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:	Steve Stanley
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

Table of Contents

Overview

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.

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 effects their lives. When they engage with data science products on their smart devices, on social media platforms, and on the web, the improved and robust privacy and safety protections and fair results increase their trust of data collection and the resulting information, allowing for broader use of data science to benefit society. In Arkansas, the educational ecosystem provides learners with a well-designed, consistent, sequenced, scaffolded, and modular education in data science with further educational and/or job opportunities available at appropriate points in their careers.

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.

1. **Big data management:** Before data streams and datasets can be used in the many kinds of learning models, they are often manually curated, or at the least, 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.
2. **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.
3. **Model interpretability:** Machine learning models often sacrifice interpretability for predictive power and are difficult to generalize beyond their training and test data. But interpretability and generalizability of trained models is critical in many decision-making systems and/or processes,

especially when learning from multi-modal and heterogeneous big data sources. There is a continuing to need to better balance the predictive power of complex machine learning models with the strengths of statistical models to better configure deep learning models to allow humans to see the reasoning behind the predictions.

These three barriers form the integrative research questions on which DART is focused. Activities in each research theme contribute to the integrative questions and the degree of interaction between themes is defined by that joint contribution. The flow of research from the research activities defined below, through the integrative questions, and finally into economic development through industry partners is captured in the Figure 2. 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.

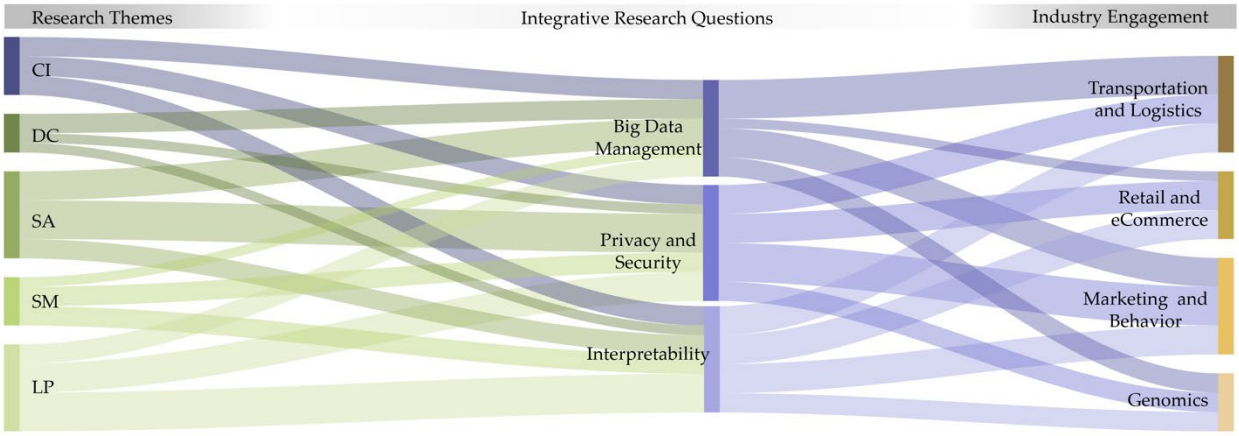


Figure 1: Sankey diagram showing 1) the contributions of each research theme to the fundamental barriers we are addressing as a project and 2) industry dependence on research to removing or mitigating the barriers. Research theme collaborations and integrations are designed to address these barriers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

Intellectual Merit

The **Learning and Prediction** research theme supports this through 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 goal of building a "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, and 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 solutions for these three particular uses cases (data cleaning, data integration, and data tracking) is an achievable goal within the 5-year time frame of the grant.

The **Social Awareness** 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 cut-edge research in socially ware 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 primarily includes 1) innovative methods, techniques, and platform 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.

The **Arkansas Research Platform (ARP)** is beginning to push 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 to researchers in the state of Arkansas, which will facilitate research, particularly big data analytics, that would have been out of reach to many researchers scattered across the smaller institutions within the state.

Broader Societal Impact

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

Big data analytics is a heavy consumer of compute 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 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, to 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 to be aware of 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 community in the endeavor to provide trustworthy technologies. The principles, methods, tools, datasets, and evaluation results will have significant effect on development of 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.

Include understanding the increasing societal polarization, amplified by the massive reach of "always on" social platforms, that 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. One of the direct applications of the research is in disaster management. 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 of well-paying jobs. As the State better aligns its investments with industry strengths and needs, more opportunities to improve the quality of life in Arkansas and steadily increase educational attainment and wages will develop. The HDR Big Idea recognizes that efforts in developing data cyberinfrastructure, education programs, and a deep workforce are most effective when linked to relevant data science research.

For the first time in AR EPSCoR history, we are working with 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

Roles and responsibilities in the Project

Science Steering Committee (SSC; also known as: Leadership Team): Comprised of CoLeads from each research theme. 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, annual conferences, and communicating progress to the external evaluation board and external evaluator via annual reports and presentations. The SSC works closely with the EAB, PI, and CoPI 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: Comprised of vice-provost level administrators from each campus receiving a subaward. The PI, CoPI, and Management Team are responsible for financial decisions and other administrative duties.

Science Advisory Committee (SAC; also known as the Arkansas EPSCoR Steering Committee): 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 EAB will include 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

Industry Advisory Board (IAB): The IAB members will serve as an intermediary between academia and industry. The IAB will include 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 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
Adita Karkera	Arkansas DIS	Deputy State Chief Data Officer
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

Summary of Year 1 Major Accomplishments

Activities in year 1 were designed to foster team coordination, collect data, review relevant literature, and build computing infrastructure. Significant accomplishments toward the integrative questions came from all research themes. We provide only a small sample here.

Big data management: The Cyberinfrastructure Theme focused on developed the necessary intra and intercampus relationships necessary to build the Arkansas Research Platforms. This has proved to be challenging and time-consuming but ultimately successful. Year 1 efforts centered on reconfiguring the UAF ScienceDMZ to create the necessary infrastructure to allow convenient but secure access to Pinnacle Portal and eventually to Grace Portal at UAMS via their Science DMZ. Data sharing among researchers is now available through Globus with endpoints at storage arrays at both UAF and UAMS and shared Git repository will soon be available for all DART resources to share code and replicate experiments as needed. An NSF mandated review of current cyberinfrastructure was led by James Deaton, executive director of Great Plains Network, under the auspices of the CC* CyberTeams project. A major test of the system involved the storing and distributing terabytes of social media data collected by the SM theme.

The DC team made very large strides implementing a variety of supervised and unsupervised clustering algorithms in their "data washing machine". For example, the machine includes sentence-BERT Zero Shot Learning for entity resolution, RoBERTa for similarity evaluation, and cluster-level data cleaning developed by this team. That code, along with large amounts of public federal contracting data on which to test it, is available in a Git repository (that will be mirrored in the DART repository) for easy access and testing by other DART researchers. This team will continue to work closely with the CI team to make this code available and easily accessible through resources on the ARP.

