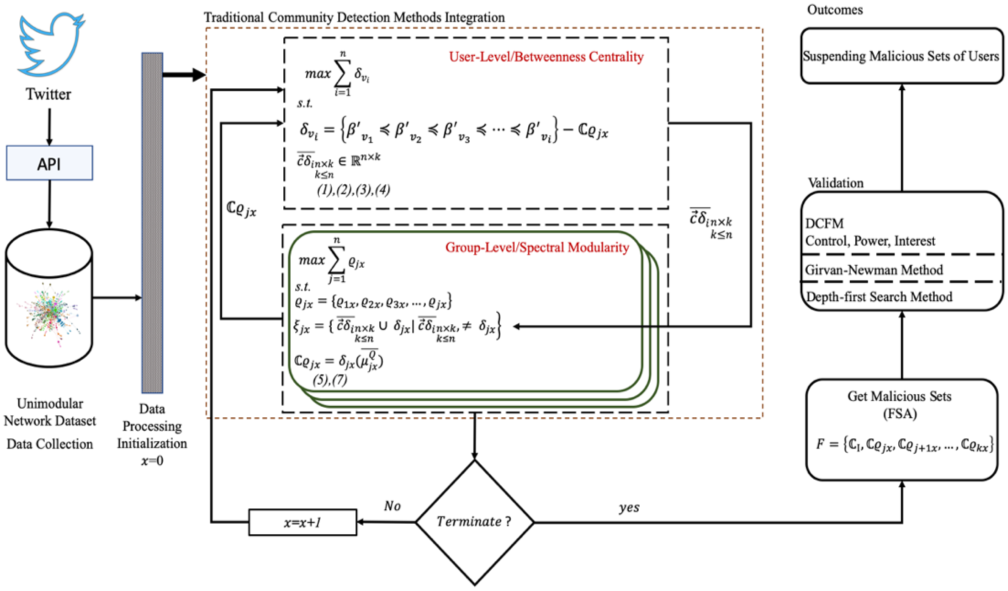

Twitter
API
Unimodular Network Dataset
Data Collection
Data Processing Initialization
x=0
Traditional Community Detection Methods Integration
User-Level/Betweenness Centrality
$max \sum_{i=1}^{n} \delta_{v_i}$
s.t.
$\delta_{v_i} = \{\beta'_{v_1} \preccurlyeq \beta'_{v_2} \preccurlyeq \beta'_{v_3} \preccurlyeq \cdots \preccurlyeq \beta'_{v_i}\} - \mathbb{C}\varrho_{jx}$
$\overline{\vec{c}\delta}_{\underset{k\leq n}{in\times k}} \in \mathbb{R}^{n\times k}$
(1),(2),(3),(4)
$\mathbb{C}\varrho_{jx}$
$\overline{\vec{c}\delta}_{\underset{k\leq n}{in\times k}}$
Group-Level/Spectral Modularity
$max \sum_{j=1}^{n} \varrho_{jx}$
s.t.
$\varrho_{jx} = \{\varrho_{1x}, \varrho_{2x}, \varrho_{3x}, \dots, \varrho_{jx}\}$
$\xi_{jx} = \{\overline{\vec{c}\delta}_{\underset{k\leq n}{in\times k}} \cup \delta_{jx} | \overline{\vec{c}\delta}_{\underset{k\leq n}{in\times k}}, \neq \delta_{jx}\}$
$\mathbb{C}\varrho_{jx} = \delta_{jx}(\overline{\mu^{Q}_{jx}})$
(5),(7)
Terminate ?
No
yes
x=x+1
Outcomes
Suspending Malicious Sets of Users
Validation
DCFM
Control, Power, Interest
Girvan-Newman Method
Depth-first Search Method
Get Malicious Sets
(FSA)
$F = \{\mathbb{C}_1, \mathbb{C}\varrho_{jx}, \mathbb{C}\varrho_{j+1x}, \dots, \mathbb{C}\varrho_{kx}\}$

Figure 6. Algebraic transformation of the model to identify key network groups coordinating deviant acts grounded in Collective Action theory. (Alassad, Agarwal et al., 2020) Journal of Information Processing and Management, Elsevier.

Figure 6 is an exemplar that demonstrates how algebraic transformation helps in model interpretability. The objective of the study is to identify key network groups that coordinate nefarious acts. A model is developed that is grounded in mathematical definitions of collective action theory. The mathematical representation of the model helps explainability and in turn, interpretability.

The model is also more accurate compared to the state-of-the-art approaches as shown in the results in Figure 7.

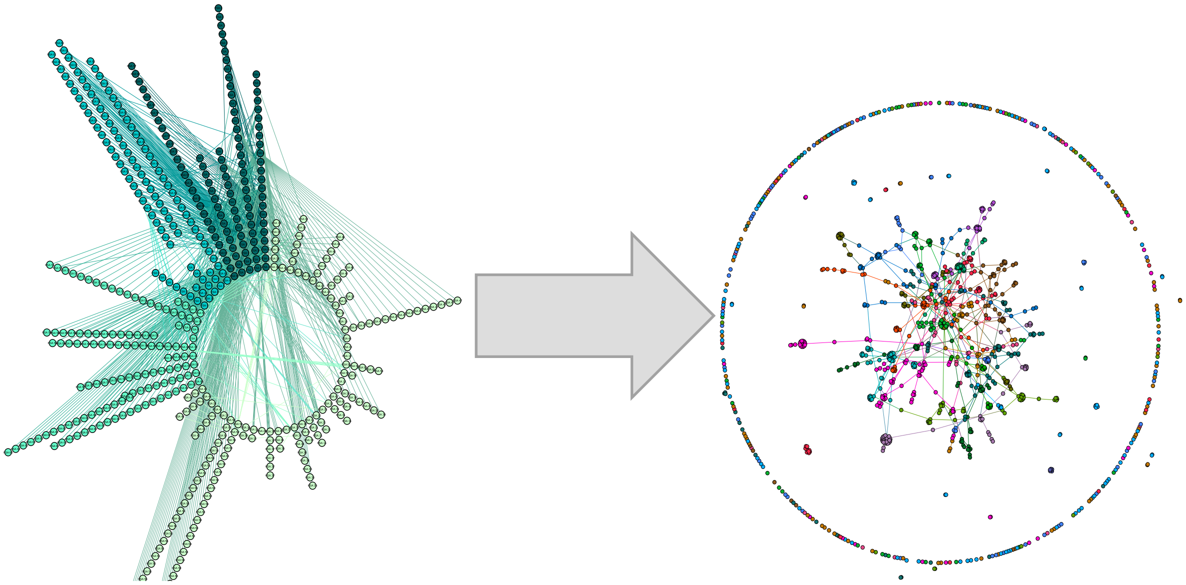


Figure 7. Before ISIS recruitment (hub an spoke) (left). After ISIS recruitment network disintegrates (right).

The results show that by removing key groups from an ISIS recruitment campaign network, how the campaign coordination efforts disintegrate. The research is published in the Journal of Information Processing and Management.

DART researchers (SA5) are investigating how misinformation spreads in an online social network. We developed a model grounded in epidemiological concepts that divides individuals in a social network into susceptible, exposed, infected, and skeptic categories. Transition between categories is modeled using differential equations, thus enhancing explainability and model interpretability. To test this model, we compiled Twitter data surrounding events on June 1st, 2020. Thousands of social media users shared #DCblackout claiming that Washington DC had experienced an outage of internet and cellular communications. The rumors came during widespread protests across the United States over the death of George Floyd, whose death while in police custody was filmed and widely shared. . The model was able to explain the spread of misinformation quite accurately with an error rate of less than 2%. The research has implications to both social media platforms and policymakers by accurately identifying which individuals are susceptible, exposed, infected, or skeptic. These studies were published in the European Conference of Operations Research this year.

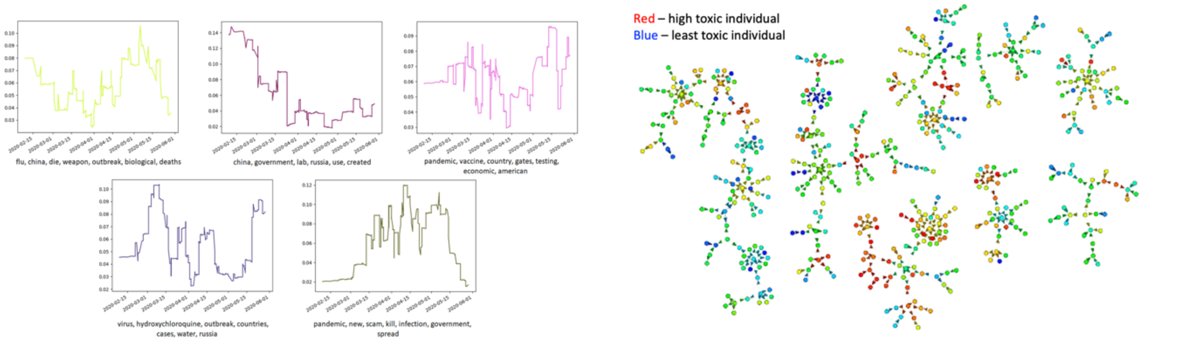
Admittedly, algebraic transformations are still hard to interpret. Therefore, DART researchers working in SM, LP, and SA have also approached the challenge of model interpretability by combining algebraic transformation with visual representation. In one example, we modeled coordination tactics of bot accounts on social media platforms like Twitter. While these botnets can disseminate humorous misinformation like sharks swimming on freeways during hurricanes, they are also known to disrupt disaster relief operations for minority communities by suggesting, for example, that the migrants are required to show formal documentation of their status before seeking help in storm shelters. In another example, US intelligence agencies released the Russian IRA botnet that played a significant role in polarizing online discourse during the 2016 U.S. federal elections by amplifying hate speech on social media platforms. Our team developed a decision tree-based model to identify coordinating botnets. Each branch of the tree helps in making the decision to assess a botnets coordination capacity, thereby making the decisions explainable - visually - and the model more interpretable. The study is published in the Social Computing and Behavioral Modeling conference proceedings.

One of the major barriers to adoption for machine learning and AI based systems is the lack of transparency. Model interpretability can help in increasing model transparency, thereby lowering the barrier to adoption. AI-based recommendation algorithms that predict our shopping behaviors, books and articles to read, videos to watch, etc. often lack transparency. Recommendation algorithms learn from behavioral data and perpetuate underlying bias.. There are several cases known to us. For instance, YouTube's recommendation algorithm is known to push its viewers down the conspiratorial rabbit hole by suggesting related videos. On Facebook, ads to recruit delivery drivers for Domino's Pizza were disproportionately shown to men, while women are more likely to receive notices in recruiting shoppers for grocery delivery services like Instacart. An interpretable model would help in identifying causes of biased recommendations, thereby enhancing the model’s interpretability. Developing methods to detect bias in algorithms can serve as a clarion call for transparency and interpretability.

In this context, DART researchers analyzed eight different contexts, multilingual and multicultural, on YouTube to study bias in its recommendation system. We analyzed 100,000 recommendations and found that there was topic drift and a decrease in relevance in the results or the recommended videos. More strikingly, top recommended videos were removed weeks or months after YouTube had recommended them. Violation of platform terms and services were stated as the reason for content removal. This study was published in the European Conference on Information Retrieval. We know that adversarial actors have begun to exploit algorithmic bias, whether these are propaganda or disinformation agents. With the advent of adversarial machine learning algorithms like generative adversarial networks (GANs), advancing research in model interpretability is more critical now than ever before. Just as it is important to detect bias in algorithms, it is equally helpful to study the algorithmic bias characteristics at a finer granularity.

Knowing which factors contribute more to algorithmic bias allows greater transparency and better model interpretability (figure 8). DART researchers are in the process of analyzing data collected from Amazon to quantify bias and unfairness of marketing strategies vis-a-vis consumer demographic attributes such as marital status, age, education status, occupation, hours worked per week, gender, country, ethnicity, etc. The bias measurement approach, known as the Shapley Additive Explanations (SHAP) helps in explaining the output of a machine learning model in terms of its sensitivity to different attributes. Are we able to explain decisions generated by the model effectively? Does it increase confidence of decision-makers? We approach the answers to these questions by working closely with domain experts or decision-makers.

The COVID-19 pandemic presents a unique example of emerging cyber social traits. While there are similarities with other misinformation campaigns, COVID-19 misinformation campaigns have nuances such as global and regional narratives, as well as high topical diversity from health, policy, religion, geopolitical effects, volume, velocity, veracity, and variety of false narratives. The COVID-19 misinformation tracker tool developed by Agarwal in collaboration with the Arkansas Office of the Attorney General, supports detection, investigation, and mitigation of cross-platform COVID-19 misinformation campaigns and scams to assist policy makers.



COVID-19 misinformation themes pushed on multiple platforms (blogs, twitter, YouTube, facebook, and other non-mainstream social media platforms). Recurring/periodic themes such as vaccine hesitancy, alternate medicines, conspiracy theories, etc. allow proactive communication strategies and policymaking to detect and mitigate emerging cyber-social threats.

Toxicity analysis on YouTube commenters. Toxic discourse causes disruption and polarization/segregation among communities, as seen above. We demonstrate that by removing highly toxic users from a network, hate speech reduces, online discourse improves, and fractured communities heal. Our findings offer guidance to policymakers within each online social network to make informed decisions about the information environment and derive appropriate and timely countermeasures to continue providing a healthy platform for their users.

Figure 8. Predictive behavioral models to assist policymaking and crisis communications.

Our efforts demonstrate that when researchers coordinate with policymakers, it can make a difference, especially when the coordination remains an ongoing process. Cross-platform narratives are detected using developed social computational models as shown. Education component of the effort provides tips to self-detect misinformation. People can notify us of scams and other misinformation cases that are not in our databases, which we then investigate. Daily reports are shared with the AG's office as well as on the website, along with investigation results and policy recommendations to enhance outreach and awareness. We observed an increased penetration of awareness into rural areas of Arkansas where corrective information may not be available in a timely manner, making the population more vulnerable to misinformation. This demonstrates the effectiveness of the model's interpretability as well as the decisions made there. The COVID-19 misinformation tracker system is publicly available and serves the mission of bridging science and society through technology. Predictive models developed using the training sample from the COVID-19 misinformation tracker system help in detecting recurring, periodic, or seasonal themes, as can be seen in the image on the left: alternate medicines, conspiracy theories, economic stimulus scams, vaccine, hesitancy, et cetera. This allows proactive communication strategies and policymaking to detect and mitigate emerging cyber social threats. Similarly, toxicity analysis on YouTube commenters showed disruption and polarization or segregation among communities as seen on the image on the right. We demonstrate that by removing highly toxic users from a network reduces hate speech and improves online discourse, which ideally will allow fractured communities to heal. Our findings offer guidance to policymakers within each online social network to make informed decisions about the information environment and derive appropriate and timely countermeasures to continue providing a healthy platform for their users. To summarize, DART team has laid solid foundations in studying model interpretability in year 1 that will be enhanced in subsequent years.

5.3. Privacy and Security

When people engage with data science products on their smart devices, on their social media platforms, and on the web, they trust their data and information they receive based on largely invisible robust privacy and safety protections. Higher levels of trust will contribute to broader use of data science to benefit society. The issues of trust harms and benefits are very real, especially given that every day we generate over 2.5 quintillion bytes of data. In this context, policymakers, consumer advocates, corporate officers, and data scientists voice the fears that the acquisition and subsequent applications of big data analytics is not secure or well-managed. The stakeholders are also concerned that harms outweigh benefits. For instance, a recent study conducted in 2019 by the Pew Research Center, noticed two things: first, 81% of Americans believe that the risks of personal data collection by companies or the government outweigh the benefits; second, 79% of Americans are concerned about their personal data used by companies. Further, 64% are concerned about the ways their data are used by the government, and 70% of Americans think their personal information is less secure today that in the past.

The study also shows important socio-demographic differences by race, age, the socioeconomic status in people's privacy, attitudes, and practices. While privacy and security concerns have multiple dimensions, personal data breaches have received much attention. Indeed, breaches affecting millions and now, even billions, of individuals have become common. In fact, the 2013 and 2014 Yahoo data breaches affected 3 billion people. Among the top 15 data breaches, the smallest incident affected over 130 million people. These data breaches compromised names, email addresses, passwords, security questions, dates of birth, passports, travel information, and of course, credit card information. In a more general sense, informational harms include economic harm, social harm, and legal harms. However, the benefits of shared data include the predictive analytics of possible pandemic outbreaks, pandemic tracing, hate speech detection, prevention of cross-media discrimination, and decision-making fairness. Consequently, our research teams have engaged in multiple research activities that enhance the highly beneficial functions of big data and big data science. The importance of these benefits is highlighted in the ongoing work undertaken by the social awareness research team.

One of our key efforts is to improve privacy and security of big data. During the first year of the grant, we began addressing this barrier by engaging in two general types of activities. First, the social awareness research teams have been working on improving practical privacy preservation and security enhancement strategies and methods. Second, we had been collecting empirical interdisciplinary studies examining the key concerns articulated by various data science stakeholders.

Researchers in SA1 completed a theoretical investigation of privacy preserving mechanisms for deep learning. Specifically, they surveyed existing attacks on deep learning models and categorized them as evasion, poisoning, and model stealing. In this regard, the SA1 team focused on adversarial examples, model inversion, and membership inference attacks. They examined four aspects of threat models: adversarial classification; adversarial knowledge; adversarial specificity; and attack frequency. They also researched representative approaches for each attack type. For instance, for the attack type of adversarial examples, the SA1 team examined the representative approaches, finding that most approaches are based on constraint perturbation with stochastic gradient descent.