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.

The SA team designed a new cryptography-based scheme for differentially private federated learning. The scheme reduces the communication cost in the training process, a known problem of federated learning especially for edge devices that have limited network resources and improves the learning accuracy while providing differential privacy. Experimental results show that this scheme performs better than existing work in convergence rate and accuracy.

The SA team also completed an extensive literature review on the definition of personal identification information (PII) from different perspectives and their privacy issues, differentiating PII attributes only by their semantics as public or protected. Protected PII attributes should be subject to more stringent control compared to public PII. Identifying sensitive information from unstructured data covers a broad area of research including named entity resolution and identification, natural language processing, privacy ontology, etc. This theme is investigating appropriate techniques for unsupervised named entity resolution (NER) to remove its dependence on training data. Another SA project is working to prevent the privacy leakage by identifying 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. 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. This research focuses on assessing the cumulative sensitivity measure of the leaked PII attributes of an individual with a goal 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.

The SM team collaborated with policy makers and practitioners, conducted cyber campaign surveys to identify characteristics and features of social media platforms used by various campaigns and found that characteristics include platform type, language, purpose (connecting, social signaling, social news, collaboration, health, gaming, entertainment, etc.), and organic/inorganic behaviors. Features include content creation, content enrichment, content engagement, content streaming, connecting (friend, follow, groups, lists, etc.). Members of the team published a paper describing a multi-taxonomy characterization of social media data.

The CI theme has investigated how to architect access to resources on a Science DMZ and manage security risks associated with accessing non-CUI research data and code. This theme will engage with Trusted CI to assess DART needs and work with the other themes to implement standards that can be applied to general computation-oriented research.

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.

The SM team 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 research is the application of disparate impact and fairness testing to quantify unfairness in data and its effect on members of the unprivileged groups. It was clear there was an unbalanced division of income prediction concerning two sensitive attributes: gender (male, female) and ethnicity (white, black, others) and the team developed methods to understand the source of these biases.

The LP team developed a novel contextual guided convolutional neural network (CG-CNN) algorithm that can be trained on outputs from other models and applied it to natural images. They find that this algorithm showed 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. The team also advanced the understanding of dimensionality reduction and developed a library of classifiers that they used to address high-dimensionality issues of malware classification. Some features used for malware classification have good performance but require extensive learning time due to their high dimensionality. An entropy measure feature from this library of classifiers showed good performance with less training time. The team will employ this feature for windows malware, android malware and IoT malware.

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 COVID-19 impact analysis.

Research and Education Program

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.

Goal 1.1 (CI1)	Establish the Arkansas Research Platform as a shared data science resource across the jurisdiction

Lead: Cothren, Prior **Team Members:** Chaffin, Deaton, Springer, Tarbox

Objective 1.1.a: Establish the Arkansas Research Computing Collaborative (ARCC)

- A-1. Create ARCC advisory board with regional partners (GPN)
- A-2. Establish ARCC governance, operations, and staff between UA and UAMS
- A-3. Expand ARCC to include UALR as a provider and other DART participants as consumers
- A-4. Create UAF CI Plan to support DART (prior to 1.1.b and 1.1.c)

Objective 1.1.b: Upgrade cluster for data science research activity and integrate with existing resources

- A-1. Specify and purchase data science cluster based on document from 1.1.a
- A-2. Test and deploy hardware elements for Pinnacle expansion for DART
- A-3. Install and configure data science cluster to work with existing resources at UAF, UAMS, UALR resources

Objective 1.1.c: Establish a science DMZ in Little Rock (UAMS, UALR) and high-speed connection with UAMS

- A-1. Specify and purchase 100Gb switch
- A-2. Install 100Gb switch
- A-3. Establish ScienceDMZ at UAMS

Objective 1.1.d: Establish a data and code sharing environment (GitLab and Globus)

- A-1. Create/identify federated identify or other authentication mechanism for all sites that provides access to core ARP resources
- A-2. Setup dedicated GitLab repository
- A-3. Setup Globus Data Management Services
- A-4. Engage other research themes to develop research-specific training modules in e.g. Python, R, Git, HPC, Singularity
- A-5. Develop and deploy training materials for code sharing, large data transfer protocols

Objective 1.1.e: Establish necessary controls to store and manage controlled unclassified, HIPAA-related, and proprietary information at UA and UAMS (other institutions if possible)

- A-1. Identify the number and type (HIPAA, proprietary economic, CUI, etc.) of private and secure data sources that will need to be accessed.
- A-2. Setup capacity for storing and managing CUI and HIPAA data at UAF

Objective 1.1.a // A-1 Create ARCC advisory board with regional partners (GPN)

Year 1 Progress

Establish CI advisory board. 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

Objective 1.1.a // A-2 Establish ARCC governance, operations, and staff between UA and UAMS

Year 1 Progress

Create a document defining the organizational structure, roles, and responsibilities of ARCC. 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.

- **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 800-171-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 REN-ISAC 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 CWG will report to the Scientific Steering Committee (SSC) directly and through the CI Research Theme.

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.

Objective 1.1.a // A-3 *Year 1 Progress*	Expand ARCC to include UALR as a provider and other DART participants as consumers
No year 1 activities	
Objective 1.1.a // A-4 *Year 1 Progress*	Create UAF CI Plan to support DART (prior to 1.1.b and 1.1.c)
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.	
Objective 1.1.b // A-1 *Year 1 Progress*	Specify and purchase data science cluster based on document from 1.1.a

Issue UAF purchase order for additional equipment. Two quotes (one each from HPE and Dell) have been received and our 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 substatnial 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.

Objective 1.1.b // A-2 *Year 1 Progress*	Test and deploy hardware elements for Pinnacle expansion for DART
No year 1 activities	
Objective 1.1.b // A-3 *Year 1 Progress*	Install and configure data science cluster to work with existing resources at UAF, UAMS, UALR resources

Collect testbed specifications and osftware/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 Pinnnacle architecture.

Objective 1.1.c // A-1 *Year 1 Progress*	Specify and purchase 100Gb switch
No year 1 activities	
Objective 1.1.c // A-2 *Year 1 Progress*	Install 100Gb switch
No year 1 activities	
Objective 1.1.c // A-3 *Year 1 Progress*	Establish ScienceDMZ at UAMS

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.

Objective 1.1.d // A-1 *Year 1 Progress*	Create/identify federated identify or other authentication mechanism for all sites that provides access to core ARP resources

Establish federated ID for all project participants. 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 thier 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.

Objective 1.1.d // A-2 *Year 1 Progress*	Setup dedicated GitLab repository

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 be added in June 2021 as a guide to all researchers on minimum standards and best practices.

Objective 1.1.d // A-3 *Year 1 Progress*	Setup Globus Data Management Services

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.

Objective 1.1.d // A-4 *Year 1 Progress*	Engage other research themes to develop research-specific training modules in e.g. Python, R, Git, HPC, Singularity

No year 1 activities

Objective 1.1.d // A-5 *Year 1 Progress*	Develop and deploy training materials for code sharing, large data transfer protocols

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.

Objective 1.1.e // A-1 *Year 1 Progress*	Identify the number and type (HIPAA, proprietary economic, CUI, etc.) of private and secure data sources that will be need to be accessed by DART researchers.

Collect research theme needs. 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.

Objective 1.1.e // A-2 *Year 1 Progress*	Setup capacity for storing and managing CUI and HIPAA data at UAF
No year 1 activities	

Goal 1.2 (CI2) Visualization for complex data in diverse data-analytics application domains

Lead: Springer **Team Members:** Conde, Cothren, Huff, and Milanova

Objective 1.2.a: Investigate state-of-the-art visualization solutions

- A-1. Investigate/define state-of-the-art visualization
- A-2. Investigate standard tools for data science visualization
- A-3. Investigate/define data exchange strategies and their relationship to other research themes

Objective 1.2.b: Define domain-specific integration of visualization solutions

- A-1. Develop and deploy visualization infrastructure software
- A-2. Develop domain-specific visualization solution for DC
- A-3. Develop domain-specific visualization solution for SA
- A-4. Develop domain-specific visualization solution for SM
- A-5. Develop domain-specific visualization solution for LP

Objective 1.2.c: Introduce/integrate visualization for shared test beds

- A-1. Integrate visualization into existing testbeds for automated data curation environment DC/SM
- A-2. Integrate visualization into existing testbeds for social media-linked, GIS platform CI/DC/SM/LP
- A-3. Integrate visualization into existing testbeds for bioinformatics workflows DC/SM
- A-4. Integrate visualization into existing transaction-based testbed LP/DC
- A-5. Engage other research themes to develop research-specific advanced visualization training

Objective 1.2.a // A-1 *Year 1 Progress*	Investigate/define state-of-the-art visualization
1 presentation/report.	
Objective 1.2.a // A-2 *Year 1 Progress*	Investigate standard tools for data science visualization
No year 1 activities	
Objective 1.2.a // A-3 *Year 1 Progress*	Investigate/define data exchange strategies and their relationship to other research themes
No year 1 activities	

Objective 1.2.b // A-1 *Year 1 Progress*	Develop and deploy visualization infrastructure software
Collect research theme needs. Deployments are planned in years 2 and 3.	
Objective 1.2.b // A-2 *Year 1 Progress*	Develop domain-specific visualization solution for DC
No year 1 activities	
Objective 1.2.b // A-3 *Year 1 Progress*	Develop domain-specific visualization solution for SA
No year 1 activities	
Objective 1.2.b // A-4 *Year 1 Progress*	Develop domain-specific visualization solution for SM
No year 1 activities	
Objective 1.2.b // A-5 *Year 1 Progress*	Develop domain-specific visualization solution for LP
No year 1 activities	
Objective 1.2.c // A-1 *Year 1 Progress*	Integrate visualization into existing testbeds for automated data curation environment DC/SM
No year 1 activities	
Objective 1.2.c // A-2 *Year 1 Progress*	Integrate visualization into existing testbeds for social media-linked, GIS platform CI/DC/SM/LP
No year 1 activities	
Objective 1.2.c // A-3 *Year 1 Progress*	Integrate visualization into existing testbeds for bioinformatics workflows DC/SM
No year 1 activities	
Objective 1.2.c // A-4 *Year 1 Progress*	Integrate visualization into existing transaction-based testbed LP/DC
No year 1 activities	
Objective 1.2.c // A-5 *Year 1 Progress*	Engage other research themes to develop research-specific advanced visualization training
No year 1 activities	

2. Data Life Cycle and Curation

Why 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.

Goal 2.1 (DC1)	**Automate heterogeneous data curation**

Lead: Talburt **Team Members:** Cothren, Liao, Liu, Rainwater, Tudoreanu, Ussery, Wang, Xu, Yang

Objective 2.1.a: Automate Reference Clustering / Automate Data Quality Assessment

- A-1. Define metrics for data quality to measure impact of unsupervised data cleansing on data standardization and reference clustering.
- A-2. Set baseline data quality for initial test datasets used in prior research and acquire additional test datasets
- A-3. Curate test datasets and make available to other researchers
- A-4. Develop a framework for collaborative data collection and cleansing for knowledge discovery
- A-5. Develop a need- and prediction-based feedback mechanism for future data collection and making scalable decisions

Objective 2.1.b: Automate Data Cleansing

- A-1. Improve the unsupervised frequency-based data cleansing method used in prior POC; Explore and test alternative methods and models for unsupervised data cleansing including ML, AI, and graph approaches
- A-2. Migrate successful data cleansing models into a scalable HDFS processes

Objective 2.1.c: Automate Data Integration

- A-1. Improve the unsupervised frequency-based data integration method used in prior POC and explore and test alternative methods and models for unsupervised data integration including ML, AI, and graph approaches
- A-2. Migrate successful reference clustering models into a scalable HDFS processes

Objective 2.1.a // A-1 Define metrics for data quality to measure impact of unsupervised data cleansing on data standardization and reference clustering

Year 1 Progress

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.

  $$E = -\sum_{i=1}^{N} p_i \cdot \log_2(p_i)$$

  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.

Objective 2.1.a // A-2 Set baseline data quality for initial test datasets used in prior research and acquire additional test datasets

Year 1 Progress

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.

Objective 2.1.a // A-3 Curate test datasets and make available to other researchers

Year 1 Progress

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.