The SA3 team completed an extensive literature review examining definitions of personal identification information from different perspectives and the associated privacy issues. Specifically, PII attributes are often differentiated only by their semantics as public or protected. In this context, the SA team has been working on assessing the cumulative sensitivity measure of the leak PII attributes. Currently, they are developing a matrix that considers both the sensitivity level of each PII attribute and the combined sensitivity of a given set of leaked attributes. SA6 also completed a literature review of privacy preserving algorithms and software. They ascertained that the privacy preserving analysis in genomics and health is broadly connected with other aspects of privacy preserving approaches, including data formats, infrastructure support, and algorithm design and development. This team has successfully disseminated information about the advantages and limitations of current approaches. They have also initiated investigation of mathematical optimization models in privacy preserving analyses in genomics and health.

Finally, the SA7 team completed an analysis of existing work that uses cryptography techniques for privacy protection and federated learning. Toward this end, they reviewed about 30 published papers. They have found that only three studies have combined cryptography with differential privacy to provide more advanced privacy protection. This team has also designed a new cryptography-based scheme for differentially private federated learning. Notably, this scheme reduces the communication cost in the training process and performs better than existing work in convergence rate and learning accuracy while providing differential privacy. In some, our colleagues approached privacy preservation from various perspectives and have already generated outcomes that improve privacy preservation and security technologies, techniques, and methods.

Across all DART themes we are advancing other aspects of socially aware data analytics. The goal is to understand and address the feedback loop between data collection technologies and the concerns their users have regarding possible harms. In addition, we must understand people's attitude towards, and perceptions of, benefits associated with big data and data science. This knowledge will assist us with communicating with users in a targeted and efficient manner. Educating socially aware data scientists and engaging in fair and socially aware policymaking. Other researchers have been conducting a social science research on the social dimensions of privacy concerns, including privacy, attitudes, behaviors, and paradoxes. Specifically, our previous collaborative research on privacy preserving mechanisms, we noted that the magnitude of references to privacy concerns in the big data literature is enormous. In fact, it is difficult to find publications discussing privacy preservation techniques that do not make a reference to privacy concerns. While the data from the Pew Research Center certainly support the notion that privacy concerns do exist, we do not have a robust understanding of what the specific concerns are and whether they translate into specific behaviors. Also, not much is known about which factors contribute to these concerns or which social groups - including members of legally protected categories based on race and gender - are more likely to be concerned with privacy. Nor do we fully understand which types of privacy really matter to different user groups. Hence, we are also studying the social side of concerns and attitudes toward data privacy. Specifically, we analyze extent empirical studies of people's practical acceptance and utilization of data analytics. And we also inquire about the social factors that affect behaviors and attitudes.

To accomplish this goal, we've been conducting a systematic literature review using two well-tested scientific protocols a spider search protocol, which is best suited for social science systematic literature reviews, and the PRISMA framework, which is an evidence-based protocol for literature review reporting and meta-analyses. Using the spider protocol, we have identified 1,324 interdisciplinary and publications. We have initially cleaned our data by removing duplicates, conference papers, books, book chapter, et cetera. Next, using the Rayyan software, we reviewed abstracts of 1,116 articles and identified 106 full texts publications for in-depth analysis. Also, since not all abstracts provided sufficient information for inclusion-exclusion decisions, we have identified another 124 articles for full text review. At this point, the designated set of articles for analysis consists of 129 articles, including 119 articles published between 2010 and 2020. We have now begun analyzing this literature using the NVivo qualitative data analysis package.

5.4. Education, Outreach, and Workforce Development

The major accomplishments for the Education Program have been to create an overall architecture for post-secondary Data Science education for the State of Arkansas, defining the approach and milestones to move from architecture to design, defining the approach and milestones to move from design to implementation, identifying “cohorts and waves” of participating academic institutions throughout the State, and beginning the pilot implementation with the first wave in the first cohort. The significance of this is that we have an agreed-upon approach, supported by the academic institutions (4-year and 2-year), the Arkansas Division of Higher Education (ADHE), the Arkansas Economic Development Commission (AEDC), the Arkansas Center for Data Sciences, and the Office of the Governor. The unique approach is based on a hub-and-spoke model where four 4-year Universities provide the “hubs” with a common core curriculum with a “Venn Diagram” of Concentrations that include special focus on regional industries with spokes of 2-year and 4-year colleges with Associate Degrees and 2+2 (and 2, then 2) programs with the first two years’ courses designed and developed to match those (including potentially using the same syllabi) of the hubs. The overall objective is to provide a high-quality, consistent, inclusive, and rigorous system to develop an educated workforce ready-to-work throughout the State, optimizing valuable teaching and laboratory resources, and minimizing cost-wasting duplication and the risk of lower quality programs.

The project is largely on schedule, but the COVID-19 pandemic has slowed the progress overall due to the inability to meet in-person. This has impacted both operations of the project team and the response time at individual campuses. In the start-up phase, high-bandwidth, in-person collaboration, workshops, and training are key to success. As this report is being written Arkansas is reducing restriction on meetings and campuses around the state are announcing plans to return to normal operations with face-to-face meetings in fall 2021. Thus, in Year 2, operating conditions will allow us to combine those areas of Year 1 that are lagging in an accelerated manner with the Year 2 activities and milestones to better track our original plan. The challenges experienced were as expected and are documented in the COVID-19 impact analysis.

We have made steady progress in our plan to employ a wide range of professional development and data science education activities to engage K20 learners in Arkansas. Our vision is that Arkansas will have a statewide educational ecosystem for learners of any age so that they can receive a designed, consistent, and scaffolded education in data science with further educational opportunities and job opportunities at appropriate points in their careers. To accomplish this, our mission is to create a model data science and analytics program for Arkansas schools that will promote problem-based and experiential based pedagogy in critical thinking and analysis, technology familiarity, and a foundation in math and statistics.

Arkansas is home to several Fortune 500 companies that rely heavily on data analytics. In 2017, Arkansas Governor Asa Hutchinson commissioned a panel of industry leaders in Arkansas to provide recommendations on advancing economic competitiveness of data analytics and computing in Arkansas. Representatives from Walmart, Tyson, JB Hunt, AT&T, Murphy Oil, Acxiom, Stephens, EZMart, FIS, and others convened to assess the supply and demand of data skilled workers. The report showed that data analytics job demand is increasing rapidly in Arkansas and nationwide.

Arkansas students are required to have completed six weeks of coding by eighth grade, and teachers are struggling to find curriculum to fill a nine week schedule block. We're working with the Arkansas Computer Science Teachers Association and Coding AR Future to develop a full nine-week coding curriculum for middle schools, which will be available for teachers nationwide, and professional development offerings for Arkansas teachers to learn how to implement the curriculum. For the post-secondary ecosystem, we've planned a hub and spoke model with four campuses serving as regional hubs for the state's primary undergraduate institutions.

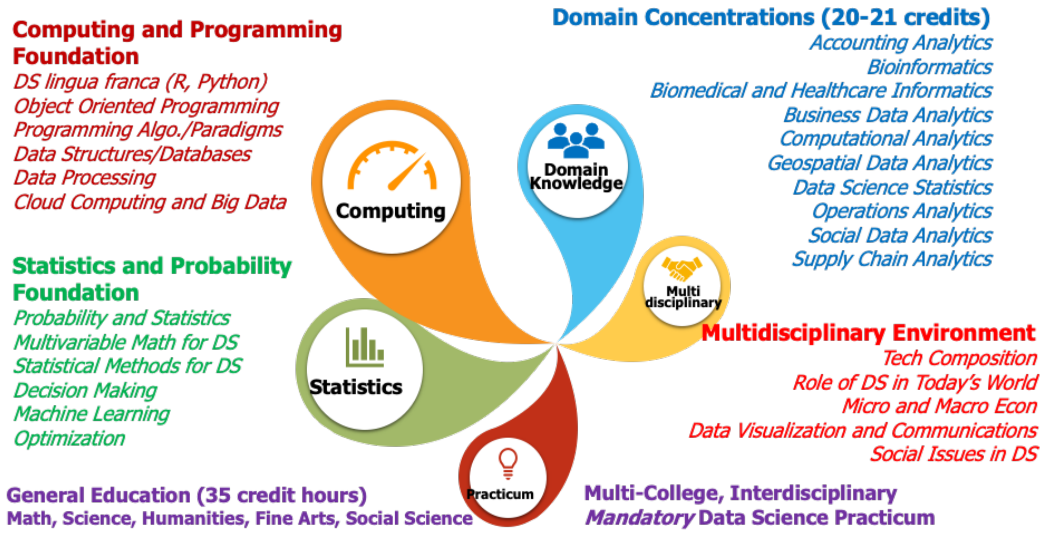

Computing and Programming Foundation
DS lingua franca (R, Python)
Object Oriented Programming
Programming Algo./Paradigms
Data Structures/Databases
Data Processing
Cloud Computing and Big Data
Statistics and Probability Foundation
Probability and Statistics
Multivariable Math for DS
Statistical Methods for DS
Decision Making
Machine Learning
Optimization
General Education (35 credit hours)
Math, Science, Humanities, Fine Arts, Social Science
Computing
Statistics
Domain Knowledge
Multi disciplinary
Practicum
Domain Concentrations (20-21 credits)
Accounting Analytics
Bioinformatics
Biomedical and Healthcare Informatics
Business Data Analytics
Computational Analytics
Geospatial Data Analytics
Data Science Statistics
Operations Analytics
Social Data Analytics
Supply Chain Analytics
Multidisciplinary Environment
Tech Composition
Role of DS in Today's World
Micro and Macro Econ
Data Visualization and Communications
Social Issues in DS
Multi-College, Interdisciplinary
Mandatory Data Science Practicum

Figure 49. UA Model curriculum.

We've developed a model curriculum that's already been implemented at the University of Arkansas at Fayetteville as a Bachelor of Science in Data Science with various concentration options (figure 9). We're working with as many campuses as possible in a “cohort and wave” sequence to use the model curriculum to develop customized, but consistent, data science programs. We work with industry partners in specific regions to make sure that the local institutions are producing graduates with the skills that the companies need. This will create a positive feedback loop for employers and educators. Figure 9 shows the breakdown of the UA curriculum model; it is based on 35 hours of general education followed by foundations in computing, probability and statistics, and multidisciplinary environment. The students can then select a domain concentration and 20 to 21 hours of specialized course content. They're also required to complete coursework on the multidisciplinary environment that data science resides in and complete a full year multi-disciplinary, multi-college practicum. Campuses can opt-in to the DART Education Project. Our team works with them very closely and in complete transparency to share all curriculum and pedagogy materials. Together, we will tailor the curriculum based on the existing strengths and needs of that campus to create a bachelor's or associate's degree and/or a technical certificate.

We've received a special fast-track authorization from ADHE for any programs developed in collaboration with us and using the UA curriculum as a base. This will enable a statewide ecosystem that ensures full transferability among institutions - a start anywhere finish anywhere model - with the option of 2 +2 or 2 then 2. Students could start a two-year program, work awhile, and then finish back at another institution to get a bachelor's degree. Arkansas recently implemented a similar 2+2 model with some of the participating campuses for a BSN in Nursing, which is providing some useful experience for our team. We also will build the programs in such a way that the campuses will be poised for accreditation through ABET or the private college equivalent and will help them through the accreditation process once graduates are produced. The first cohort includes three of the State's four HBCUs, as well as community colleges and two other four-year campuses. As we develop the programs for each of the participating campuses, we hold pedagogy workshops to establish a train-the-trainer pipeline of faculty with skills to teach the courses we develop together. Our first pedagogy workshop will take place in June with the faculty from the first cohort taught by UA faculty. Next year, the first cohort or participants will teach these topics to the second cohort participants while UA faculty train the next wave of faculty from the first cohort. We hold three workshops per year devoted to the educational component and the attendees include faculty, administrators, IT support, and CIOs from the campuses all over the state. Through these forums, we've established a process for identifying the various needs of each campus. We put together a series of surveys to help each campus figure out their infrastructure, hardware, and software gaps. Another survey polled local employers for the exact skills they're looking for. The institutional needs survey has been distributed statewide, and we are still receiving submissions. This survey addresses connectivity, hardware, software, and IT on each campus.

Most campuses in Arkansas are connected to our research and educational optical broadband network, also known ARE-ON, for middle-mile Internet, but the connection speeds and the port sizes vary. Even on some of the larger campuses, Wi-Fi is neither ubiquitous nor robust. Many of the smaller campuses are lacking physical infrastructure, such as computers and computer labs, to support the data science courses. Many of the faculty are not up to speed on Python, R, or other skills taught in the curriculum. The employer needs survey was distributed in summer 2021 from AEDC and is based on a similar survey initially developed by ADHE. The survey asks a series of questions about how many candidates they're looking for and what skills and education level they want from the candidates, as well as any support their existing employees might have to take data science courses. We'll use the response data to inform the curriculum and programs on a regional level. ADHE has endorsed the state's academic institutions using the results of this survey in their program RFPs to avoid needless duplication. We're also collecting information on specific needs for each course, including which operating system is used; the type of computers that are available or being used; and developer environments and related packages used in the courses.

In order to engage and support a broad group of participants, DART has planned a number of K20 professional development initiatives. We're providing student research assistantship positions, working with industry partners to connect with students with internships, funding under-represented minority students to participate in summer undergraduate research experiences, and hosting the annual Arkansas Summer Research Institute. We are funding mini-seed grant projects from community partners and K12 schools and we provide a number of technical skills training and career development workshops throughout the year for all of our participants. We're implementing two avenues of professional development for K12 educators. One is focused on the middle school coding block described earlier. The other is a large-scale program in partnership with EAST Initiative, a local non-profit organization. They have a large footprint, they are in over 260 schools in four states, and they serve thousands of educators and tens of thousands of students. Over the next four years EAST will train a total of 448 teachers though 32 workshops. Training will focus on two areas: Student leadership and Integration of 3D printing, coding, and virtual reality technology in the classroom.

We've also developed a three-pronged approach to support our post-secondary faculty to receive technical and pedagogy training in data science related fields.

We offer at least three career development workshops throughout the year that cover a range of topics from technical skills, science communication, mentoring, and grantsmanship. They're free and open to the EPSCoR community to attend. We had a great workshop on communicating science to policymakers in April. We had a workshop with Mike Morrison on the #BetterPoster to teach our students how to use the #BetterPoster format. We hosted Dr. Barbara Bruno from Hawaii to learn more about individual development plans and mentorship assessments. We plan to have workshops on software carpentry, deep learning, and NSF grants this summer. Some past examples include NIH and DOD grantsmanship, SBIR, STTR, iCorps, and entrepreneurship. We are financially supporting 40 graduate students and 15 undergrads annually through the research assistantship program. We also set aside $80K per year to fund summer research experiences for undergraduate students who belong to groups that are under-represented in CISE-related disciplines. We're leveraging our extensive industry relationships and industrial advisory board to secure at least five internships per year for our students. We also hold regular informal student forums where we update the students on project progress, celebrate accomplishments like submitting a dissertation proposal, graduation, or getting published and share pet and baby pictures.

In 2021, we're hosting the 7th annual Arkansas Summer Research Institute. This is a short, intense professional development experience for undergrads. In the past, it was held in person on the campus of our partner, the Arkansas School for Mathematics, Sciences and the Arts, or ASMSA, a residential high school for academically talented students. In 2020 due to the pandemic, we shifted to a virtual space and decided that while it did have some limitations, it actually gave us more room to expand the length of the event, invite more presenters, and engage more students. This year will be the second virtual ASRI, and we're expecting around 100 students to attend. We also discovered that it's not just undergrads who want training in data science, so this year we're accepting select high school juniors and seniors, graduate students, and even allowing faculty to audit and participate. . The experience level and background of the attendees is diverse, so we have some plenary sessions and some breakout sessions that are broken out either by discipline or experience level. The technical sessions are taught by DART faculty. We also have a few panel discussions, like data science in today's world where students will hear from a variety of industry representatives and how they used data analytics in their fields. There's also an entrepreneurship panel with local STEM entrepreneurs discussing their journeys. A session on science communication with Mike Morrison of #BetterPoster fame, and a session on equity and inclusion in research.

6. Summary of Significant Problems, Novel Opportunities, and Changes in Strategy

Overall, there are now significant problems facing the program other than maintaining a cohesive experience for faculty and students in the midst of pandemic-mitigating directives. However, as these hopefully ease in the fall and face to face meetings resume, we expect to achieve more dialog that will lead to new opportunities and possible changes in strategy.

In the start-up phase, high-bandwidth, in-person collaboration, workshops, and training are key to success. In Year 2, we will combine those areas of Year 1 lagging in an accelerated manner with the Year 2 activities and milestones to better track the original plan. The challenges experienced were as expected and documented in the risk mitigation analysis.

Access to computer infrastructure has emerged as a barrier to the implementation of the statewide education program. This includes access to high-speed internet between campuses and low-bandwidth connections on campus. Additionally, not all campuses use imaging technology to deploy software packages to their computer laboratories and faculty may not have administrative rights that allow software installation. Thus, in addition to sharing curricula we will begin to share best practices for administering the deployment of software across campuses. This addresses part of the problem, however, while DART does include some hardware at large campuses, it does not include hardware support at the state's PUIs and private colleges. All public institutions are connected to The Arkansas Research and Education Optical Network (ARE-ON). ARE-ON provides 1Gb, 10Gb, or 100Gb ethernet connections to its members. While private institutions are eligible, most have not yet joined ARE-ON and so are unable to access the state's fastest connections. Thus, even though the majority of campuses have the appropriate high-speed broadband they do not have the campus infrastructure (including Science DMZs, communications, protocols, and client-server systems) to support access to HPC systems at UAMS and UA. To address this problem DART personnel and others have written and submitted a proposal to support DART activities to NSF under OAC Campus Cyberinfrastructure. This proposal entitled *CC* CIRA: Shared Arkansas Research Plan for Community Cyber Infrastructure (SHARP CCI)* has UA Research Computing Director Donald DuRousseau as PI and DART PI Jackson Cothren as a co-PI. Dr. Addison represented the Education Group in the development of this proposal and is identified under senior personnel. Additional infrastructure support is being investigated to provide classroom tools at campuses where need exists.