- vendor_contractingOfficerBusinessSizeDetermination
- vendor_vendorHeader_vendorAlternateName
- vendor_vendorHeader_vendorDoingAsBusinessName
- vendor_vendorHeader_vendorLegalOrganizationName
- vendor_vendorHeader_vendorName
- vendor_vendorSiteDetails_ccrRegistrationDetails_registrationDate
- vendor_vendorSiteDetails_ccrRegistrationDetails_renewalDate
- vendor_vendorSiteDeta__fEducationalEntity_is1862LandGrantCollege
- vendor_vendorSiteDeta__peOfGovernmentEntity_isPlanningCommission
- vendor_vendorSiteDetails_typeOfGovernmentEntity_isPortAuthority
- vendor_vendorSiteDeta__typeOfGovernmentEntity_isTransitAuthority
- vendor_vendorSiteDetails_vendorAlternateSiteCode
- vendor_vendorSiteDetails_vendorDUNSInformation_DUNSNumber
- vendor_vendorSiteDetails_vendorDUNSInformation_cageCode
- vendor_vendorSiteDetails_vendorLocation_ZIPCode
- vendor_vendorSiteDetails_vendorLocation_city
- vendor_vendorSiteDeta___vendorLocation_congressionalDistrictCode
- vendor_vendorSiteDetails_vendorLocation_countryCode
- vendor_vendorSiteDetails_vendorLocation_faxNo
- vendor_vendorSiteDetails_vendorLocation_phoneNo
- vendor_vendorSiteDetails_vendorLocation_state
- vendor_vendorSiteDetails_vendorLocation_streetAddress
- vendor_vendorSiteDetails_vendorLocation_streetAddress2
- vendor_vendorSiteDetails_vendorOrganizationFactors_annualRevenue
- vendor_vendorSiteDeta__rganizationFactors_countryOfIncorporation
- vendor_vendorSiteDeta__ndorOrganizationFactors_numberOfEmployee
- vendor_vendorSiteDeta__dorOrganizationFactors_organizationalType
- vendor_vendorSiteDeta__nshipWithFederalGovernment_receivesGrants
- vendor_vendorSiteDetails_vendorSiteCode

Objective 2.1.a // A-4 Develop a framework for collaborative data collection and cleansing for knowledge discovery

Year 1 Progress

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.

Objective 2.1.a // A-5 *Year 1 Progress*	Develop a need- and prediction-based feedback mechanism for future data collection and making scalable decisions

Formulate a framework for sequential data collection on an as-needed basis. 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.

Objective 2.1.b // A-1 *Year 1 Progress*	Improve the unsupervised frequency-based data cleansing method used in prior POC; Explore and test alternative methods and models for unsupervised data cleansing including ML, AI, and graph approaches

Document and train team on data cleansing methods developed in prior research.

- 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.

Objective 2.1.b // A-2 *Year 1 Progress*	Migrate successful data cleansing models into a scalable HDFS processes
No year 1 activities	
Objective 2.1.c // A-1 *Year 1 Progress*	Improve the unsupervised frequency-based data integration method used in prior POC and explore and test alternative methods and models for unsupervised data integration including ML, AI, and graph approaches

Document and train team on reference clustering method developed in prior research. 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.

- Min-Max ratio test for unlabeled paired samples
- 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.

Objective 2.1.c // A-2 *Year 1 Progress*	Migrate successful reference clustering models into a scalable HDFS processes
No year 1 activities	

Goal 2.2 (DC2) Explore secure and private distributed data management

Lead: Talburt **Team Members:** Talburt, Wang, Tudoreanu, Pierce, Liu, Rainwater

Objective 2.2: Build a POC and demo for Positive Data Control (PDC)

A-1. Build a POC and demonstration code for a Positive Data Control system layer forcing all the tools read/write operations to synchronize with the platforms metadata tool

A-2. Data exchange between PDC systems

Objective 2.2 // A-1 *Year 1 Progress*	Build a POC and demonstration code for a Positive Data Control system layer forcing all the tools read/write operations to synchronize with the platforms metadata tool
No year 1 activities	
Objective 2.2 // A-2 *Year 1 Progress*	Data exchange between PDC systems
No year 1 activities	

Goal 2.3 (DC3)	Harmonize multi-organizational and siloed data

Lead: Talburt **Team Members:** Wang, Tudoreanu, Pierce, Liu, Rainwater

Objective 2.3.a: Standardize pipelines for genome and proteome storage, retrieval, and visualization

- A-1. Define and download datasets to be curated
- A-2. Optimize data storage and retrieval
- A-3. Develop visualization methods

Objective 2.3.b: Automate quality scores for biological sequence data

- A-1. Develop pan-genome and Pan-proteome databases
- A-2. Develop taxonomy links to downloaded genomes/proteomes
- A-3. Develop a genomic database for Arkansas genomic pathogen surveillance of antimicrobial resistance

Objective 2.3.c: Apply machine learning methods to systems biology

- A-1. Define training sets to be used for ML
- A-2. Integrate multi-omic models for ML
- A-3. Benchmark ML results

Objective 2.3.a // A-1 Define and download datasets to be curated

Year 1 Progress

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.**

Objective 2.3.a // A-2 Optimize data storage and retrieval

Year 1 Progress

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.

Objective 2.3.a // A-3 *Year 1 Progress*	Develop visualization 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.

Objective 2.3.b // A-1 *Year 1 Progress*	Develop pan-genome and Pan-proteome databases

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.

Objective 2.3.b // A-2 *Year 1 Progress*	Develop taxonomy links to downloaded genomes/proteomes

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.

Objective 2.3.b // A-3 *Year 1 Progress*	Develop a genomic database for Arkansas genomic pathogen surveillance of antimicrobial resistance
No year 1 activities	
Objective 2.3.c // A-1 *Year 1 Progress*	Define training sets to be used for ML

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 meta-proteomics 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.

Objective 2.3.c // A-2 *Year 1 Progress*	Integrate multi-omic models for ML

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.

Objective 2.3.c // A-3 *Year 1 Progress*	Benchmark ML results
No year 1 activities	

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.

Goal 3.1 (SA1) Privacy Preserving and Attack Resilient Deep Learning

Lead: X. Wu **Team Members:** Q. Li. Zajiciek

Objective 3.1.a: Identify potential vulnerabilities of deep learning algorithms

- A-1. Research existing attacks including model inversion attacks and data poisoning attacks and capture mechanisms behind the threat models.
- A-2. Study the potential risks due to correlations among input data features, parameters, output, target victims, and latent feature space in deep learning algorithms
- A-3. Study the sensitivity and impact of input data features, parameters, and the objective functions on the model output and identify appropriate differential privacy preserving mechanisms for different computational components in a variety of deep learning models

Objective 3.1.b: Develop a universal threat- and privacy-aware deep learning framework

- A-1. Investigate the tradeoff of achieving privacy, resilience to adversarial attacks, and utility
- A-2. Study the mechanisms of redistributing injected noise across input data features, model parameters, and coefficients of objective functions based on their vulnerability and impact on the model output
- A-3. Develop and implement threat- and privacy-aware deep learning models

Objective 3.1.c: Conduct comprehensive evaluations of the proposed framework and models

- A-1. Evaluate the developed framework and models against baselines using benchmark datasets
- A-2. Evaluation and validation with participating companies

Objective 3.1.a // A-1 Research existing attacks including model inversion attacks and data poisoning attacks and capture mechanisms behind the threat models.

Year 1 Progress

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).

Objective 3.1.a // A-2 *Year 1 Progress*	Study the potential risks due to correlations among input data features, parameters, output, target victims, and latent feature space in deep learning algorithms

Initiate theoretical investigation on the risks of deep learning algorithms (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.

Objective 3.1.a // A-3 *Year 1 Progress*	Study the sensitivity and impact of input data features, parameters, and the objective functions on the model output and identify appropriate differential privacy preserving mechanisms for different computational components in a variety of deep learning models

Initiate theoretical investigation of privacy preserving mechanisms (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.

Objective 3.1.b // A1 *Year 1 Progress*	Investigate the tradeoff of achieving privacy, resilience to adversarial attacks, and utility
No year 1 activities	
Objective 3.1.b // A-2 *Year 1 Progress*	Study the mechanisms of redistributing injected noise across input data features, model parameters, and coefficients of objective functions based on their vulnerability and impact on the model output
No year 1 activities	
Objective 3.1.b // A-3 *Year 1 Progress*	Develop and implement threat- and privacy-aware deep learning models
No year 1 activities	
Objective 3.1.c // A-1 *Year 1 Progress*	Evaluate the developed framework and models against baselines using benchmark datasets
No year 1 activities	
Objective 3.1.c // A-2 *Year 1 Progress*	Evaluation and validation with participating companies

No year 1 activities

Goal 3.2 (SA2) Socially Aware Crowdsourcing

Lead: Hu **Team Members:** N. Wu, X. Wu

Objective 3.2.a: Improve crowdsourcing data quality with considerations of uncertainty

A-1. Allow uncertain labels in crowdsourcing data collection
A-2. Aggregate raw labels after label collection
A-3. Filter out possible noises to further improve data quality

Objective 3.2.b: Enhance available inference and learning models with novel algorithms for improved effectiveness and efficiency

A-1. Build theoretic foundations
A-2. Develop learning models and inference algorithms
A-3. Test and apply these learning models and algorithms

Objective 3.2.c: Verify and validate the robustness and trustworthiness of information from crowdsourcing data

A-1. Establish additional evaluation metrics
A-2. Develop algorithms to calculate the metrics
A-3. Verify and validate computational results

Objective 3.2.a // A-1 Allow uncertain labels in crowdsourcing data collection
Year 1 Progress

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.

Objective 3.2.a // A-2 Aggregate raw labels after label collection
Year 1 Progress

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.

Objective 3.2.a // A-3 Filter out possible noises to further improve data quality
Year 1 Progress

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.

Objective 3.2.b // A-1 *Year 1 Progress*	Build theoretic foundations

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.

Objective 3.2.b // A-2 *Year 1 Progress*	Develop learning models and inference algorithms

No year one activites were defined; however, 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.

Objective 3.2.b // A-3 *Year 1 Progress*	Develop learning models and inference algorithms

No year one activites were defined; however, 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.

Objective 3.2.c // A-1 *Year 1 Progress*	Establish additional evaluation metrics
No year 1 activities	
Objective 3.2.c // A-2 *Year 1 Progress*	Develop algorithms to calculate the metrics

No year 1 activities

Objective 3.2.c // A-3 Verify and validate computational results

Year 1 Progress

No year 1 activities

Goal 3.3 (SA3) User-centric Data Sharing in Cyberspaces

Lead: N. Wu **Team Members:** Q. Li, Hu

Objective 3.3.a: Investigate on personal identifying information and their privacy issues

A-1. Research state-of-art entity identification techniques for non-structure data

A-2. Investigate appropriate techniques for identifying context-aware sensitive information

A-3. Develop appropriate text analysis techniques to identify sensitive information from unstructured data

Objective 3.3.b: Investigate appropriate multimodal deep learning techniques to identify discriminative and stigmatizing information

A-1. Research state-of-art multimodal deep learning techniques for identifying private sensitive information Develop learning models and inference algorithms

A-2. Investigate appropriate techniques for identifying discriminating and stigmatizing information

A-3. Develop appropriate deep learning text analysis techniques to accurately remove discriminating and stigmatizing information

Objective 3.3.c: Develop a user-centric privacy monitoring and protection framework

A-1. Develop appropriate techniques for monitoring personal information disclosure on the Internet such as those from government records, news reports, and online documents

A-2. Develop a risk assessment method for possible privacy breach given the amount of personal identifying information disclosed/published

A-3. Develop appropriate techniques for safeguarding sensitive information by helping end users monitor and proactively control the release of their personal information

Objective 3.3.a // A-1 Research state-of-art entity identification techniques for non-structure data

Year 1 Progress

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.

Objective 3.3.a // A-2 *Year 1 Progress*	Investigate appropriate techniques for identifying context-aware sensitive information

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.

Objective 3.3.a // A-3 *Year 1 Progress*	Develop appropriate text analysis techniques to identify sensitive information from unstructured data

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.

Objective 3.3.b // A-1 *Year 1 Progress*	Research state-of-art multimodal deep learning techniques for identifying private sensitive information Develop learning models and inference algorithms

Study and document state of art multimodal deep learning techniques. Nothing to report

Objective 3.3.b // A-2 *Year 1 Progress*	Investigate appropriate techniques for identifying discriminating and stigmatizing information

No year 1 activities

Objective 3.3.b // A-3 *Year 1 Progress*	Develop appropriate deep learning text analysis techniques to accurately remove discriminating and stigmatizing information
No year 1 activities	
Objective 3.3.c // A-1 *Year 1 Progress*	Develop appropriate techniques for monitoring personal information disclosure on the Internet such as those from government records, news reports, and online documents
No year 1 activities	
Objective 3.3.c // A-2 *Year 1 Progress*	Develop a risk assessment method for possible privacy breach given the amount of personal identifying information disclosed/published
No year 1 activities	
Objective 3.3.c // A-3 *Year 1 Progress*	Develop appropriate techniques for safeguarding sensitive information by helping end users monitor and proactively control the release of their personal information
No year 1 activities	

Goal 3.4 (SA4) Deep Learning for Preventing Cross-Media Discrimination

Lead: L. Zhang **Team Members:** X. Wu, Sha, Zajicek

Objective 3.4.a: Explore deep learning-based techniques to detect cross-media discrimination

A-1. Use deep convolutional neural networks (CNN) to recognize discrimination-sensitive objects from images

A-2. Adopt long short-term memory (LSTM) network to model the text

A-3. Utilize bilinear model to capture the implicit relationship between the detected discrimination-related objects and the text

Objective 3.4.b: Design generative adversarial models to remove cross-media discrimination

A-1. Adopt mixture Generative Adversarial Nets framework for generating perceptually similar and discrimination-free image patches

A-2. Applies the encoder-decoder mechanism to automatically generate the discrimination-free text based on the original text

Objective 3.4.c: Develop a joint multi-modal deep learning framework to detect and prevent cross-media discrimination. Test and evaluate the proposed techniques and models with large-scale social media data

A-1. Develop a comprehensive multi-modal deep learning framework for jointly learning from both images and text to detect and prevent cross-media discrimination

A-2. Test and evaluate the proposed techniques and models from available data sources in social networks like Facebook, Instagram, and Foursquare

Objective 3.4.a // A-1 *Year 1 Progress*	Use deep convolutional neural networks (CNN) to recognize discrimination-sensitive objects from images

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.