To better understand the current state of facilities, faculty, and instruction at the state's academic institutions, Drs. Schubert and Addison and Ms. Fowler developed an "Institutional Needs Assessment" survey that is being distributed to all state institutions of higher education to understand better the needs as described above. This also includes information on instructional capabilities. A second survey, "Software and Package Use," has been developed to identify the software-in-use from the faculty perspective and the student perspective. This survey has been distributed to the DART Education team and will subsequently also be distributed to all post-secondary institutions enabling us to evaluate readiness and to development of a gap analysis. This survey will also assist in the skills development, training, and software planning for the institutions as they join the cohorts and waves.

7. Research and Education Program

7.1. Coordinated CyberInfrastructure

What is Coordinated CyberInfrastructure? Coordinated CyberInfrastructure is a the underlying support for the development, optimization, and management of analysis pipelines from each of the research themes. This can include containerized pipelines for image curation, genomics analysis, machine learning, and much more.

A newly established CI working group (CWG), chaired by James Deaton, executive director of GPN and composed of the GPR CyberTeam CoPI has been organized by UAF Information Technology Services. The CI Working Group will advise ARCC in filling the gaps identified in the initial analysis and in ongoing analysis.

The administrators, engineers, and researchers listed below are all involved in the design and implementation of the network architecture required to make the ARP available of more members of the jurisdiction.

James Deaton	Executive Director, CyberTeam Co-PI, Great Plains Network
Kevin Brandt	Director of Research Computing, CyberTeam Co-PI, South Dakota State University
Brian Berry	Administrator, MT, UALR
Jan Springer	Director of Emerging Analytics Center, DART CI Theme Co-Lead, UALR
David Merrifield	Interim Executive Director, ARE-ON
Scott Gregory Ramoly	Chief Technology Officer, ARE-ON
Guy L Hoover	Manager of Network Engineering, UAMS
Shawn Bynum	Director of Unified Communications, UAMS
Stephen Cochran	Chief Information Security Officer, UAMS
Matthew Reiss	Network Capacity Engineer, UAMS
Eric Wall	Assistant Director of IT Security, UAMS
Fred Prior	Chair of Bioinformatics, DART CI Theme Co-Lead, UAMS
Stephen L. Tycer	Chief Information Security, UAF
Elon T. Turner	Network Director , UAF
Lisa Richardson	Director, Project Management Office, UAF
Michael E. Davis	Network Architect, UAF
Nick Salonen	Senior Information Security Analyst, UAF
James McCarthy	Project/Program Manager (Enterprise Services), UAF
Don DuRousseau	Associate CIO for Research , UAF
Jackson Cothren	Director AHPCC, DART CI Theme Co-Lead, UAF

The CyberInfrastructure (CI) Plan for DART was recently accepted and outlines the organizational structure, roles, and responsibilities of the CWG and ARCC, as well as how these groups interact with the DART CI Research Theme. The CWG and inclusive subgroups have been meeting monthly since October 2020 to discuss current network configurations, rational for requested changes to that network, and the security impacts of the changes. One of the subgroups has specifically addressed plans to manage CUI data at UAF in accordance with NIST SP 800.171.

The CWG will assist the ARCC and DART to eventually address all five recommendations listed below and within the DART CI Plan. However, the makeup of this particular group is targeted to immediately address recommendations 1, 2, and 3. The CWG will report to the Scientific Steering Committee (SSC) directly and through the CI Research Theme.

- **Recommendation 1:** Monitoring and measuring the capabilities of the current state of the network and as adjustments and upgrades are introduced needs to be implemented. Each of the individual universities and ARE-ON have practices in place to collect network telemetry but aspects need to be coordinated to provide a thorough view of the state of the network for ARP-related activities. In addition to the telemetry, additional perfSONAR nodes need to be deployed to assess the performance of the network and to gauge the impact of network changes. Such deployments need to be a coordinated activity to assure consistency in the testing processes and assure efficient operation and stable measurement archives.
- **Recommendation 2:** The importance of federated identity practices grows as resources via ARCC are utilized across institutional boundaries. As resources are consumed (and shared) via the broader goals of ARP across state borders and nationally with the GPN Research Platform, Pacific Research Platform and XSEDE, InCommon membership and practices become very important. The state of federated identity at the institutions is mixed with only UAMS operating as an InCommon identity and service provider and none of the institutions registered as Research and Scholarship adopters. Efforts to address this should be guided by InCommon's Baseline Expectations for Trust in Federation Version 2 and REFEDS Research and Scholarship practices.
- **Recommendation 3:** Review of the ARCC resource providers data controls included requests for documentation regarding NIST 800171-related System Security Plans as well as regulatory compliance efforts associated with HIPAA and FERPA. Responses were mixed with nominal effort underway addressing university efforts toward dealing with CUI. The breadth of research to be addressed within DART, the diversity of participating institutions and the broader impacts of addressing a strategy for regulatory compliance make this project an intriguing potential engagement opportunity for Trusted CI. An upcoming engagement application window should be leveraged to garner the insight of this NSF Cybersecurity Center of Excellence to identify opportunities to address security and compliance aspects of the project. In addition, all the ARCC resource providers, ARE-ON and several of the other institutions are RENISAC members. REN-ISAC provides a peer assessment service which has recently shifted from an in-person endeavor to working remotely. It can also provide substantial insight in this area.
- **Recommendation 4:** As resources are shared across institutions, local facilitation will play a valuable role. The XSEDE Campus Champions program provides a structure for the identification and a community of support for individuals serving in these roles. Aside from UAF, there are only 2 other Champions identified within these institutions participating in the project, one at ASU and one at UALR. Individuals who will interface with researchers more directly need to be identified. The outreach and support of mentors within the GPR CyberTeam will work with these individuals to help identify more granular gaps in the CI as the project progresses.

- **Recommendation 5:** Of the universities involved in DART, only UAF and UAPB have received funding from the NSF CC* program. All the other institutions remain eligible for funding within the program's area 1 and/or area 2. Significant improvements in research CI should be funded through this program and can occur in parallel with other activities within DART.

Challenges are scattered throughout these recommendations. The most prominent challenge we face in year 1 is developing a secure Science DMZ that serves researchers but protects sensitive data and campus enterprise systems and networks. This requires close collaboration between campus IT leadership and the campus research communities. Design and implementation of secure Science DMZs is evolving to meet security needs and campus CIO's and CISO's new to the concept are eager to better understand not only the how, but the why of the concept. The organizational structures described in section 1.3 are intended to bridge the gap between enterprise IT and research computing needs. Furthermore, DART will take advantage of exiting NSF resources to learn and leverage current best practices in the design of coordinated Science DMZ's. DART applied for and received an engagement award with the Trusted CI program (# 1920430 CICI: CCoE: Trusted CI: Advancing Trustworthy Science) and engaged with the CyberTeams program (#1925681 CC* Team: Great Plains Regional CyberTeam) early in the program. The CyberTeams collaboration is described in some detail below. As of the submission of this report, the Trusted CI engagement was underway.

The CI Research Theme is also working through the CWG to stage a series of proposals over the next two years to coordinate network improvements and further leverage DART funding. The first of these proposals was submitted in March 2021, led by UAF Director of Research Computing, and submitted in collaboration with CoPIs from UAMS, UCA, and UALR. While DART is not dependent on this proposal it would be advanced the award. The purpose of this CC* CIRA: Shared Arkansas Research Plan for Community Cyber Infrastructure (SHARP_CCI) proposal is to develop a statewide CI plan for Arkansas that focuses on eight (8) degree granting institutions performing science and engineering research on campuses across the state. Each school has a growing demand for federated access to high-speed networks, shared storage arrays, high-performance compute clusters, technical training and managed support services, and a coordinated plan for providing these capabilities and services does not currently exist.

Create UAF CI Plan. In collaboration with the CWG, UAF developed a CI Plan that is consistent with both the UAMA CI Plan and the recommendations made by the CI review process. This plan, along with the UAMS CI Plan, will be published and used a guide for other institutions using ARP resources.

Issue UAF purchase order for additional equipment. Two quotes (one each from HPE and Dell) have been received and are currently under legal and technical review.

- Target date for installations of the additional nodes is June 1, 2021.
- Specifications (DART purchase only):
- 20 nodes dual AMD 7543, 1024 GB, NVMe local drive, single PCI 40 GB A100 GPU;
- 4 nodes dual AMD 7543, 1024 GB, NVMe local drive, four SXM 40 GB A100 GPU;
- 100 Gb Inifiniband connection and 10 Gb Ethernet connection;

- 3 Enclosed cooled racks.

This purchase will provide approximately 68 Teraflops of CPU (13.6% increase of current capacity at UAF) and 350 Teraflops of GPU (50% increase). This purchase also served as catalyst and incentive for non-DART researchers by providing an opportunity to "piggy-back" on the large purchase order and receive substantial discounts for an additional 200 Teraflops (~$400K) in condo node capacity. The new nodes represent a significant addition - not only in pure Teraflops - but in the ability for researchers to run interactive sessions and operations on very large datasets.

Collect testbed specifications and software/platform needs. Target date for beginning the installation of the additional nodes is June 1, 2021. The specifics of this purchase are such that they are easily integrated into the existing Pinnacle architecture.

Create UAMS CI Plan. In collaboration with the CWG, UAMS developed a CI Plan that is consistent with both the UAF CI Plan and the recommendations made by the CI review process. This plan, along with the UAF CI Plan, will be published and used a guide for other institutions using ARP resources.

Establish federated ID for all project participants. We anticipate that full InCommon registration as Research and Scholarship providers will take longer to implement at UAF, UAMS, and UALR. However, the CoLeads are working within their respective institutions to discuss how to address this problem. In the meantime, access to ARP will be available through the ScienceDMZ at UAF. An ARP-wide System Security Plan, along with cluster-specific Information Security Plans, are being developed in collaboration with the CWG. Approval of these plans by University IT services is expected by June 30, 2021. Funds budgeted in year 2 and year 3 ($80,000 total) for the Globus Standard subscription and Globus for Box subscription will be instead used during those years to support federated identify improvements (such as enrollment in InCommon) at various campuses that are needed to support the three major services being provided and promoted through the ARP.

Create and publish document outlining GitLab user guidelines, minimum standards for code repositories, and best practices. A Gitlab repository with a dedicated server has been implemented behind the ScienceDMZ at UAF. It is accessible by all participates and mergers with other, existing repositories, are expected to begin in May 2021. An example project will be added in June 2021 as a guide to all researchers on minimum standards and best practices.

Globus Data Management contract executed. A Globus Basic server (no contract required for Globus Basic) has been established behind the existing Science DMZ at UAF with endpoints at storage arrays at UAF and UAMS. The purchase of additional services has been delayed based on needs identified in the CI review and will be further reviewed during year 2. The recently approved CI Plan will support a budget modification that re-allocates funds saved by using the no contract Globus Basic service to supporting the ScienceDMZ.

Quick Reference Guides (QRG). QRG's are under development for Globus and GitLab. Example repositories are set up in the DART git repository. Two software carpentries workshops are planned for Spring '21.

Collect research theme needs. There are no current research initiatives underway that require HIPAA, propriety economic, or CUI information. However, we expect that as the project matures these types of datasets will be required for research. ARCC is working with the CWG to develop a System Security Plan (SSP) that will define security levels and govern use of this data at various institutions. In addition, Information Security Plans (ISP) are being developed for Pinnacle and Cluster. ISPs will guide how data is stored and protected on that system.

7.2. Data Life Cycle and Curation

Why focus on Data Life Cycle and Curation? The three most time-consuming data preparation processes are data cleaning, data integration, and data tracking (data governance). The vision for the research is a "data washing machine." People are accustomed to throwing their dirty laundry into the washer along with some soap, setting the dials for the type of clothes, and letting the washer operate automatically. A data washing machine would work in a similar manner on dirty data - simply 'throw in dirty data', push a button, and out comes 'clean' or curated data. Below is the description of progress to date on Year 1 milestones.

Define at least one metric for completeness, standardization, and clustering quality of unstandardized reference data. The assumption for the initial research on unsupervised data curation is that the data being processed comprises references to real-world objects such as customers, patients, products, parts, or locations. Furthermore, it assumes there is a relatively high level of data redundancy, i.e., many references to the same object. This case was selected because "multiple sources of the same information" is generally acknowledged as one of the leading data quality problems faced by organizations.

Based on this assumption, the first-year research has focused on an unsupervised clustering (entity resolution or ER) first approach to address data redundancy and to organize the data into clusters of references to same object. This is the reverse of the current approach to apply a supervised standardization (ETL) process to each source before applying a supervised ER process. Based on this model the following metrics can be defined

- **Completeness data quality metric:** Borrowing from the genomics, the reference tokens sequences can be used to find gaps that potentially represent incompleteness. Given a cluster of unstandardized references, suppose that one reference has the sequence of tokens ABCD, and another the sequence ABD. Given the tokens A, B, and D, match in order between the two references, C can be understood as a missing value in the second references. The percentage of missing tokens found across all clusters can be used as a completeness metric.
- **Standardization data quality metric:** This metric is very problematic and has not been addressed for the first year of the project. It is problematic because of the unsupervised ER-first approach which pushes standardization to the end of the process instead at the beginning in the supervised ETL approach. While there is not yet an algorithm, there is some discussion about position alignment of tokens similar content. For example, is the first token of a standardized reference is a person's first name, then the first tokens across all references should have "name-like" characteristics, and those that do not represent deviance from standardization. While concept could have some merit, it has not yet been developed into an

algorithm. This type of alignment analysis seems to suggest unsupervised clustering techniques might be used as both a metric and tool for standardizing the references.

- **Clustering data quality metric:** The current approach to evaluating the quality of a cluster of references is to use a modified form of Shannon entropy.

  $$EE = \underset{ii=1}{\overset{NN}{-pp_{ii}}} \cdot \log_2(pp_{ii})$$

  Given a cluster of R references, the algorithm takes the first token (T) from the first reference in the cluster and searches for a single instance of T in each of the other references in the cluster. Only one instance of T is counted in each cluster. Based on the number instances found (N), then the probability of T is N/R. As each token is counted, it is removed from further consideration. Once all tokens in the first reference have been counted, the algorithm checks to see if there are any remaining tokens to count in the second reference, and so on, until all tokens have been counted. The rationale for this measure is that in a perfect cluster, all R references would have an identical set of tokens, i.e., the references all share the same tokens. In this case, each token would have a probability of 1 and the overall entropy would be 0. As different references in the cluster have different tokens, the entropy increases representing higher levels of disorganization (non-uniformity) in the cluster.

  This cluster entropy metrics has been implemented in the Data Washing Machine (DWM) proof-of-concept described in Activity 1, of Objective 2.1.c, Automate Data Integration.

  The goal of clustering is to group the reference data according to the similarity so that the similar reference is clustered together. One metric is to total similarity for references in the same cluster in comparison with the similarity of different clusters. The silhouette coefficient contrasting the average distance to elements in the same cluster with the average distance to elements in other clusters, will be used.

Design and implement an unsupervised algorithm for each metric: The deep learning autoencoder/decoder model based on pretrained language model such as BERT will be used to transform data into vectors in high dimensional space. The lost function will be developed based on the silhouette coefficient to minimize the total intra-cluster distance and maximize inter-cluster distance.

Establish baseline quality using supervised methods for existing datasets:

- **Hyperparameter tuning for the Proof-of-concept Data Washing Machine (DWM) using Bayesian Optimization.** As the proposed Data Washing Machine (DWM) uses an array of parameters that affect the performance and are set before observing the data, the selection of values that guarantee good results becomes crucial. In this sense, we have treated these parameters as hyperparameters, and the search for the optimal setting is made possible. For this purpose, some test code was built on the Beta branch of the GitHub repository of the DWM to use Bayesian Optimization for the tuning of hyperparameters. Technically, Bayesian Optimization uses an iterative search (within a finite search budget) by mapping Gaussian Process Regression models to the input hyperparameters and finding the best next point to test through an acquisition function that balances exploration and exploitation of good regions in the search space. The current results show that for a finite search, the parameters used in the DWM are close to optimal and are not improved by this search for the proof-of-concept data.

- **Single-cell sequencing data.** Single cell sequencing has emerged as a powerful set of technologies for elucidating biological systems in detail. Several large-scale single-cell sequencing projects have generated a large volume of data. These datasets can potentially help us to understand the heterogeneity of complex diseases and offer better treatment. Meanwhile, new computational methods are necessary for addressing several challenges in single-cell data analytics. The single-cell RNA sequencing dataset contains substantial missing values due to low capture efficiency and stochastic gene expression. Recovering the missing values will be essential for downstream analysis.

  We are developing a deep learning-based data imputation model for recovering missing values in scRNA-seq. The metrics for model evaluation include MSE, purity, entropy as well as the effectiveness for cell-type identification.

  We also built and compared classification models using classical and deep learning algorithms for Alzheimer's disease prediction and biomarker identification.

  A classification downstream task model based on pretrained BERT language model will be developed to minimize the predictive loss that measures the difference between the model result with the ground truth label.