Objective 3.4.a // A-2 *Year 1 Progress*	Adopt long short-term memory (LSTM) network to model the text

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.

Objective 3.4.a // A-3 *Year 1 Progress*	Utilize bilinear model to capture the implicit relationship between the detected discrimination-related objects and the text
No year 1 activities	
Objective 3.4.b // A-1 *Year 1 Progress*	Adopt mixture Generative Adversarial Nets framework for generating perceptually similar and discrimination-free image patches
No year 1 activities	
Objective 3.4.b // A-2 *Year 1 Progress*	Applies the encoder-decoder mechanism to automatically generate the discrimination-free text based on the original text
No year 1 activities	
Objective 3.4.c // A-1 *Year 1 Progress*	Develop a comprehensive multi-modal deep learning framework for jointly learning from both images and text to detect and prevent cross-media discrimination
No year 1 activities	
Objective 3.4.c // A-2 *Year 1 Progress*	Test and evaluate the proposed techniques and models from available data sources in social networks like Facebook, Instagram, and Foursquare
No year 1 activities	

Goal 3.5 (SA5) Marketing Strategy Design with Fairness

Lead: Sha **Team Members:** L. Zhang, X. Wu

Objective 3.5.a: Text mining and sentiment analysis of user-generated data from social media and consumer shopping records to extract customer-desired product features

A-1. Data collection from social media data, e.g., Amazon and Facebook, using vacuum product as the application context

A-2. Data collection and analysis of consumer panel data from Nielsen datasets at the Kilts Center for Marketing

A-3. Collection and analysis of marketing cases and/or ads with unfairness and exclusions

A-4. Text mining and sentiment analysis of the collected data for the development of metrics of fairness in marketing, and the identification of customer-desired product features

Objective 3.5.b: Network-based modeling of customer preference incorporating marketing parameters

A-1. Quantification and rating of bias and unfairness of marketing strategies and relate it to customer-desired product features

A-2. Network-based approach for choice modeling by incorporating customer preferences and perception to marketing (e.g., price) (un)fairness

A-3. Validation of the network-based choice modeling via demand prediction

Objective 3.5.c: Design of marketing strategies with fairness consideration and validate the approach

A-1. Parameterize marketing strategies and incentive design for improved advertisement with fairness consideration

A-2. What-if scenario analysis with the simulation of recommended marketing strategies and quantify the impact of those strategies on brand value and demand

A-3. Validation through human-subject experiment and surveys

Objective 3.5.a // A-1 Data collection from social media data, e.g., Amazon and Facebook, using vacuum product as the application context

Year 1 Progress

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.

Objective 3.5.a // A-2 Data collection and analysis of consumer panel data from Nielsen datasets at the Kilts Center for Marketing
Year 1 Progress

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.

Objective 3.5.a // A-3 *Year 1 Progress*	Collection and analysis of marketing cases and/or ads with unfairness and exclusions
No year 1 activities	
Objective 3.5.a // A-4 *Year 1 Progress*	Text mining and sentiment analysis of the collected data for the development of metrics of fairness in marketing, and the identification of customer-desired product features
No year 1 activities	
Objective 3.5.b // A-1 *Year 1 Progress*	Quantification and rating of bias and unfairness of marketing strategies and relate it to customer-desired product features

Document and disseminate the findings of researching the existing methods for fairness quantification and adverting parameterization (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, $DI_{ref.}$ 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 $DI_{ref.}$. 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.

$$DI_{ref} = \frac{P(A = unpriviliged)}{P(A = privileged\ )} \quad (1)$$

$$DI = \frac{P(A = unpriviliged)}{P(A = privileged\ )} \quad (2)$$

Fairness Testing. A fairness test was conducted based on the calibration scores using predicted probability score s 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 $Y = 1$ (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.

$$P(S = s, A = m) = P(S = s, A = f) \quad (3)$$

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.

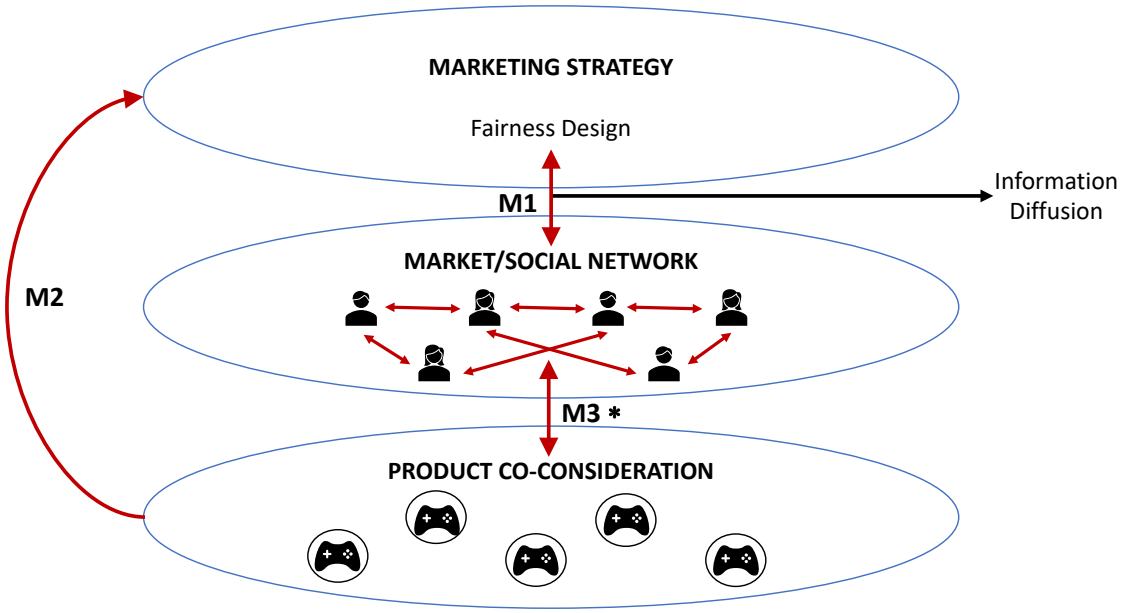


Figure 2: Research framework on Design for Market Systems

Objective 3.5.b // A-2 *Year 1 Progress*	Network-based approach for choice modeling by incorporating customer preferences and perception to marketing (e.g., price) (un)fairness
No year 1 activities	
Objective 3.5.b // A-3 *Year 1 Progress*	Validation of the network-based choice modeling via demand prediction
No year 1 activities	
Objective 3.5.c // A-1 *Year 1 Progress*	Parameterize marketing strategies and incentive design for improved advertisement with fairness consideration

No year one activites were defined; however, the team did begin 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.