Compare results of unsupervised quality metrics developed in Activity 1 to supervised results: Using compression rate as a proxy for data quality. As the data quality assesses in some way a lack of uniformity in the records, the premise of this approach is that records with poor quality will be compressed into longer representations than those with good quality. Experiments were conducted with some compression algorithms and, for the sample files, those that were manually labeled as poor quality seem to be statistically associated with higher values in compression rate. Further experiments with more data samples are being conducted.

The cluster entropy described in the Activity 1 or Objective 2.1.a. was compared to cluster of precision, recall, and F-measure for the fully annotated datasets of synthetic name and address references S1 to S18 used to test the DWM POC. For these datasets, the entropy metric tracks closely with cluster F-measure for those clusters describing the same person provided there were no spouse references in the same source. Whereas the F-measure of the annotated references is based entirely on the ground truth regardless of reference similarity, the entropy tended to allow (miss) false positive links where a couple with the same last name and same address were very similar and incorrectly linked. Similarly, the entropy measure would often allow (miss) false negative errors where the references had different addresses for the person. These errors did not happen in cases, especially when there where other identity attribute values to correctly discriminate such telephone number or data of birth. More research is underway to refine the entropy measure to overcome these defects.

The results of unsupervised quality metrics will be compared with the result from the supervised results using measures such as K-L divergence, Rand index and other statistical metrics that measures the difference between statistical distributions.

Establish a repository for the reference datasets and make available to other researchers. A GitHub repository for the DWM has been created and is being maintained that includes proof-of-concept code. It has a "master" branch with the original code and a "beta" branch for additional tests like the Bayesian Optimization. Additionally, a "development" branch can be added for the full Python version. This is useful as GitHub is widely used for sharing code and databases. Two datasets were created based on real world vendor data published by the Federal government.

All contracts and awards made by Federal agencies are required to be made public. This repository has millions of transactions, and we focused on vendor information because it provides an easy-to-understand problem for entity resolution, namely who is providing the services or products to the Federal government. Additionally, this information is likely entered by a multitude of people over a period, and so it is likely to be subject to real world data quality issues.

The sizes of the two datasets were set to 1,000 and 10,000, and they contain 29 columns for each vendor (listed below). A separate ground-truth file was also created to allow researchers to easily compare the results of any algorithm applied to this data to the actual entity represented in the main file. The results and data sets of developed machine learning models will make available to other researchers using GitHub.

Formulate a hierarchical and as-needed data collection and cleansing strategy. In order to allow the extraction of additional data from the publicly available repository of all contracts and awards from the Federal government, the following framework was setup and tested. Although this framework can produce large data sets, we only used it for small scale production. Two separate files can be produced, (a) the actual data regarding vendors for the US departments and agencies, and (b) a ground truth file to precisely identify each vendor. The steps of this framework are:

- Download the desired years and US Departments from the online archives provided Federal Procurement Data System - Next Generation (fpds.gov)
- Populate a SQL database with the data. We were able to convert the XML format from fpds.gov into .csv, and then use optimized queries to populate the schema of the database.
- Formulate queries to determine appropriate vendors. Not all vendors can be uniquely identified because, for example some are not based in the USA.
- Compute the ground-truth files.

Formulate a framework for sequential data collection on an as-needed basis. The Min-Max ratio test for unlabeled paired samples described in Objective 2.1.c: Automate Data Integration, Activity 1 is one of the preliminary studies for this activity.

Refine the formulation by including various practical constraints and test on small-scale problems. Nothing to report. This part of the activity is awaiting further work on the Min-Max study.

Document and train team on data cleansing methods developed in prior research. The first attempt at unsupervised data cleaning for this project was used in the DWM POC. The DWM POC process including the global (file-level) data cleaning was described in a working paper which was later published as "An Iterative, Self-Assessing Entity Resolution System: First Steps toward a Data Washing Machine" (International Journal of Advanced Computer and Applications, Vol 11, No 12, 2020) The working paper and later publication were made available to the team documentation for the process as well as making the Python code available on BitBucket.org.

The previous work of the team includes unsupervised clustering algorithms DBSCAN and SCAN, both have been successfully applied to entity resolution and many other applications. The current data are semi-structured or free text. The team is developing novel approaches based on machine learning and AI to address the challenges of automating data cleaning.

Design and implement in Python or Java improvements to the prior frequency-based approach.

- **Sentence-BERT Zero Shot Learning for Entity Resolution.** Using an artificial intelligence model trained for Natural Language Processing (NLP) tasks, we used an already trained model to find multidimensional vector embeddings for each text record in each data sample. As the model was trained on a different task, this is known as Zero Shot Learning. After projecting the records into the vector space, a distance matrix is computed using either cosine similarity or L2 distance. With the distance matrices as input, the performance of unsupervised clustering algorithms is explored including Agglomerative Hierarchical Clustering, Hierarchical DBSCAN and Affinity Propagation. In some samples, the performance is like the DWM and in others it is significantly inferior. The time for a complete run seems to grow with the square of the number of observations. Further developments will be focused in gaining scalability and improving recall.
- **Using RoBERTa for similarity evaluation.** Using an approach similar to the previous point, a Natural Language Inference tool is used to evaluate the similarity of two text records. This could replace the embedding and distance matrix portions of the previous procedure. As Dr. Talburt has pointed out, it can also be integrated in the current DWM.
- **Transitive closure port to Python.** Created an implementation of the transitive closure Java routines in Python. This might allow for further ports of Java code into Python.
- **Cluster-Level Data Cleaning.** The initial POC implemented a global (file-level) data cleaning routine. It successively compared the similarity between high-frequency tokens and low-frequency tokens. When the similarity was within one Levenshtein edit distance (one character difference), the high-frequency token was a candidate to replace the similar low-frequency token. However, to avoid introducing errors, certain constraints were put into place to restrict the replacement operation. These constraints include a lower limit on the high-frequency search, an upper-limit of the low-frequency search, and a minimum length of a replaced token (usually 3 characters) in order prevent situations like "A" frequency 100 from replacing "Z" frequency 1. While helpful, these constraints still created incorrect replacements such as "RONALD" with frequency 50 replacing "DONALD" with frequency 3. The problem is that correct low-frequency tokens can legitimately occur and be very similar to correct high-frequency tokens. For this reason, an additional constraint was added to prevent a low-frequency token from being replaced if it appeared in a dictionary of familiar words and names. The dictionary built by extracting single-token phases from an open-source Python English dictionary, then

supplementing these with name lists taken from U.S. Census data. The use of the dictionary also excluded the possibility of numeric token replacements. In a large file, there will often be a wide range of very similar numeric tokens, and frequency alone is not a sufficient constraint. Just because “123” has a high frequency, it doesn’t mean it should replace “1234” occurring with a low frequency. In case of numeric tokens, the dictionary does not provide any additional guidance. However, the reasonably accurate clustering results of the DWM POC provide a new opportunity for data cleaning at the cluster level. As described in Activity 1 of Objective 2.1.a, the algorithm for assessing completeness can be extended to data cleaning. Again, borrowing from the genomics, the reference tokens sequences can be used to find gaps that potentially represent incompleteness, misspelling, or data corruption. The advantage to cleaning at the cluster level is knowing the references in the same cluster are, or are very likely, to be for the same entity, and are already judged to be similar. There are two scenarios:

- The first scenario as discussed in Activity 1 of Objective 2.1.a, is for incompleteness. As in that example, suppose one reference has the sequence of tokens ABCD, and another the sequence ABD. Given the tokens A, B, and D, match in order between the two references, C can be understood as a missing value in the second references. In addition to using this to measure incompleteness, it could also be used to correct incompleteness by inserting the token C into the corresponding position of the second reference so that both references share the same sequence of tokens ABCD.
- The second scenario is very similar, but in this case the sequence ABCD in the first reference aligns with the sequence ABED in the second reference. Here the issue not as clear as the first scenario. The possibility is that either C should replace E, or visa versa, E should replace C. When the clusters are small, making the judgement by token frequency within the cluster may not provide sufficient justification for either. Frequency within the cluster may need to be supplemented with the global frequency of both tokens. One advantage of this approach is that it potentially allows for the correction of numeric tokens. Whereas, replacing “123” with frequency 3 with “1234” with frequency 100 at the file level is risky, replacing “123” with frequency of 1 with “1234” with frequency of 3 within a cluster 6 references would be a more confident decision.

Additional research is currently underway to evaluate various rules for cluster-level cleaning. One factor under consideration is applying this same type of sequence logic at the block level in place of, or in addition to, the cluster-level cleaning. Cleaning at the block level prior to clustering may provide a way to significantly increase the accuracy of the linking by make key token replacements that increase the similarity between references to the same object. If successful, it will provide a powerful lever for improving not only data quality, but the data integration. This could lead to several new washing cycles for the DWM starting with:

Global-Level Cleaning ○ Block-Level Cleaning ○ Clustering ○ Cluster-Level Cleaning

- Repeat Steps 2, 3, and 4 until no changes are made

The team is developing Python and Java improvements based on Machine Learning and AI. More specifically, the deep learning autoencoder and decoder model based on pretrained language model will be implemented.

Document and train team on reference clustering method developed in prior research. The DWM POC process including the unsupervised clustering algorithm was described in a working paper which was later published as "An Iterative, Self-Assessing Entity Resolution System: First Steps toward a Data Washing Machine" (International Journal of Advanced Computer and Applications, Vol 11, No 12, 2020) The working paper and later publication were made available to the team documentation for the process as well as making the Python code available on BitBucket.org.

Network clustering (aka community structure detection or graph partitioning) is an important task for the discovery of underlying structures in networks. Many algorithms find clusters by maximizing the number of intra-cluster edges. While such algorithms find useful and interesting structures, they tend to fail to identify and isolate two kinds of vertices that play special roles – hubs that bridge clusters and outliers that are marginally connected to clusters. Identifying hubs is useful for applications such as viral marketing and epidemiology since hubs are responsible for spreading ideas or disease. In contrast, outliers have little or no influence, and may be isolated as noise in the data. We developed a novel algorithm called SCAN (Structural Clustering Algorithm for Networks), which detects clusters, hubs, and outliers in networks. SCAN clusters vertices based on a structural similarity measure. The algorithm is fast and efficient, visiting each vertex only once. An empirical evaluation of the method using both synthetic and real datasets demonstrate superior performance over other methods such as the modularity-based algorithms. SCAN has received over 821 citations since published in ACM SIGKDD'07 according to Google Scholar. It is adapted in some popular data mining textbooks.

Design and implement in Python or Java improvements to the prior frequency-based approach. We derived the distribution of the ratio of the minimum to the maximum of two paired normal random variables with non-zero means. This led to defining a Likelihood Ratio Test (LRT) procedure that enables identifying differences in parameters of the two distributions. Due to the min-max ratio being invariant to the order of the variables, it allows for the estimation of the test even when the labels of the paired observations are missing or unknown. For the context of this work, "label" refers to knowing if the observation comes from the first or the second population. This might be useful in contexts where the labeling of each pair is unreliable, or it is desirable to not share the labels. Analogous to privacy protection machine learning methods, the developed LRT might be useful to test for differences in parameters while hiding sensitive information from the analyst. This work has been submitted to Statistics and Probability Letters for potential publication.

The team is currently working on novel machine learning and AI models and algorithms to automate data cleaning. The models and algorithms are implemented using Python. The preliminary experiment shows a superior performance.

The original DWM POC written as a combination of Python and Java code is now being refactored as an entirely Python program. The DWM Refactor codes is available on BitBucket.org and reported in ERCore.

Build genomics database, including quality scores, and gene/ protein annotation. We have decided on using the two largest genome databases, which contain most (nearly all) of the publicly available prokaryotic genomes. We have downloaded a set of more than 300,000 bacterial and archael genomes from the NCBI (The National Center for Biotechnology Information, which is part of The U.S. Library of Medicine). We have also downloaded the complete set of genomes from the Integrated Microbial Genomes and Microbiomes project, which is part of the U.S. Department of Energy's Joint Genome Institute (JGI) at https://img.jgi.doe.gov.

Use Elastic Cloud Storage for fast retrieval. We have built a **structured, organized genome database** that is stored on the high-performance computer system at UAMS. We have performed quality score analysis for more than 300,000 bacterial genomes in GenBank.

For each genome, we have used a standardized pipeline for finding genes. We run the Prodigal gene finder, with the same settings, for protein identification; we have used RNAmmer for finding ribosomal RNA genes, and tRNAscan-SE for identifying tRNA genes. For each genome, we have stored all Pfam domains and architectures from the predicted proteins.

Compress data and remove duplicate entries. We have more than one billion proteins (!) from our collection of 300,000 bacterial genomes; out of these, we find roughly 170 million proteins that are present in identical copies (that is 100% the same amino acid sequence).

Use Pfam domains and compression of sparse matrices for optimal retrieval in proteome comparison studies. Visanu Wanchai, a Ph.D. student in our group, has developed (and published) a program that does this (ProdMX), which can speed up genome comparison more than a million-fold, compared to the traditional all against all alignment methods.

We have used heat maps to visualize the comparisons of more than a hundred-thousand E. coli genomes. We utilized Mash, a program that approximates similarity between two genomes in nucleotide content, and an in-house Python script to create a matrix of distances, which can be displayed as a heat map. This work was done primarily by Kaleb Abram (a PhD. Student) and Zulema (a post-doctoral fellow), and was published in January 2021. We have decided on using the two largest genome databases, which contain most (nearly all) of the publicly available prokaryotic genomes.

Prototype of R-BioTools for visualizing genomes. We are developing methods and tools for visualization of pan- and core-genomes. This is part of the R-BioTools package, which is an R-Studio version of our previously published CMG-BioTools. We have recently released a python pipeline for pan-genome-based functional profiles for metagenomics samples from microbial communities.

Develop architecture / structure for rapid storage/retrieval of taxa-specific pan- and core-genomes. We are in the process of building a carefully curated set of type strain genomes to be used as an anchor or reference for mapping taxonomy of bacterial species, in a consistent and reproducible manner, using DOIs to point to the current taxonomy, as well as synonyms and previous names. From this we will add all current genomes, and a distance map to the nearest type-strain. This will be stored in a graph database, for rapid access of all the genomes of a given taxonomic group. Further, their proteomes will be approximated by using Pfam architectures, which can allow fast functional identification across all genomes (that is, less than one second to search for a given function across several hundred thousand genomes). As part of this database, we will be able to quickly pull up taxa-specific architectures.

Compare duplicate, known type strain genomes using ANI, Mash, 16S rRNA. In principle, duplicate genomes from the same strain should have identical, or nearly identical values in terms of genome-derived properties that are often used for taxonomy. We have tested three commonly used sequenced-based methods for predicting an organism's taxonomy: Average Nucleotide Identity (ANI), Mash, and 16S rRNA, on a set of about 3700 genomes that have two or more sequences deposited to GenBank (this represents 1610 unique type-strain species). Kaleb Abram (Ph.D. student) has a near-finished manuscript describing the results; we hope to submit the paper in April or May.

Build a novel proteo-genomics database for type strains. We will incorporate species-based pan- and core-genomes to build a novel protein database to identify and quantify novel protein isoforms from bacterial type strains. This work is being done by a group of Ph.D. students, and will provide information at the protein functional level.

Identify key datasets and problems for ML. We are utilizing the bacterial genomes described in activity 2 to build a machine learning approach for meta-proteomics analysis in order to understand the host/bacteria response within a biological system. A challenge for metaproteomics is the conserved peptide sequences for taxa identification. The novel curated genomes and the Pfam domain information will be included to clearly identify and quantify protein changes among bacterial species under various biological conditions.

Integrate genomic / microbiome / taxonomy datasets (petabytes). We are starting a set of experiments where we grow replicate bacterial samples of the Escherichia coli type strain [DSM 30083], and perform genomic, transcriptomic, and proteomic analysis. The curated type strain databases will be utilized to test and correct where necessary the taxa annotations. We will use lessons learned from this to assist in building novel protein databases by incorporating the gene level information. We have recently published a paper in the journal "Molecular Omics", which gives an overview of multi-omics approaches.

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

7.3. Social Awareness

How does Social Awareness impact data analytics? Social awareness is pivotal for those who work with data analytics and is a key factor that affects the uses, benefits, and risks of big data. It is a common practice for both government agencies and private entities to collect and integrate volumes disparate data, process it in real time, and deliver the product or service to consumers. There are increasing worries that both the acquisition and subsequent application of big data analytics could cause various privacy breaches, render security concerns, enable discrimination, and negatively affect diversity in our society. All these concerns affect public trust regarding big data analytics and the ability of institutions to safeguard against such negative social outcomes.

Document literature research of attack models and mechanisms behind attacks. We have conducted a survey of existing attacks on deep learning models which can be broadly categorized into evasion, poisoning, model stealing. We particularly focused on adversarial examples, model inversion, and membership inference attacks. For each type of attacks, we examined threat models from four aspects, adversarial falsification, adversary’s knowledge, adversarial specificity, and attack frequency, and then researched representative algorithms. For example, for the attack type of adversarial examples, we researched representative approaches (such as fast gradient sign method, projected gradient descent adversarial patch attack, and so on) and examined their mechanisms on how to generate adversarial examples. We found most approaches are based on constraint perturbation and involve stochastic gradient descent (SGD).

Initiate theoretical investigation on the risks of deep learning algorithms (Complete). We have conducted theoretical studies of the potential risks of deep learning models. We examined the correlations among input data features, parameters, latent feature space, and output of deep learning models and studied their sensitivities under adversarial attacks. We found the stochastic gradient descent algorithm widely used for training deep learning models is generally sensitive to input data perturbations, which incurs potential risks of trained deep learning models, e.g., changing the prediction output under adversarial attacks.

Initiate theoretical investigation of privacy preserving mechanisms (Complete). We have conducted theoretical investigation of privacy preserving mechanisms for deep learning. To achieve differential privacy in deep learning models, we can adopt different mechanisms, such as perturbing input data, using differential privacy preserving SGD via gradients clipping and perturbation, or adding noise in the objective functions used for training deep learning models. We studied the utility-privacy tradeoff and applicability of each mechanism on different types of deep learning models. The research findings from Activity 1-3 lays out a solid foundation for developing the universal threat- and privacy-aware deep learning framework in Objective 3.1.b.

Selected the approaches through literature review. Crowdsourcing has been an emerging machine learning paradigm. It collects labels from human crowds, typically through internet, as inputs for further learning. Due to its open nature, there are various uncertainties related to human factors such as participants' knowledge level, intention, social-economic status, etc. Commonly used binary-valued labeling scheme forces a worker to either accept or reject an instance completely even with ambiguity. In this work, we investigate interval-valued labels to enable a worker specifying both type-1 and type-2 uncertainties in his/her label without information loss.

Computational schemes are identified. Collected labels on a given instance reflect opinions of multiple crowd workers. These raw labels should be aggregated for a reasonable inference. The commonly available aggregation strategies, including majority voting and others, assume binary-valued labels mostly. Studying statistic and probabilistic properties of interval-valued labels, we have developed algorithms that is able to aggregate interval-valued labels as an inference with a preferred probability of matching above 50% computationally.

Identified possible sources of noises. To further improve data quality, we have established strategies to pre-process collected interval-valued labels. These strategies can be divided into two categories. One is data cleaning; and the other is normalization. In data cleaning, we do the followings:

1. eliminating neutral and/or near neutral interval-valued labels;
2. fixing out-of-range labels if possible or otherwise abolishing them; and
3. identifying likely unreliable workers either without sufficient knowledge or possibly with adversarial intention.

In data normalization, we make all interval-valued labels with a unified type-2 uncertainty through a delta-normalization. For these not near neutral interval-valued labels but containing 0.5, we perform an optional symmetric cancellation to adjust their type-1 uncertainty.

Specified mathematical requirements. Interval-valued labels are interval-valued datasets. To manage uncertainty with interval-valued labels in crowdsourcing, we must clearly define statistic and probabilistic properties specifically for interval-valued datasets. We have extended traditional statistic and probabilistic concepts for point-valued datasets to interval-valued ones. These concepts include mean, variance, standard deviation, and probability density function for interval-valued labels.

We have investigated two learning algorithms on deriving inferences from interval-valued labels. One is majority voting; and the other is with preferred matching probability. The former is a simple extension from commonly used binary-valued labels to interval-valued ones. The latter is based on the theory of probability distribution of interval-valued datasets. That leads to a preferred above 50% probability matching ground truth. Furthermore, we established an uncertainty index that quantitatively measures the level of overall uncertainty of collected labels from crowd workers.

We have carried out computational experiments to test the effectiveness of applying interval valued labels in managing uncertainty in crowdsourcing. Our experiments have successfully verified our theoretical and algorithmic results. The testing datasets are synthetic. Our computer implementation is in a current version of Python.

We have documented our results and submitted to the 2021 Annual Conference of the North American Fuzzy Information Processing Society NAFIPS 2021, which is a top conference in the field according to Guide2Research. We received the acceptance notification after anonymous peer review on March 21, 2021. The manuscript is expected to be published by the Springer in this summer. We are slight ahead of our planned year-1 objective on this goal.

Document and disseminate the findings on personal identifying information and their privacy issues (Complete). The team has done an extensive literature review on the definition of personal identification information (PII) from different perspectives and their privacy issues. When considering PII attributes only by their semantics, PII attributes can be categorized as either public PII or protected PII. Public PII is available in public sources such as telephone books, public websites, business cards, etc. Public PII usually does not require redaction prior to document submission. Protected PII is defined as individual’s name in combination with any one or more types of information, including SSN, password numbers, credit card numbers, biometrics, medical records, and so on. So protected PII attributes should be subject to more stringent control compared to public PII. We discussed some of our findings in one project-wide zoom meeting, and will document our findings in a report. When PII attributes are embedded in unstructured documents, then they need to be extracted first and then their semantics need to be identified before any downstream analysis. This process is actually a typical named entity resolution process. Researching the state-of-art techniques in NER is one of the objectives in Activity 3.

Identify and disseminate the findings on the sensitivity of information in different context. A key challenge to identifying PII attributes and preventing the privacy leakage is to identify the sensitive information embedded in unstructured documents. The sensitivity of a PII attribute might be varying in different context. A PII attribute can be private by itself or by combining with other information. For example, a nine-digit number might be privacy sensitive if it is a valid SSN; however, it will be very sensitive if it is combined with its owner's name. Several frameworks have been proposed to assess the sensitivity of information in different context. One approach is based on the linguistic constructs of sentences to capture different types of PII sensitivities. By viewing linguistic constructs as three-part structure, namely, the subject, the predicate, and the extension, the sensitivity measure of a construct is defined as a weighted sum of sensitive measures of three parts. As this research focuses on assessing the cumulative sensitivity measure of the leaked PII attributes of an individual, we are currently working to develop a metric that considers both the sensitivity level of each PII attribute and the combined sensitivity of a given set of leaked PII attributes.

Research appropriate text analysis techniques to identify sensitive information from unstructured data. Identifying sensitive information from unstructured data covers a broad area of research including named entity resolution and identification, natural language processing, privacy ontology, etc. Privacy ontology is primarily used in designing and managing privacy policies, so it is out of the scope of this study.

Named entity resolution (NER) systems have been studied and developed widely for decades, but accurate systems using deep neural networks (NN) have been introduced in recent years and shown promising performances compared to the traditional methods. One system with great potentials is a bidirectional LSTM-NN architecture that is able to automatically detect word- and character-level features using a hybrid bidirectional LSTM and CN architecture. Standford CoreNLP is one of the widely used tools for core natural language analysis by both research community and commercial products. One of the utilities is NER that recognize both named and numeric entities such as person, location, organization, number, date, time, etc. Most current NER systems are supervised-based that requires the training data to tune the system. We will investigate appropriate techniques for unsupervised NER to remove its dependence on training data.

One special challenge in this research is, when a document contains PII attributes of multiple entities, how to associate PII attributes with their corresponding entity. Researching appropriate techniques to match PII to its entity will be one of our future research tasks.

Initiate theoretical investigation on using CNN to recognize discriminatory objects (Complete). We have conducted theoretical investigation on using deep learning models such as CNN to identify discriminatory objects from social images. The features in the last layer of CNN models can be used as the semantic representations of discriminatory objects.

Initiate theoretical investigation on using LSTM to model discriminatory text (Complete). We have conducted research on adopting long short-term memory network to model the text such as captions, tags, and discussions of social images. We use word embeddings to represent each word in the text and capture its hidden semantic and grammatical meanings. The LSTM captures the whole sequence information via its maintained hidden state vector. We have also studied attention mechanisms to derive the correlation weight between each word and the text label.

In addition to the above theoretical studies, we have conducted empirical studies on multimodal hate speech detection. In particular, we compared unimodal algorithms (e.g., CNN and image-grid on images, LSTM and BERT on text) and multimodal algorithms (e.g., Late Fusion, Concat BERT, and MMBT-grid) on two multi-modal hate speech detection datasets, MMHS150K and Facebook Meme Challenge datasets. The preliminary results showed the effectiveness of using deep learning based techniques to detect cross-media discrimination. We also conducted research of detecting coded words in hate speech detection. For example, on Twitter, “Google” is used to indicate African American, and ``Skittles'' is used to indicate Muslim. As a result, it would be difficult to determine whether a hateful text including “Google” targets African-American or the search engine. We developed a coded hate speech detection framework, called CODE, to detect hate speech by judging whether coded words like Google or Skittles are used in the coded meaning or not.

Document and disseminate the findings of literature research and evaluation of the target products for case study. The team selected the gaming industry as the primary area for the analysis. The originally proposed vacuum cleaner industry may be used in the future for validation purposes. The decision is made after discussing with our external collaborator, Sam M. Walton College of Business faculty, Dr. Dinesh Gauri, at the University of Arkansas. We found that the gaming industry is more appropriate than the vacuum cleaner industry for our proposed research agenda in terms of the availability of datasets and the rich marketing interests in this field. The chosen products are video games, e.g., Call of Duty, God of War, Grand Theft Auto, etc., and game consoles, e.g., PlayStation, Xbox, and Nintendo series, released in the past 15 years (2005 - 2020). We have started analyzing social media data from Twitter, Reddit, Tumblr, and Gamespot, using Brandwatch, and currently reaching out to various external collaborators to collect consumer panel data. Brandwatch statistics provide data visualizations, metrics, and other data syntheses. We collected sentiment over time, demographics, content sources overtime and breakdown, online behavior, geography, and product mention overtime. Such information was very useful to learn about the product diffusion and adoption for the respective products and the sentiment based on the marketing campaign for the product. From various game console and video games market analyses performed on December 2020, the following results were concluded. On average, 81% of male users review the gaming products than the female population, which only makes up 19% of the social media for the gaming industry. Based on the Entertainment Software Association, women make up to 48% of the gamers, which is only 4% less than the male gamers. Yet, based on the market research, most game publishers appear to believe that making cognizant and inclusive design choices in their games will result in lower sales and revenue (as a result of alienating their male player-base), but in reality, the opposite is true. This is because women make up a larger portion of the video-game players, and hence, it is as important that female game players' preferences should also be considered when designing gaming products. Understanding customer choice behaviors for both males and females in the gaming market, their likes and dislikes also inform optimal design decisions in product development. The will in turn, help to develop marketing strategies that are more personalized campaigns with the intent to avoid any discrimination rooted in the tactics and consideration of customer preferences.

Previous literature has answered the question regarding the software-hardware relationship of video games and game consoles, the gaming product lifecycle, and the effect of each stage on market demand. The first stream of literature focused on the amount of available gaming software and showed that a greater amount of available software increases hardware demand. The second stream of research examines the software-hardware demand link separately for different types of software. In particular, this literature categorizes software into superstars (i.e., software of exceptional quality) and non-superstars (i.e., the balance of software that may even include good-to-average quality products). In our second study, the market was examined for gaming software and hardware in US households. Results show that both quality and social network effects are significant factors in determining market share in high-tech markets. In our study, we want to dive into the market based on sensitive attributes and find the link as well as the impacts of potential biased advertising on demand for the product and its effect on the brand name.

We are now working on collecting product information and reviews from dynamic websites, e.g., Amazon, GameStop, and Bestbuy, to evaluate and make comparisons of the product features and conduct a study of their current marketing strategies for the target products. A much thorough text mining and sentiment analysis will be performed using Python 3 over the Brandwatch data to capture more detailed reviews of the top-ranked product features, remove noise and neutral comments and ensnare any forms of bias embedded in the advertising and marketing campaigns.

Document and disseminate the findings of processing the data from Nielsen datasets and extract the information needed (e.g., demographics, product market segment, etc.) for this project. To help progress in the research, the initial project plan was broken down into the following steps: i) Perform sentiment analysis on the social media reviews data and product information for select products; ii) analyze customer reviews data to identify pros/cons of product features or unfairness in social media advertising content; iii) study the effects of this unfairness on the market segments and the consumer's buying behavior; iv) identify unfairness by combining the analysis of the two datasets and create fairness measures, and v) assess the existing fairness measures introduced in the conference paper and develop a metric that best fits the design for the fair market system.

In this plan, to collect data for step (iii), the team reached out to external collaborator, Nielson Company, to collect data for the first industry selected: the automotive industry. Based on communication with our external collaborator, Sam M. Walton College of Business faculty, Dr. Dinesh Gauri, we arrived at a decision to exclude Nielson Company as a primary source for data collection as there was limited data availability for the automotive industry. The data in Nielsen is not appropriate, and they do not have individual customer-level information and only have the scan of shoppers' receipts. Dr. Gauri recommended the team explore and research various industries offered through the NPD analytics platform that provides access to a large repository of marketing data that can better help us answer the proposed research questions. NPD data description seemed more appropriate for the marketing analysis, and the team decided to request NPD Datasets for the gaming industry. NPD datasets will aid in (i) identifying customer demographics, (ii) analyzing purchasing behaviors by age, income, gender, (iii) monitoring sales by retailer, region, or territory, and (iv) velocity and distribution of the respective products. This information will provide information of the market segment such as market campaign target audience, exclusions of the marketing campaign, sales and demand of the product, sales and demand trends throughout the lifecycle of various game consoles, e.g., the product introduction, growth, maturity, and decline phases of a product generation. The team has created a comprehensive list of game consoles and video games released in the past 15 years for data collection and analysis. Currently, the team is collaborating with Dr. Dinesh Gauri and the NPD Group to gather real datasets for the products selected.

Document and disseminate the findings of researching the existing methods for fairness quantification and adverting parameterization (Complete). The team has made significant progress on fair machine learning literature review. A research paper (currently under review) performed an exploratory study on fairness-aware design decision-making. This paper explored existing statistical fairness metrics such as disparate impact, calibration fairness, group fairness, demographic parity, equalized odds, predictive rate parity, and fairness through unawareness to quantify potential unfairness in Adult Income data. The major highlight of this paper is the application of disparate impact and fairness testing to quantify unfairness in data and its effect on members of the unprivileged groups. From our initial analysis of the dataset, it was clear there was an unbalanced division of income prediction concerning two sensitive attributes: gender (male, female) and ethnicity (white, black, others). Based on such disparities, we divided each sensitive attribute into binary classes, privileged (white, male) and unprivileged groups (black, female). We trained Logistic regression and CatBoost classifiers on the pre-processed dataset to predict individuals' income in the test data using a 10-fold cross-validation approach.

Disparate Impact. In the Disparate Impact (*DI*) analysis, we observed that gender attribute had a severe disparate impact index value than ethnicity attribute using Equation (1). Equation (1) is based on the actual outcome, Y, and set as a reference of the disparate impact index value, $\mathrm{DI}_{\mathrm{ref}}$. We then calculate *DI* based on binary predictor C using Equation (2). Classifiers used to make predictions reinforced gender unfairness in the test data as well in comparison to the $\mathrm{DI}_{\mathrm{ref}}$.. In the second observation, we tried to remove the data's sensitive attributes and determine if the disparate impact index value could increase to conclude that we have an unbiased dataset. However, disparate impact index value aggravated even further, indicating that it is insufficient to remove discrimination by removing sensitive attributes from data-driven approaches.

Fairness Testing. A fairness test was conducted based on the calibration scores using predicted probability score ss to determine the Gender attribute discrimination using Equation (3). In this study, we established that females would be at a disadvantage and get a lower prediction even when the female has an actual outcome of $YY = 1$ (income greater than $50,000) and high predicted probability scores ($s \geq 0.5$). One of the many reasons for the discovered discrimination could be the lack of data for the unprivileged group. Few practical strategies suggested from previous ML literature can be applied to achieve fairness. They are (i) optimizing training data, (ii) lowering the unprivileged group's threshold ($P(S = s, A = f) \geq 0.3$), and (iv) receive an equal number of privileged and unprivileged groups' data.

$$PP(SS = ss, AA = mm) = PP(SS = ss, AA = ff) \qquad (3)$$

From the analysis, the following results are concluded. A severe disparate impact can impact a member of the unprivileged group and put them at a disadvantage during several decision-making processes. This could mean the exclusion of their perspectives in the design process (an exclusive design that caters only to members of privileged groups) or rejecting them as target audience since their predicted income is less than $50,000. In our case study, we observed that a young female woman (membership in two sensitive attributes) would be at a loss even when her actual income is greater than $50,000. This is because when using ML prediction models, this individual received a negative outcome even at a higher predicted probability due to the sensitive attributes. After determining the results from disparate impact and test-fairness, it is clear that we can quantify unfairness from data. However, direct application of the knowledge from computer science literature to design decision-making may not work due to the unique characteristics and challenges in the product design and development process, e.g., prototyping designs, evaluating design features, production, and planning, etc. We need to conduct further analysis to develop a metric that can fit the decision-based design framework from the knowledge gained in fair ML.

Following the literature review and the work conducted, a research framework mapped out in Figure 2 has identified gaps in the existing literature and the engineering design field to apply fairness application. The team is working on various design solutions to bridge the gaps between marketing strategies' design with fairness consideration in (i) product co-consideration (M3) and (ii) social networks (M1). One potential research direction is to document and disseminate the literature review in Decision-Based Design and provide a data-driven analytical approach to integrating consumer preferences into engineering design for fairness-aware decision making.