Objective 3.5.c // A-2 *Year 1 Progress*	What-if scenario analysis with the simulation of recommended marketing strategies and quantify the impact of those strategies on brand value and demand
No year 1 activities	
Objective 3.5.c // A-3 *Year 1 Progress*	Validation through human-subject experiment and surveys
No year 1 activities	

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

Lead: Huang **Team Members:** Q. Li, M. Yang, Pierce, Ussery, CI Team

Objective 3.6.a: Design and develop machine learning algorithms and software, and advanced security and privacy technologies, for privacy-preserving data analytics

A-1. Design and develop machine learning and deep learning algorithms and software for privacy-preserving data analytics; Data and Infrastructure request and preparation

A-2. Develop privacy-preserving analytics algorithms, which will be based on high-dimensional tensor mathematical optimization model and combinatorial models

A-3. The optimization models will be incorporated with machine learning and deep convolutional neural network models

Objective 3.6.b: Train, test and validate the models and algorithms with publicly available data and some controlled genomics and health data; develop innovative frameworks and practical privacy-preserving techniques

A-1. Train, test and validate the models and the algorithms with publicly available data and some controlled genomics and health data

A-2. Study and develop innovative frameworks and practical privacy-preserving techniques, to facilitate centralized as well as distributed data analysis

Objective 3.6.c: Test the algorithms and technologies to work with a wide range of data types and high-dimensional heterogeneous data sources; Develop and deploy bioinformatics workflows into the private cloud environment, the ARP

A-1. Test the algorithms and technologies to work with a wide range of data types and high-dimensional heterogeneous data sources

A-2. Test these algorithms on available data sources from other research projects from this Track-I

A-3. Deploy bioinformatics workflows into the private cloud environment, the Arkansas Research Platform ARP, proposed in this Track-I

Objective 3.6.a // A-1 *Year 1 Progress*	Design and develop machine learning and deep learning algorithms and software for privacy-preserving data analytics; Data and Infrastructure request and preparation

Document and disseminate the findings of literature research of privacy-preserving data analytics 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.

Objective 3.6.a // A-2 *Year 1 Progress*	Develop privacy-preserving analytics algorithms, which will be based on high-dimensional tensor mathematical optimization model and combinatorial models

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.

Objective 3.6.a // A-3 *Year 1 Progress*	The optimization models will be incorporated with machine learning and deep convolutional neural network models
No year 1 activities	
Objective 3.6.b // A-1 *Year 1 Progress*	Train, test and validate the models and the algorithms with publicly available data and some controlled genomics and health data
No year 1 activities	
Objective 3.6.b // A-2 *Year 1 Progress*	Study and develop innovative frameworks and practical privacy-preserving techniques, to facilitate centralized as well as distributed data analysis
No year 1 activities	
Objective 3.6.c // A-1 *Year 1 Progress*	Test the algorithms and technologies to work with a wide range of data types and high-dimensional heterogeneous data sources
No year 1 activities	
Objective 3.6.c // A-2 *Year 1 Progress*	Test these algorithms on available data sources from other research projects from this Track-I
No year 1 activities	
Objective 3.6.c // A-3 *Year 1 Progress*	Deploy bioinformatics workflows into the private cloud environment, the Arkansas Research Platform ARP, proposed in this Track-I
No year 1 activities	

Goal 3.7 (SA7) Cryptography-Assisted Secure and Privacy-Preserving Learning

Lead: Q. Li **Team Members:** Huang, N. Wu

Objective 3.7.a: Develop privacy-preserving federated learning methods by combining cryptography techniques and privacy models

A-1. Research the hybrid use of existing cryptography techniques and differential privacy in federated machine learning
A-2. Develop new applied cryptography techniques to use in combination with differential privacy for federated machine learning
A-3. Develop unified security models for theoretical analysis of hybrid solutions

Objective 3.7.b: Explore how to protect the privacy of classification input data from the server hosting machine learning models

A-1. Develop methods for building/perturbing the model so that it can respond to encrypted or perturbed classification input
A-2. Study whether and how differential privacy can be achieved for classification input

Objective 3.7.c: Assess/Protect the trustworthiness of training data and machine learning models

A-1. Develop methods to assess the trustworthiness of training data and machine learning models

A-2. Develop methods to protect the trustworthiness of training data and machine learning models

Objective 3.7.a // A-1 Research the hybrid use of existing cryptography techniques and differential privacy in federated machine learning

Year 1 Progress

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.

Objective 3.7.a // A-2 Develop new applied cryptography techniques to use in combination with differential privacy for federated machine learning

Year 1 Progress

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.

Objective 3.7.a // A-3 *Year 1 Progress*	Develop unified security models for theoretical analysis of hybrid solutions
No year 1 activities	
Objective 3.7.b // A-1 *Year 1 Progress*	Develop methods for building/perturbing the model so that it can respond to encrypted or perturbed classification input
No year 1 activities	
Objective 3.7.b // A-2 *Year 1 Progress*	Study whether and how differential privacy can be achieved for classification input
No year 1 activities	
Objective 3.7.c // A-1 *Year 1 Progress*	Develop methods to assess the trustworthiness of training data and machine learning models
No year 1 activities	
Objective 3.7.c // A-2 *Year 1 Progress*	Develop methods to protect the trustworthiness of training data and machine learning models
No year 1 activities	

4. Social Media and Networks

What can we learn from Social Media and 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/dis-information, 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.

Goal 4.1 (SM1) Mining cyber argumentation data for collective opinions and their evolution

Lead: Zhan **Team Members:** Adams, S. Yang

Objective 4.1.a: Develop a cyber discourse social network platform

- A-1. Brainstorming the features needed for the cyber discourse social network platform.
- A-2. Software design for the platform
- A-3. Implement the platform
- A-4. Test the platforms
- A-5. Deploy and disseminate the platform

Objective 4.1.b: Collect data using the developed cyber discourse social network platform

- A-1. Design the "hot button" questionnaire items
- A-2. Develop Individual and network question measures and submit IRB consent form
- A-3. Data collection by engaging students in the "hot button" issue discussions

Objective 4.1.c: Develop natural language processing algorithms to analyze discourse data collected by the platform and existing data

- A-1. Develop advanced natural language processing algorithms
- A-2. Test the natural language processing algorithms using the existing data
- A-3. Validate the natural language processing algorithms using the data collected by the developed platform

Objective 4.1.a // A-1 Brainstorming the features needed for the cyber discourse social network platform.
Year 1 Progress

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.