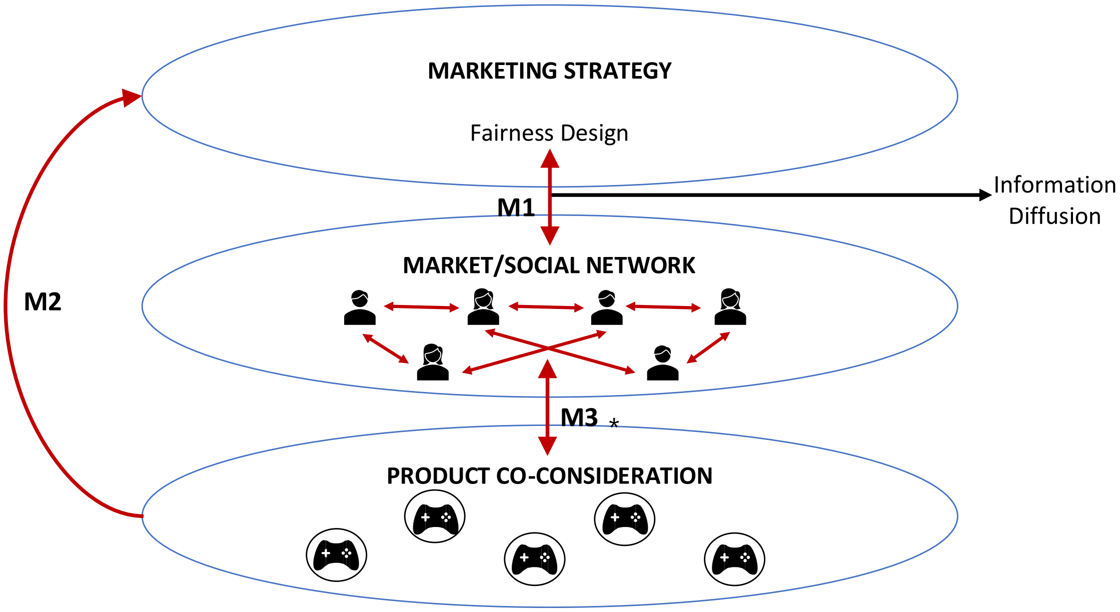


Figure 2: Research framework on Design for Market Systems

The team began developing code that will supports a data-preprocessing treatment on the Adult Income dataset to remove missing values, columns that don't affect the ML process, and change categorial features to binary and integer values. Then the data is split into 90% training and 10% test using a 10-fold validation approach. After splitting the data, supervised machine learning algorithms, Logistic regressions, and CatBoost classifiers were trained on the data to predict the outcome for the test data. The predicted probabilities and predicted outcome used for further analysis to detect bias and discrimination in the dataset. The observations are then used to analyze the effects of bias on members in unprivileged groups. This set of codes will be further developed and refined as the project unfolds in the following years. **Document and disseminate the findings of literature research of privacy-preserving data analytics algorithms and software (Complete).** For year one, our team members have conducted a survey of existing frontier work related to the privacy-preserving analysis in genomics and health. We recognized that this is broadly connected with the different aspects of the privacy-preserving analysis, including data formats and availability, infrastructure support, as well as algorithm design and development. We have shared and discussed the advantages and limitations of the current approaches, especially for privacy-preserving data analytics algorithms and software, directly related to this proposed activity.

Initiate investigation on mathematical optimization models (Complete). We conducted the initiate investigation of the mathematical models for privacy-preserving analysis in genomics and health. With the current infrastructure support and data availability, many algorithms are built based on models related to computational phenotyping, mathematical optimization and statistics models. This investigation helps lay a solid foundation for the further work on our proposed algorithms and technologies, which we anticipate work with a wide range of data types and high-dimensional heterogeneous data sources, such as text data, genomic data, imaging and video data, electronic health records, and medical imaging.

A survey of existing cryptography techniques and their applications in differentially private federated learning (Complete). We have conducted a survey of existing work that uses cryptography for privacy protection in federated learning. The survey covered around 30 papers published in the recent several years. We analyzed each work along multiple dimensions including the machine learning model/algorithm/method studied (e.g., neural networks, gradient descent, linear regression, deep learning, logistic regression), the type of dataset partition considered (e.g., horizontal partition and vertical partition), the cryptography method used (e.g., homomorphic encryption, secure multi-party computing), and also whether differential privacy is provided or not. Our main finding is that although there is much work that uses cryptography for privacy protection in federated learning, only three studies have combined cryptography with differential privacy to provide more advanced protection. One of the three studies relies on selected participants to add noise to achieve differential privacy and hence suffers from trust issues. In the second study, participants add excessive noise than needed for differential privacy, and thus the solution suffers from unnecessary loss in learning accuracy. The third study considers a special shuffling model and also induces high noise in the learning process. The survey calls for more attention to integrated designs leveraging both cryptography and differential privacy for privacy protection in federated learning.

Design of preliminary new cryptography techniques used for differentially private federated learning (Complete). We have designed a new cryptography-based scheme for differentially private federated learning. The scheme is designed with two goals in mind. The first goal is to reduce the communication cost in the training process, a known problem of federated learning especially for edge devices that have limited network resources. The second goal is to improve the learning accuracy while providing differential privacy. Considering the distributed stochastic gradient descent method, our scheme makes two contributions to meet the two goals. First, in each training iteration, whether a participant uploads its gradients to the aggregation server or not depends on a new factor not considered before, i.e., whether the direction of its gradients is well aligned with the collaborative convergence trend. That can speed up the convergence of the learning model and reduce the number of iterations needed, hence reducing the communication cost. Second, in each training session, an efficient homomorphic encryption scheme is used in together with a distributed noise generation method, such that each participant only adds a little amount of noise to its gradients but the total amount of noise accumulated in the aggregate gradients is enough for differential privacy. The amount of noise is less than existing solutions where each participant adds sufficient noise for differential privacy to its gradients and the total noise in the aggregate gradients is more than necessary. Experimental results show that our scheme performs better than existing work in convergence rate and also in the learning accuracy. This work will be submitted to a conference before the end of Year 1 for publication.

7.4. Social Media and Networks

What can we learn from Social Media and Digital Networks? Social media and networking platforms have billions of active users and leverage significant impacts on society. New types of social media and networking platforms or new features of existing platforms continue to be developed to meet users' demands. With an increasingly large amount of unstructured social data on these platforms, social media and networking analytics research has many scientific challenges including: detection of mis/disinformation, the ways in which it is disseminated, and the scope of impact; analytically assessment of the collective impact of social media and networking on societal polarization and other social phenomenon; and visualization of large social network data. Below is the description of progress to date on Year 1 milestones.

Key features for the platform are determined (Complete). We brainstormed various designs of cyber discourse social network platforms that integrate individual and social network measures. We also conducted literature reviews of various related platforms. We then determined the key features that we need for our platform. We plan to implement the platform in different stages. First, we will incorporate the most important features needed for our data collection. We will add additional features that are valuable for future data collection that will emulate features of widely used discourse apps.

Software design document is finalized. In order to implement the platform in a consistent way, we prepared a software design guideline document which will serve as a short-term and long-term guide for the platform development. We refer to a standard software design procedure with special focus on our needs for this project. The baseline user and discourse measures for the cyber-discourse platform functionality and user interface are determined. The enhanced training of natural language processor has been determined.

Develop questionnaire for collecting discourse data. We developed and designed a new social issue generator for social network data collection processing. Those generators reflect new up to date hot button social issues, such as COVID and face mask, COVID and collegial sports, racial inequality, and policing. We believe this new set of generators better capture current social discussion topics that draw individual students' social network. We plan to implement those generators in our Fall cohort 2021 social network data collections.

The question measures are determined, and IRB protocol is approved. The above-mentioned social issue generators received IRB approval for our Fall 2020 Data Collection, and will be ready for implementation in the Fall of 2021. The baseline individual user and Social Network measures for cyber-discourse platform have been determined.

The advanced natural language processing algorithms are developed. We spent significant amount of time collaborating on data production. Basically, thanks to a previous work, our team inherited a platform called Intelligent Cyber Argumentation
System (ICAS). Using ICAS, we are able to collect social network data from several cohorts of students taking a General Sociology class from 2018-2020. However, those data are in the form of tens of thousands of line of English chat threads. We worked together to identify key variables, and mine the

massive data to populate statistical compatible datasets (SPSS). Such an outcome is impossible without contributions from all participants including social scientists who identify key dimensions to be codified, computer engineers, and data scientists, who have the technology and skill sets to processing large quantity of data and produce customized datasets.

Social media platforms identified (Complete). We conducted literature survey, news article survey, and Internet research to identify various social media platforms used in different cyber influence campaigns. Several prominent platforms have been identified including mainstream social media (blogs, YouTube, Twitter, Facebook, WhatsApp, Telegram) and alternative social media platforms (Parler, BitChute, Rumble, Gab, MeWe, etc.). Contextual and geographic/regional differences were observed. Resulting publications have been uploaded on dartreporting.org website. Research was conducted in close collaboration with practitioners and policy makers.

Cyber campaigns identified (Complete). We conducted literature survey, news article survey, and Internet research to identify various cyber influence campaigns from different contexts (security, health, politics, foreign affairs/diplomacy) and regions (Canada, US, Europe, Turkey, China, Australia and the Indo-Pacific region). Some examples are COVID-19 misinformation cyber campaigns around the world and in Arkansas, Canadian Prime Ministerial Elections, cyber influence campaigns targeting NATO's military exercises, anti-US/anti-West campaigns in Indo-Pacific region. Other examples of cyber campaigns that we identified were specific to Australia-China foreign affairs and NATOTurkey-Russia diplomatic relations. These campaigns were identified along with our partners from Arkansas Office of the Attorney General, Canada PMO, Canadian Royal Forces, US Department of Defense, NATO, Australian Department of Defence Science and Technology Organisation (DSTO), University of Sydney, Turkey, among others. Resulting publications have been uploaded on dartreporting.org website. Research was conducted in close collaboration with practitioners and policy makers.

Characteristics and features identified (Complete). Social media and cyber campaign surveys helped in identifying characteristics and features of social media platforms used by various campaigns. Succinctly, characteristics include mainstream/alternative platforms, regional/national/global platforms, language (English, French, Spanish, Russian, Turkish, Chinese, Arabic, Tagalog, etc.), purpose (connecting, social signaling, social news, collaboration, health, gaming, entertainment, etc.), organic/inorganic (bot, social bot, botnet) behaviors. Features include content creation (text, video, image, audio), content enrichment (tagging, hashtagging, mention - @, etc.), content engagement (like, dislike, share, dig, bury, view, comment, up/down vote, etc.), content streaming, connecting (friend, follow, groups, lists, etc.). Resulting publications have been uploaded on dartreporting.org website. Research was conducted in close collaboration with practitioners and policy makers.

Taxonomy developed (Complete). Based on the characteristics and features identified, a multi-taxonomy characterization of social media data was conducted. This is a multi-year activity and will be revised as new characteristics and features are observed. Succinctly, dimensions of the taxonomy include user actions (blogging, vlogging, podcasts, networking, content engagement, content enrichment, etc.), behaviors (organic/inorganic, group formation, bridging/brokering, information solicitation, information diffusion), content-based characterization (using unsupervised clustering, topic modeling, color theory-based movie barcode approach), coordination-based characterization (single vs. multiplatform coordination, platform orchestration). Research was conducted in close collaboration with practitioners and policy makers. Resulting publications have been uploaded on dartreporting.org website. This includes a best paper award.

Data sources identified (Complete). Data sources have been identified. These include thousands of blogs, YouTube videos and channels, Twitter accounts, Parler, Rumble, BitChute, WhatsApp public groups, Telegram public groups, etc. Research was conducted in close collaboration with practitioners and policy makers.

Data acquisition procedures established (Complete). Data collection methodology has been developed and published. For each data source mentioned above, data acquisition methods have been identified that include API access and web scraping. For Twitter, an academic data collection license has been procured. This involved submitting a data access proposal and review. Proposal has been accepted. A multithreaded, parallelly distributed, fault-tolerant, scalable, resilient, accurate, data collection framework has been developed, tested, and deployed. The framework provides a live and real-time dashboard to monitor progress with alerting capability for excessive API usage, bottleneck in hardware infrastructure, exceptions, etc. Resulting publications have been uploaded on dartreporting.org website. Research was conducted in close collaboration with practitioners and policy makers.

Database setup (Complete). Database has been created after reviewing all the fields that can be captured from a variety of data sources. Database schema is extensible to accommodate new fields with changes in data sources or their characteristics. Recognizing the 4Vs (volume, velocity, variety, value) of the big 'social' data, database is designed to be efficient, scalable, redundant, and fault-tolerant. Research was conducted in close collaboration with practitioners and policy makers.

Key characteristics defined (Complete). While multimedia data on social platforms is rich and diverse, important features towards the broader goals of the program have been identified. These characteristics, such as classification, perceptive quality, and scalability play a major role and their identification and accuracy play a key role in the success of future program goals relevant to this objective. Overall, two graduate students and one recruit has been involved with the ongoing activities. One of the planned publications is currently being prepared for publication.

Identify and define three major learning objectives document. Building learning mechanisms based on heterogenous, multi-modal data sets is a particularly challenging problem. In this activity, we have identified one of the learning objectives that will later be tested and verified in subsequent years. We expect this activity to be completed in the second year with the rest of the objectives defined that will also clear the path towards the establishment of the learning mechanisms.

Three key applications defined. Although we have a pretty clear idea on the applications that will be the basis for the testing and verification of the activities in Objectives 4.3.a and 4.3.b, we have majority of this activity towards the end of the first and the beginning of the second year to better align the program objectives with each other. Nevertheless, we have identified one of the key applications as "damage assessment and verification based on image and video data" in disaster response situations. Other key applications will be determined based on the learning objectives defined in Objective 4.3.b and other program activities such as 4.4.a and 4.4.b. This particular goal (SM3) is so far on schedule. The potential threat that may hinder progress as planned may result from hardships in recruiting future graduate students under pandemic conditions.

Identify and define social platform content types of interest (e.g., image, video, text, etc.). The team has made healthy progress on identifying content types on social platforms that can describe transportation infrastructure status after disruptions due to a disaster. The team is using a framework for real-time humanitarian logistics data from the literature to characterize the content types according to attributes such as logistical content, timeliness and accuracy. Further, the team has documented the workflow surrounding how to manually transform data from individual content elements posted to social platforms into information regarding the transportation infrastructure status. The team is in discussions regarding how to move this manual workflow to automated processes, given team member capabilities.

Identify and define content types of interest (e.g., satellite imagery, traffic cameras) from sources other than social platforms. The team has made progress identifying other data sources at various levels (e.g., state, federal) that contain real-time humanitarian logistics data. For example, satellite imagery is available from the National Oceanic and Atmospheric Administration (NOAA) for all domestic regions. Traffic cameras are available for some states from state departments of transportation, like the iDrive program in Arkansas. The team is using the same framework for real-time humanitarian logistics data from the literature, mentioned under Objective 4.4.a, to characterize the content types according to logistical content, timeliness and generalizability. The team is in discussions regarding how access these data streams in formats that enable automated processing. These discussions will lead to a refinement of the list of content types covered explicitly in this project.

Obtain testing data from social platforms. This data has not yet been obtained. Some project team members have Twitter data from past hurricanes in their possession, however, these data are not current and were not obtained for the specific purpose of detecting road status information. The research team is currently exploring whether the Academic Research product track from Twitter, which provides free access to historical data, will meet project needs. Application for this product track will be pursued if deemed appropriate. Further, the team is investigating whether and how raw data from other social platforms (e.g., Facebook, YouTube) can be mined automatically (versus identified and extracted manually).

Select at least two disaster response routing problem variants using Milburn's existing qualitative interview data (Complete). A first routing problem with important disaster response application involves developing a least cost path for the transport of critical resources from an origin to a destination, and a second includes extending this problem for the case of multiple origins and/or destinations. These mimic the real-world practice of staging critical resources (e.g., truckloads of water, meals, etc.) upstream from the disaster theater and then moving them from the staging area to final destinations (e.g., Points of Distribution, food banks, mass care shelters, etc.). What these two routing problem variants have in common is the traversal of a road network that includes disruptions (e.g., flooded roads) that are only partially known. These applications are consistent with problems described in Milburn's qualitative data from interviews with humanitarian logisticians. A third routing application under consideration is the problem of selecting, for each vehicle (or team) in a fleet, a set of locations to visit, and planning the sequence of those visits, such that the demand served at visited locations is maximized, and operational constraints such as shift limits are not violated. This pertains to applications such as search and rescue. Similar to the two path-finding problems discussed above, this application requires the traversal of a road network with only partially known disruptions.

Conduct literature review for identified routing problem variants and publish journal article synthesizing review with qualitative data from 4.4.c.1. The problems described under Activity 1 can be modeled as Canadian Traveler Problem (CTP) and Orienteering Problem (OP) variants. The team has completed reviews of the CTP and OP academic literature. The "qualitative data from 4.4.c.1" listed in the above bullet point is a typo; the strategic plan should instead reference qualitative data from 4.4.d.1. As the qualitative analysis is still underway, journal article synthesizing the literature review with the qualitative data is incomplete. This effort will continue for the remainder of Year 1 and continue into Year 2.

Define GIS system requirements. Initial discussions of GIS system requirements have begun. The requirements are still evolving as content types and workflows from Objectives 4.4.a-4.4.c are defined. To speed progress in this area, a monthly meeting among all personnel contributing to SM4 has been established. In addition to the monthly meetings, PI Milburn will meet with PI Angel bi-weekly to translate project goals to specific GIS requirements

7.5. Learning and Prediction

How does Learning and Prediction impact data analytics? A major challenge in building secure and widely adopted deep learning systems is that they sometimes make wrong, unexplainable, and/or unpredictable misclassifications. In addition to confusing examples of very different classes, they are also vulnerable to adversarial examples. These systems are often trained as large feed-forward error-backpropagating black boxes and thus we have no way of interpreting the meanings of their features and understanding the causes of misclassifications, a situation that can be exploited by attackers. Research in this theme focuses on applying statistical learning techniques alongside more advanced deep learning techniques while investigating the challenges surrounding high-dimensional, dynamic, and unstructured data sets and exploring solutions in the domains of genomics, transaction scenarios in eCommerce, and supply chain logistics. Below is the description of progress to date on Year 1 milestones.

Complete the preliminary theoretical investigation on the proposed modeling approach. The team has made healthy progress on literature review and identifying the research directions. In particular, the team is able to replicate one of the existing methods in the literature for intelligent food-borne disease investigation (based on event data). The team has started working on the extension of the existing approaches for regularized large-scale problems. In addition, the team has reached out for external industry collaborators for the possibility of getting a real dataset. Team members have explored the RF-SRC method and the use of NHPP to model the recurrent event. Following the literature review, gaps in existing methods have been identified. The team has started investigating the potential of tree-based methods for capturing interactions among features for large problems.

Submit conference paper with initial model. The definition of the approach is on schedule. The approach utilizes a neural network to predict the parameters of a probability distribution that accurately models the dynamic probability of failure in an observed system. The paper is targeted for submission in April 2021 to the 2021 INFORMS conference Service Science to be held in August 2021.

Present conference paper with preliminary results of implementation. The proof-of-concept implementation is under development in Keras. An abstract is being submitted and the work will be presented at the 2021 INFORMS conference in Anaheim, CA. The results will focus on the comparison of the machine learning approach defined in Activity 1. The results will provide a comparison of fitting using a variational Gaussian versus a nonhomogenous Poisson process.

Publish curated data to GitHub. The teams has curated a dataset of sensor data, system attributes, and failure/repair data of 8232 oil and gas wells installed between 2007-2017. This dataset is being used in Year 1 work on the model framework and implementation. Candidate civil infrastructure datasets have been identified and are being evaluated for use in Objective 5.2b in Year 2. The team is also attending Data Curation thrust meetings to learn about datasets available in the healthcare industry from our collaborators at the University of Arkansas for Medical Sciences.

Development of the first unsupervised convolutional area and the first flow-based deep learning approach. Using local contextual guidance, an algorithm for extraction of broad-purpose maximally-transferable representations has been proposed. The proposed CG-CNN algorithm uses a single-layer CNN architecture; however, it can be applied to any type of data, besides imaging, that exhibit significant contextual regularities. Furthermore, rather than being trained on raw data, a CG-CNN can be trained on the outputs of another CG-CNN with already developed pluripotent features, thus using those features as building blocks for forming more descriptive higher-order features. Multi-layered CG-CNNs, comparable to current deep networks, can be built through such consecutive training of each layer.

Development of linear dimensionality reduction methods. In our experiments on natural images, we developed a library of classifiers from simple linear classifiers such as SVMs and linear autoencoders to complex deep learning approaches including AlexNet, ResNet, and GoogLeNet. The library also has some standard machine learning classifiers such as random forests, naïve Bayes, multi-layer perceptrons.

Application of the developed methods with applications on natural images and textures, and classification of a malware dataset. In our application to natural images, we find that our proposed contextually guided features (CG-CNN) show the same, if not higher, transfer utility and classification accuracy as comparable transferable features in the first CNN layer of the well-known deep networks AlexNet, ResNet, and GoogLeNet. We also explored malware classification by addressing high dimensionality issues. There are various features used for malware classification. Some features have good performance but require a lot of learning time due to their high dimensionality. The team proposes a low-dimension feature with entropy information. The feature shows good performance with less training time. The team will employ this feature for windows malware, android malware and IoT malware.

Another application the team studies is the fingerprinting of websites. This work explores the anonymous network vulnerability by analyzing network traffic data. There are thousands of features that can be extracted from network traffic data. The teams can extract about 100 features due to the feature important. The teams are studying how features affect learning models. One conference paper was accepted in 7th Annual Conf. on Computational Science & Computational Intelligence (CSCI'20), December 16-18, 2020. One poster paper was accepted in The Network and Distributed System Security Symposium (NDSS) 2021, Feb 21-25, 2021.

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). We investigated the efficacy of group structure on generalized neural network (GNN) architecture beginning with the smallest finite simple nonabelian group A5 (the alternating group of even permutations on five labels) by means of its isomorphic group of symmetric rotations of a regular icosahedron. Initial rounds of experimental results applying the A5 group action to random and clustered synthetic data of small size met with limited success. We then applied these techniques to the color channel data of three-dimensional fundamental topological structures (e.g., spheres, cubes, and tori). These applications yielded slightly more promising results with a linearization of the data being produced from several distinct images. The resulting data is currently being studied for any significance by means of exploratory data analysis and statistical inferential methods. This particular activity involves higher risk than our other activities in LP3. We, however, expect to close this activity by the end of June 2021.

An important aspect of the aforementioned activity included the training of six UGRA students who did not have any prior knowledge in the areas of machine learning, deep learning, or group theory. During the training, the students worked on prediction and regression problems and learned working with python and/or R programming languages.

Development of a generalized model of reward function in DRL addressing the issues with both sparse and dense feedback. In RL (DRL), an agent's navigation is guided by its acquired rewards. If the rewards, however, are too sparse, or too frequent (dense), it can hinder the progress and the learning process of the agent. Both of those scenarios can slow the progress and deviate the agent from the path towards the goal causing it to spend most of its time navigating the environment with little progress. We have analyzed the existing approaches to exploration-based methods for sparse reward in DRL and developed a generalized method to address this issue. A survey paper on exploration-based reward design, and another paper on a generalized (exploration based) reward model is in progress. Both are expected to be completed by summer 2021.

Develop Teacher - Student Distillation Deep Learning Algorithms. The team developed multiple low-cost deep learning methods, including Teacher-Student Distillation Deep Learning, Distilled ShuffleNet, Self-Knowledge Distillation Algorithms.

Develop analytic approaches to the proposed methods in Activities 1.1. The team already analyzed the current issues of distillation methods for various computer vision dataset, such as face recognition, action recognition, medical imaging, etc. and propose to develop improved frameworks.

Obtain and prepare cleaned data for research. The team has requested APCD data. In the same time, we have obtained another dataset (MIMIC-III) for our preliminary analyses. We anticipate to submit a manuscript using the MIMIC-III data by April 2021. We are also processing a private insurance claims dataset for predicting opioid overdose.

Acquire features that are highly representative. Using the MIMIC data, we have identified important features that are related to patient outcomes under ventilators. We have also created descriptive statistics to be included in the feature set. We have presented preliminary results at the 2020 INFORMS Annual Meeting.

Complete selection and testing of deep learning models. We are in the process of constructing deep learning models for prediction using the insurance claims data.

7.6. Education

What does data science education look like in Arkansas? Programs in data science are not currently offered at most Arkansas IHEs. On the other hand, courses and programs in computer science, information science, and statistics are widely available across a spectrum of institutions. Through this project we expect to ensure that all collaborating IHEs will gain a better understanding of the nature of data science and the appropriate resources that such programs require. Our vision is to create a model Data Science and Analytics program for colleges and universities in Arkansas to promote problem-based, and experiential-based pedagogy in critical thinking and analysis, technology familiarity, and foundation in math and statistics. This will form the basis of an educational ecosystem where learners receive a designed, consistent, sequenced, scaffolded, and modular education in data science with further educational and/or job opportunities available at appropriate points in their careers. Below is the description of progress to date on Year 1 milestones.

Middle School Coding Block- Initial Workshop completed, plan finalized and disseminated to stakeholders. Due to the prolonged pandemic, we are postponing the face-to-face workshop until 2022. Some initial meetings have been convened to being planning and contact is being established with master teachers and stakeholders around the state. The representatives from ASMSA that are collaborating on this project, Daniel Moix and Carl Frank, are working to develop a list of 30 K12 teachers that will be invited to the first face-to-face workshop. Additional virtual short meetings will take place this spring and summer to establish the timeline, and we plan to report additional progress in the Year 2 report.

Plan disseminated to stakeholders. The 5-year plan was outlined for stakeholders during the workshop that took place in November 2020 and will be further detailed in the April 2021 workshop. The November workshop hosted approximately 50 attendees from 40 campuses and organizations around the state. All information regarding the curriculum and plan are published for participants on a OneDrive site.

Cohorts identified; all collaborators assigned. An institutional needs assessment was distributed to all campuses in Arkansas to be completed by CS/DS faculty and IT support at each campus. The survey is helping the team establish the needs at each campus ranging from classroom technology and internet access to faculty experience and department sizes. Approximately half of campuses have responded, and we are working to collect the additional responses needed and fulfill this milestone.

The Education team has met periodically to review the UA and UCA programs and discuss progress as described. At the University of Arkansas, the Data Science program accepted its inaugural class which included 45 students. Of those students, approximately one-third are not calculus-ready, about a third are "standard 8-semester plan," and a third are transfer students from other majors or from other academic institutions. The latter group are a mix of calculus ready and not. As students began considering the program it became apparent that the issue of calculus-readiness would need to be addressed. Computer Science programs across the country have been developing alternative mathematical tracks the reduce calculus requirements for Computer Science majors. As we develop recruiting materials, we will need to highlight the need for appropriate mathematical preparation to enable students to complete the program in a timely manner. UA has also developed a suggested 6-semester plan for students who change their majors to Data Science and that can be adapted for the second two years of 2+2 programs. UA also has a representative on ABET's accreditation workgroup and is closely tracking requirements for readiness at the time of first graduates of the program, expected to be May 2023.

At the University of Central Arkansas (UCA) significant progress has taken place. Prior to the beginning of DART UCA had Data Science Tracks in its mathematics program and in its ABET-accredited Computer Science Program as well as significant coursework in business analytics within its College of Business. Working through the summer and the early fall a standalone BS in Data Science was developed. The new BS degree includes concentration in computer science, statistics, and business, and is constructed to allow the inclusion of additional concentrations. This will enable UCA to fully participate as a hub as its programs will not be based on completing a disciplinary track in mathematics or computer science.

Data Science for Arkansas Workshops. 3 Workshops completed. All three have invited all post-secondary academic institutions in the state and have been well-attended. In each, the Arkansas Division of Higher Education (ADHE), Arkansas Economic Development Commission (AEDC), and the Arkansas Center for Data Science (ACDS) have been actively involved, engaged, and participated. Additionally, the Office of the Governor has been supportive and informed.

Info disseminated to stakeholders. Overall Data Science Objectives and Outcomes have been presented and discussed at the previous workshops and this, along with ABET accreditation, was discussed at the workshop in April 2021. Our plan is to have a consistent and complementary set of objectives and outcomes across the state's implementation of data science programs. The UA Program has been broadly distributed. The existing tracks at UCA have been distributed at statewide meetings. The new UCA standalone program has been shared internally with cohort participants – it will be shared more broadly after program approval by the Arkansas Higher Education Coordinating Board.

Identify "Wave 1" of accreditation candidates. The UCA BS program has now received all academic approvals on campus and has been approved by the UCA Board and submitted for review and approval by the Arkansas Higher Education Coordinating Board. This degree program was developed around the available advice from ABET which focuses on computer aspects of data science programs. Pilot-year ABET accreditation for data science is scheduled for 2021-22. This process will be monitored to ensure that the program is adjusted for any changes that might occur as the result of the pilot and is serving as a model for other campuses nearing readiness for accreditation. Arkansas State University plans to join the first cohort of accreditation, and other campuses may be identified over the summer. Additional progress will be reported in Year 2.

Begin "Cohort 1" Proposal Preparation. The University of Central Arkansas has developed a program for a standalone degree based on its existing tracks. All on campus approval have been received. The program has been submitted to the Arkansas Higher Education Coordinating Board.

At the University of Arkansas of Pine Bluff (UAPB), discussions have begun as to how to begin a program in Data Science. Under the leadership of Aslam Chowdhury, UAPB will begin their efforts by adding a new concentration in their Computer Science degree. Dr. Addison is scheduled to visit Dr Chowdhury in April and will share his experience with using degree concentrations to develop enrollments so that a standalone degree can be developed at a later date after the concentration has matured. Significant progress has taken place at two other campuses that were not included in the original cohort. As a result of discussions that were initiated at workshops sponsored by ACDS prior to the grant being funded, Dr. Schubert assisted both Arkansas State University and North Arkansas College in the development of data science programs.

Arkansas State University has developed a BS in data science that was approved by the Arkansas Higher Education Coordinating Board in December and will begin accepting students in Fall 2021. These efforts were led by Jason Causey, Dr. Causey is the Associate Director of Arkansas State's Center for No-Boundary Thinking. North Arkansas College has developed an Associate's degree that is designed to be articulated with the four-year program at the University of Arkansas. The North Arkansas College efforts were led by Laura Berry. Dr. Berry is the Dean of Arts, Sciences, and Business and Information Technology. Representatives of both programs already attend the regular meetings of the DART Education Group as we continue our efforts to develop the state data science infrastructure. Arkansas State will be able to serve as a hub in northeast Arkansas and its participation will significantly strengthen our efforts to develop programs statewide. University of Central Arkansas serves as our hub in central Arkansas, and the University of Arkansas serves as our hub in northwest Arkansas.

Philander Smith College got approval for three courses- Intro to data science using python, Machine learning, and Ethics in data science. These three have been approved by the curriculum committee, and are now waiting on a board of trustees meeting for final approval to be added to course calendar. Intro to data science with python will be offered for the first time in fall 2021. PSC also has established a partnership with IBM and 3 faculty have completed IBM data science training.

Shorter College has finalized the curriculum for two courses and faculty are waiting to meet with the Dean and Associate Dean to discuss the approval and implementation plan. Shorter also partnered with IBM and 4 faculty completed the data science and artificial intelligence badge programs.

Create shared resources with UAF UCA existing materials and establish cataloging methodology. All course materials and curriculum are shared via a OneDrive to participants. As campuses adapt and implement the curriculum, they share those with the group as well, so the database grows continually with new variations on the base UA curriculum.

Identify "Opt-In" Research Theme Researchers & Collaboration Types & Timing. Opt-In partners among the research teams have been identified, and Schubert and Addison will work with them over the summer to identify integration opportunities.

8. Workforce Development and Broadening Participation

Why does Arkansas need a workforce skilled in Data Science? Data science is targeted in this NSF EPSCoR project because it is a strategically important technology for a significant and growing part of the State's economy. Arkansas companies, including Walmart, Tyson, J.B. Hunt Transport Service Inc., Stephens Inc., First Orion, and Acxiom, make decisions from data, employ large numbers of data scientists, and are recruiting a workforce with a higher-level of broad and integrated data science skills. The grand challenge of the workforce development and broadening participation initiatives is to create a larger, more diverse pipeline of people with rich educational experiences and skills in data science and computing graduating and entering the workforce in Arkansas.

EAST Initiative Annual Conference. Preparations with EAST Initiative began in fall 2020 to lay out the 4-year plan for the professional development workshops. The application for teachers was launched in March 2021 and also during late March, EOD Fowler will host an exhibit booth during the Virtual 2021 EAST Conference. Pictures and updates from the virtual event can be found on the @arepscor Facebook and Twitter pages. Engagement activity will be reported in Year 2.

10 seed grants awarded (4 awards complete by end of Year 1). When the strategic plan was prepared, the number of awards to issue each year was miscalculated. We budgeted $20,000 per year which results in four awards of $5,000 each, not 10. We decided to split the awards into two solicitation rounds during the year. The first round was offered in October 2020 and two awards of $5,000 were issued. One was awarded to the Arkansas Regional Innovation Hub for a virtual field trip project entitled "STEM Saturdays", and one was awarded to the Henderson State University STEM Center for a project entitled "ConneCTED: Developing Communities of Practice with an Emphasis on Computational Thinking and Engineering Design". The dates of performance for both awards take place in Spring 2021 and will be reported on in the Year 2 Annual report. The second round for Year 1 funding was announced in March 2021 and will be awarded later in the spring.

1 awardee presentation at All Hands meeting. One of the Year 1 awardees will be invited to present at the first DART conference in September 2021.

Cohort 1 established. The campuses that will submit faculty for the first cohort of training have been identified: Shorter College, Philander Smith College, University of Arkansas at Pine Bluff, North Arkansas College, and Arkansas Tech University. Each institution will choose two faculty to participate in Y1 training. The faculty will be identified in spring 2021 and updates will be provided in the Year 2 report.

Host annual workshops on a variety of grantsmanship and entrepreneurship topics. 3 Workshops completed. This activity will be completed by the end of the year. An email request was distributed to the DART students and faculty to solicit topics of interest for the Year 1 workshops. The first workshop will take place on April 7 and the topic will be "Communicating Science to Legislators" with Dr. Jory Weintraub of Duke University. The second workshop will take place on May 4 and with the topic "Individual Development Plans (IDPs)" with Dr. Barbara Bruno from University of Hawaii. The third workshop will focus on a few NSF programs of interest to our group including EPSCoR Track-4, CAREER, and RUI and will be scheduled for summer 2021.

15 UG supported (Complete). Please see participant funding table.

40 GA supported (Complete). Please see participant funding table. All 55 student research assistantship positions were filled during Year 1. DART student forums began in March 2020 and will continue to be offered every other month. The student poster competition will take place during the September conference and will be reported on in the Year 2 report.

5+ capstones identified. Capstone partners among the research teams have been identified, and Schubert and Addison will work with them over the summer to develop the first series of capstone projects. We plan to complete this by the end of Year 1, publish the capstones in time for the fall semester, and report progress in the Year 2 report.

Arkansas Summer Research Institute. This activity will be completed by the end of Year 1. The 2021 ASRI is scheduled to take place June 14-25 virtually. At the time of this report, the recruiting process is well underway with approximately 40 applicants so far. We plan to continue to recruit and are reaching out to 2-year schools and other campuses with low representation at previous ASRIs. We plan to accept 100 students and have developed a new list of interactive courses including introductory data science concepts, tools in the data science toolkit, programming in Python and R, genomics and bioinformatics, and career development topics such as resume and CV building, ethics, equity, and representation in research, and research literacy and presentation. The presenters will be largely from the DART faculty group and will include additional faculty and panelists from other institutions. We also have licensed a new platform called UpSquad which serves as an online community with teleconferencing and telework functionality that will provide a good basis for the longitudinal observations and provide better networking opportunities with the attendees and presenters. We look forward to reporting results in the Year 2 report.

10 UG supported. This activity will be completed by the end of Year 1. The SURE program was announced in March 2021 and awards will be issued in May. We will discuss the awards and results in the Year 2 report.

20+ scholarships provided. This activity will be completed by the end of Year 1. Scholarships will be provided as described for the 2021 ASRI which takes place June 14-25. We will discuss this in the Year 2 report.

1+ workshop completed. While regular meetings between Drs. Addison and Schubert and ACDS Director Bill Yoder have taken place, we have not yet established a regular meeting focused solely on opportunities for students in ACDS. Like DART, the operations of ACDS have been COVID-impacted - in the current year ACDS has focused much of its attention on developing its apprenticeship programs. The University of Central Arkansas hosts the Arkansas Coding Academy. Dr. Addison initiated meetings between Director Yoder and Coding Academy Director, Dr. Don Walker to facilitate the development of an introductory Data Science track to be offered through the Coding Academy for ACDS. Dr. Addison also worked with Dr. Walker on the content of the class which was offered in fall

2020. The course was successfully run, but Dr. Walker chose leave UCA after its completion. This has slowed progress of the initiative. Allison Wish, the newly hired Director of the Coding Academy has already been in contact with ACDS to collaborate on additional joint initiatives. Dr. Addison and Mrs. Wish have also met to formulate plans for future joint collaboration with ACDS and UCA's academic and outreach programs in the development of apprenticeship and internship opportunities. ACDS is still developing its programs. Most ACDS initiatives to date have been focused on retraining or providing opportunities for initial employment for those not intending to pursue higher education. As activities suitable for DART participants are developed, they will be disseminated to DART participants.

Develop apprentice and hosting company feedback and evaluation methodologies and instruments. Under COVID-impacted operations ACDS has focused on the development of apprenticeship opportunities through partners like the Arkansas Coding Academy and the Forge. We stand ready to aid them in the development of evaluation methodologies and instruments and anticipate that we will embark on these activities as we emerge from COVID imposed virtual operations. To date collaborations on the development of feedback and evaluation instruments have focused on the needs of the education partners in the DART program, in particular through the development of employer surveys focusing on their data science needs and institutional capability surveys. We anticipate that the direction of this collaboration will pivot toward the development of evaluative instruments in the second year of operation.

9. Communication and Dissemination

How do we grow DART and gain public interest? Great consideration is given to how project communications are executed to ensure that all DART participants are aware of their roles and responsibilities to the project, and to make the public aware of DART and its success.

Platform established and participants onboarded (Complete). A Slack group was established for DART and the faculty utilize Slack and email for daily communication. The DART GitLab was also established and as of March 2021 efforts were being made to ensure cross-campus participation. We are also exploring a license to UpSquad, a new telework and networking community that is being used for the 2021 ASRI. Virtual office hours have been held periodically over Fall 2020 but experienced very low participation, so this has been halted until need is reestablished.

DART Monthly Seminars. 11 webinars complete. The DART Monthly Webinar Series kicked off in October 2010. There have been 5 webinars hosted to date with 2 additional webinars planned by the end of Year 1. No webinars were hosted during the months of July – September 2020 while the strategic plan was being completed. No webinar was hosted in December due to numerous scheduling conflicts and we do not plan to host a webinar in May 2021 since the Annual All-Hands meeting will be held in May 2021.

Webinar Offerings:

- October 2020: Welcome to DART, presented by Dr. Jackson Cothren
- November 2020: Coordinated Cyberinfrastructure, presented by CI Co-Leads
- January 2021: First Steps toward a Data Washing Machine, presented by Data Curation and Life Cycle Co-Leads
- February 2021: Socially Aware Data Analytics, presented by Social Awareness Co-Leads
- March 2021: Social Media and Networks, presented by Social Media and Networks Co-Leads
- April 2021 (planned): Learning and Prediction, presented by Learning and Predication Co-Leads
- June 2021 (planned): Education and Outreach, presented by Education and Outreach Co-Leads

All webinars have been recorded and are available on the DART website. All DART faculty, staff, and students are invited to the webinar series. During the year 2 we plan to more widely advertise the webinar series to increase participation.

Monthly team meetings. 11 meetings per research team completed (6 teams).

1 All Hands & Poster Competition. Due to the prolonged pandemic, we were not able to hold a face-to-face meeting during Year 1. Many of the DART faculty and students received the COVID-19 vaccine in Spring 2021 and most campuses plan to transition back to in-person classes for the Fall 2021 semester. The first in-person meeting has been scheduled for September 13-14, 2021 and will take place in Little Rock for those who are vaccinated and feel comfortable doing so. A virtual meeting will take place with the External Advisory Board in May 2021 to provide them with information needed to compile the EAB report as part of the annual report process, but they will also be invited to the September conference to meet the participants and judge the student poster competition.

Retreat. Year 1 event cancelled due to prolonged pandemic

Project website published (Complete): A basic website presence has been published for the project at https://dart.cast.uark.edu/ and discussions are underway to further develop the DART web presence. Planned improvements include 1) adding details about each research theme and the faculty/graduate students involved; 2) adding relevant documents (such as the Strategic Plan, Arkansas S&T Plan, and Annual Reports). Additional improvements will be made as they are identified.

AEDC Blog. 4 blogs published. This activity will be completed by the end of Year 1. Two blogs have been published at the time of this report and two additional ones will be published before the end of Year 1. Blogs are posted at https://www.arkansasedc.com/news-events/arkansas-inc-blog.

Social Media Following increased by 10%. Content on Facebook and Twitter has been updated frequently during Year 1 and the following has been increased by 7% at the time of this report, but we are confident to meet the 10% goal by the end of Year 1. YouTube videos have typically been made to cover events and due to the lack of events, we have not made any new videos in Year 1. We plan to resume video production in Summer 2021. A communications intern has been hired at the central office to assist in content generation and publicity of DART, and will work with the AEDC marketing team to maximize exposure.

Campus Communications Committee formed; host first meeting.. Initial contact has been made with communications offices at most of the participating campuses. An initial meeting is planned for summer 2021. Efforts have also been made to collect social media accounts and blogs of DART faculty for cross-posting. Three listservs have been established: one each for the DART SSC; DART Project Faculty and Staff; and DART students. Additional listservs may be developed, as needed.

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 in ER Core.

Participants onboarded; 3 training webinars complete. Five trainings were held from September to January and 70% of users attended at least one training. Additional trainings will be conducted during the summer of 2021. The central office is currently participating in the ER Core Consortium and working with the hired developers to make continuous improvements and upgrades to the platform.

Scientific publications. The team did publish 9 peer-reviewed articles and juried conference papers that will be included in the publication list for Year 1 and reported in NSF PAR.

Statewide Workshops for Cohorts and Waves. 2 Workshops complete.

Science Journalism Committee formed; host first meeting. This activity has been postponed to Year 2.

10. Broadening Participation

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 will be reported in the Year 2 report.

The goals we committed to for diversity in the Broadening Participation (BP) plan are:

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

The project's Starting Participant Diversity* is:

- Faculty- 29% female, 6% URM
- Graduate Students- 38% female, 7% URM
- Undergraduate Students- 67% female, 25% URM
- Advisory Boards- 27% female, 12% URM

** As self-reported by participants*

The project's Y1 Participant Diversity is:

- Faculty- 29% female, 6% URM (no additional faculty to report)
- Postdocs- 50% female, 50% URM (two postdocs were hired)
- Graduate Students- 38% female, 11% URM
- Undergraduate Students- 52% female, 29% URM
- Advisory Boards- 27% female, 12% URM (no additional advisors to report)

** As self-reported by participants*

At the time of this report, DART has 135 participants and 16 confirmed advisory board members. Additional advisory board members are still being recruited.

Table 4. All Y1 DART Participants by Race and Ethnicity

Race and Ethnicity	**Count**
Asian	48
Asian, Caucasian	1
Black or African American	11
Black or African American, Native American, White or European American	1
Caucasian	20
Caucasian, Native American	1
Caucasian, White or European American	6
Hispanic	1
Hispanic, Latina / Latino	1
Hispanic, Latina / Latino, Native American, White or European American	1
Latina / Latino	1
Middle Eastern or North African	6
Middle Eastern or North African, White or European American	1
Prefer Not to Say	5
White or European American	31
Grand Total	135

First Generation Students. 14% of the DART undergraduate students identified as first-generation college students, and 38% of the graduate students identified as first-generation college students. 58% of the graduate students are first-generation graduate students.

Disabilities. None of the participants reported disabilities.

Veterans. Two participants identified as US Veterans.

Mentorship Program. Individual Development Plan (IDP) templates and guidelines were developed for the categories of participants outlined in the BP plan (SURE students, graduate students, early career faculty seed grant recipients and postdocs. They are included at the end of this document as appendices. The IDPs will be implemented starting in Year 2 at the beginning of the project participation for each role and reviewed at the end of each person's participation period, culminating in a survey. Additional progress on this will be reported in Year 2.

DART Research Seed Grant Program. At the time of this report filing, the project leadership team is working to finalize the request for proposals documents for the first round of seed grant solicitation. Each seed grant applicant will be required to submit a letter of support from an existing DART participant. Upon award, the SSC will assign mentors as appropriate to the awardees. Additional progress on this will be reported in Year 2.

Career Development Workshops. DART will host at least three career development workshops annually with three rotating topics: mentorship, grantsmanship, and science communication. These workshops will be open to all project participants and free to attend. Three workshops were held during Year 1: Mentorship as a Tool for Diversity, Communicating to Policymakers, and Individual Development Plans.

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. At the time of this report filing, the application window is open and the central office is accepting applications. Award information will be reported in Year 2.

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. Additional information will be reported in Year 2.

11. Expenditures and Unobligated Funds

Available funds shown on research.gov (ACM$) (as of July 9, 2021)	**$2,675,803.00**
Total of obligated Y1 funds	**$1,896,729.00**
Total of unobligated Y1 funds	**$779,074.00**

Y1 DART Funding Obligation

Category	Obligated	Timeline
Campus subagreement (9)	$1,379,888.00	August
All-hands in person meeting and Admin expense (including in-state participants travel)	$85,000.00	September
Research Seed Grant	$318,000.00	August
Summer Undergrad Students Support	$55,792.00	September
Education outreach summer activities	$2,500.00	August
K12 Activities in summer (joint effort with EAST Initiative)	$30,000.00	August
Education outreach Mini Seed Grant	$25,549.00	September
Total Y1 Obligated	**$1,896,729.00**	

12. Tabular/Graphic representation of progress to date (Attachment- Stoplight Tables)

13. Appendices

13.1. DART Mentorship Program Guidelines and Materials: For Seed Grant Faculty & Postdocs

Overview

An Individual Development Plan (IDP) is a personal action plan designed to help you take ownership of your training and professional development, set and achieve realistic goals, and clarify your academic responsibilities and expectations. You can tailor your IDP to your individual needs, with input and advice from your mentor. It can also be a useful launching point for discussing your long-term career interests.

Completion of your plan and assessments are considered mandatory as part of the DART Broadening Participation and Mentorship Plan, and progress related to this program will be included in each year's annual report to NSF. If you have any questions about these materials or the program in general, please contact the central office.

Process & Timeline

The first step is to complete the pre-survey form to assess your skills and articulate your goals. Next, you will make an action plan by setting specific goals and milestones to help you be more deliberate about your professional development and stay on track. At the end of the designated time period, you will complete a post-survey to assess your progress and help you plan for the future.

The plan template should be completed within 2 months of your DART participation start date and uploaded to your profile page in www.dartreporting.org. The post-survey should be completed within one month of the completion of your DART participation end date.

Tips & Best Practices

- Self-assessment: Try to be realistic when identifying your strengths and defining the areas that need development. Be sure to ask your mentor or other colleagues familiar with your work for feedback!
- Identify the skill sets you will need to pursue opportunities that interest you. What are your main priorities and how will you develop the necessary skills? What resources are available to help you? (Consider resources in your department, professional organizations, online courses/webinars, etc. If you can't find what you're looking for, ask around – your peers and your mentor may have ideas.)
- The best milestones are SMART:
  - Specific (Is the milestone focused and unambiguous?)
  - Measurable (What product or outcome will show you have achieved the milestone?)
  - Actionable (What action is required on your part?)
  - Realistic (Considering difficulty and timeline, is the milestone achievable?)
  - Timely (By what date will you complete the milestone?)

Project Website

Information about the DART project can be found on the website www.dartproject.org

Step 1: Self Assessment Pre-Survey

Review the questions in each section and respond in the space provided. Don't overthink it. Then, rate your skills on the scale given at the bottom of the page. After you have completed the written self-assessment, use it to complete your plan template.

Teaching & Mentoring

What are your teaching and mentoring goals? Have you ever taught, guest-lectured or served as a TA, or plan to in the next 12 months? What feedback have you received on your teaching? What teaching skills or knowledge would you like to improve? Have you ever served as a mentor? What qualities/skills do you associate with good mentoring? What mentoring skills would you like to improve?

Please evaluate your skills and abilities in the area of teaching and mentoring.

Use a scale of 1 (not at all proficient) to 5 (highly proficient). If something is not applicable, type N/A

Competency or Skill	Score
Familiarity with inquiry-based learning best practices	
Familiarity with active learning strategies	
Encouraging student participation	
Use of instructional technologies	
Providing constructive feedback	
Thoughtful listening	
Providing a culturally inclusive and supportive environment	
Providing career guidance	
Serving as a role model	
Seeking teaching/mentoring help when needed	

Research

What are your research goals? How do your research goals/objectives/activities contribute to DART's research themes (model interpretability, big data management, security and privacy, workforce development and education)? What research-related skills have you acquired to date? What research related skills would you like to develop or improve? What feedback have you received on your research project(s)?

Please evaluate your skills and abilities in the area of research.
Use a scale of 1 (not at all proficient) to 5 (highly proficient). If something is not applicable, type N/A

Competency or Skill	Score
Awareness of how your research contributes to DART	
Knowledge of concepts/theories related to your project	
Knowledge of past and current literature related to your project	
Laboratory skills	
Programming Skills	
Project management	
Data analysis including statistics	
Critical evaluation of data	
Creativity in designing experiments and new research directions	
Seeking research help/feedback when needed	

Collaboration
What leadership experiences have you had (e.g., organized a workshop, chaired a meeting)? What leadership experiences would you like to have? What collaborations have you established in the past? What new collaborations could benefit your research? How can you pursue them? What experience do you have with negotiation and conflict resolution?

Please evaluate your skills and abilities in the area of collaboration.
Use a scale of 1 (not at all proficient) to 5 (highly proficient). If something is not applicable, type N/A

Competency or Skill	Score
Identifying possible collaborations or collaborators	
Identifying support for collaborations	
Strategic planning or project management	
Ability to work in a team	
Ability to lead and motivate a team	
Respecting contributions and ideas of others	
Dealing with and resolving conflict	
Negotiating with a peer	
Negotiating with a more senior person (e.g., advisor)	

Communication & Dissemination

What writing or presentation skills would you like to improve? What resources are available? What research papers, proposals, or fellowship applications would you like to write in the next 12 months? Where could you present your research to peers within the next 12 months (e.g., at a lab meeting, seminar, conference)? Where could you present your research to a general audience within the next 12 months (e.g., blog, outreach event, local school presentation)? Do you use social media professionally?

Please evaluate your skills and abilities in the area of communication.

Use a scale of 1 (not at all proficient) to 5 (highly proficient). If something is not applicable, type N/A

Competency or Skill	Score
Communicating effectively in everyday conversation	
Presenting research to peers (e.g., seminar)	
Sharing research with a general (non-specialist) audience	
Effectively writing under time constraints	
Writing a peer-reviewed publication on your research	

Giving peer feedback on communication	
Receiving peer feedback on communication	
Social media communication & etiquette	
Writing a grant proposal or fellowship application	
Seeking communication help/feedback when needed	

Professional Development
What is your long-term career goal (e.g., college professor, environmental consultant, researcher in a government lab)? If you aren't sure, what information would help you decide? Are your CV and professional webpage up to date? Are you prepared for a job interview? If not, what should your next steps be to prepare? Are you aware of grant opportunities and how to submit proposals? Are you confident in your grant writing ability? Do you evaluate your past performance and consciously think about self-improvement? Do you reflect on your feelings and reactions to your work and work environment?

Please evaluate your skills and abilities in the area of career development.
Use a scale of 1 (not at all proficient) to 5 (highly proficient). If something is not applicable, type N/A

Competency or Skill	Score
CV/Resume writing	
Establishing career goals	
Grant writing and proposal development	
Awareness of career opportunities in your field	
Networking inside your academic environment	
Networking outside your academic environment	
Carving out time for career development	
Interviewing for a job	
Negotiating a job offer	
Self-reflection and evaluation	

Seeking career-related help/guidance when needed	

Diversity, Equity, & Inclusion

What does diversity mean to you? When have you or someone you know been treated inequitably? How can you help others to feel included/valued? How has your background or identity influenced your research? What implicit biases have you identified in yourself? What viewpoints do you tend to be dismissive of, and what resources would help you better understand those viewpoints? What would you like to learn about other cultures/viewpoints in your community? What do you wish people outside your community knew about your culture?

Please evaluate your skills and abilities in the area of DEI.

Use a scale of 1 (not at all proficient) to 5 (highly proficient). If something is not applicable, type N/A

Communicating respectfully with others	
Awareness of one’s own worldview/identity	
Knowledge of & respect toward other worldviews/identities	
Cultivating awareness of one’s own implicit biases	
Seeking out opposing viewpoints to improve understanding	
Effectively contributing to an inclusive work climate	
Incorporating cultural protocols and ethical standards into your research	
Awareness of DEI concerns in your discipline/community	
Familiarity with your organization’s diversity office and code of conduct/related policies	
Seeking help/guidance on DEI issues when needed	