Objective 4.1.a // A-2 *Year 1 Progress*	Software design for the platform

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.

Objective 4.1.a // A-3 *Year 1 Progress*	Test the platform

No year 1 activities

Objective 4.1.a // A-4 *Year 1 Progress*	Implement the platform

No year 1 activities

Objective 4.1.a // A-5 *Year 1 Progress*	Deploy and disseminate the platform

No year 1 activities

Objective 4.1.b // A-1 *Year 1 Progress*	Design the "hot button" questionnaire items

Develop questionnaire for collecting discourse data (this milestone for YR1 and YR2). 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.

Objective 4.1.b // A-2 *Year 1 Progress*	Develop Individual and network question measures and submit IRB consent form

The question measures are determined, and IRB protocol is approved (this milestone for YR1 and YR2). 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.

Objective 4.1.b // A-3 *Year 1 Progress*	Data collection by engaging students in the "hot button" issue discussions

No year 1 activities

Objective 4.1.c // A-1 *Year 1 Progress*	Develop advanced natural language processing algorithms

The advanced natural language processing algorithms are developed (this mileston for YR1 and YR2). 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.

Objective 4.1.c // A-2 *Year 1 Progress*	Test the natural language processing algorithms using the existing data
No year 1 activities	
Objective 4.1.c // A-3 *Year 1 Progress*	Validate the natural language processing algorithms using the data collected by the developed platform
No year 1 activities	

Goal 4.2 (SM2) Socio-computational models for safer social media

Lead: Agarwal **Team Members:** Trudeau, Zhan, Milburn

Objective 4.2.a: Characterize online information environment (OIE)

- A-1. Study social media spaces and cyber campaigns to identify characteristics and features
- A-2. Create a taxonomy of dimensions to characterize social media spaces
- A-3. Revisit and adjust taxonomy as social media space evolves

Objective 4.2.b: Develop socio-computational models to identify key actors and key groups of actors

- A-1. Review cyber campaigns and social media data
- A-2. Identify behavioral traits for key actors and key groups by leveraging OIE characterization
- A-3. Develop computational model(s) for key actor and key group discovery
- A-4. Evaluate model(s)

Objective 4.2.c: Study tactics, techniques, and procedures (TTPs) of deviant cyber campaigns

- A-1. Review campaigns, social media platforms, and involved actors and groups
- A-2. Identify and document tactics, techniques, and procedures (TTPs) (e.g., platform orchestration, botnets, inorganic behaviors, stalking, pacing, leading, threadjacking, hashtag latching, boosting, echo chambers)
- A-3. Examine tactics, techniques, and procedures (TTPs) vis-a-vis OIE characterization

Objective 4.2.d: Develop socio-computational models to measure power of a cyber campaign

- A-1. Review OIE characterization and TTPs to identify campaign attributes
- A-2. Develop computational model to measure power of a campaign by integrating attributes, key actors, key groups, collective action theory
- A-3. Evaluate model(s)

Objective 4.2.a // A-1 *Year 1 Progress*	Study social media spaces and cyber campaigns to identify characteristics and features

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 NATO-Turkey-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.

Objective 4.2.a // A-2 *Year 1 Progress*	Create a taxonomy of dimensions to characterize social media spaces

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 moviebarcode 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.

Objective 4.2.a // A-3 *Year 1 Progress*	Revisit and adjust taxonomy as social media space evolves
No year 1 activities	
Objective 4.2.b // A-1 *Year 1 Progress*	Review cyber campaigns and social media data

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 multi-threaded, 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.

Objective 4.2.b // A-2 *Year 1 Progress*	Identify behavioral traits for key actors and key groups by leveraging OIE characterization
No year 1 activities	
Objective 4.2.b // A-3 *Year 1 Progress*	Develop computational model(s) for key actor and key group discovery
No year 1 activities	
Objective 4.2.b // A-4 *Year 1 Progress*	Evaluate model(s)
No year 1 activities	
Objective 4.2.c // A-1 *Year 1 Progress*	Review campaigns, social media platforms, and involved actors and groups
No year 1 activities	

Objective 4.2.c // A-2 *Year 1 Progress*	Identify and document tactics, techniques, and procedures (TTPs) (e.g., platform orchestration, botnets, inorganic behaviors, stalking, pacing, leading, threadjacking, hashtag latching, boosting, echo chambers)
No year 1 activities	
Objective 4.2.c // A-3 *Year 1 Progress*	Examine tactics, techniques, and procedures (TTPs) vis-a-vis OIE characterization
No year 1 activities	
Objective 4.2.d // A-1 *Year 1 Progress*	Review OIE characterization and TTPs to identify campaign attributes
No year 1 activities	
Objective 4.2.d // A-2 *Year 1 Progress*	Develop computational model to measure power of a campaign by integrating attributes, key actors, key groups, collective action theory (and other theoretical constructs)
No year 1 activities	
Objective 4.2.d // A-3 *Year 1 Progress*	Evaluate model(s)
No year 1 activities	

Goal 4.3 (SM3) Auto-annotation of multimedia data

Lead: Milburn **Team Members:** Dagtas, Trudeau, Cothren

Objective 4.3.a: Develop multimedia indexing methods for social media data

A-1. Define priorities and characteristics for multimedia data on social platforms
A-2. Design and build algorithms for efficient retrieval of nontraditional data
A-3. Advanced querying methods and implementation of interfaces with multimedia capabilities

Objective 4.3.b: Design and implement deep learning methods for multimedia data

A-1. Define learning objectives for social data from multimodal sources
A-2. Develop detection and classification methods
A-3. Deep learning applied to multimedia data and related indexing mechanisms
A-4. Verification of learning objectives and methods developed in Activity 2 and 3

Objective 4.3.c: Build Integrated smart applications based on unstructured multimedia data

A-1. Define key applications for the implementation and testing of the indexing and retrieval mechanisms
A-2. Integration with disaster response and other applications defined in Activity 1
A-3. Define ethical and legal perspectives for the use of multimedia data

Objective 4.3.a // A-1 *Year 1 Progress*	Define priorities and characteristics for multimedia data on social platforms

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.

Objective 4.3.a // A-2 *Year 1 Progress*	Design and build algorithms for efficient retrieval of nontraditional data
No year 1 activities	
Objective 4.3.a // A-3 *Year 1 Progress*	Advanced querying methods and implementation of interfaces with multimedia capabilities
No year 1 activities	
Objective 4.3.b // A-1 *Year 1 Progress*	Define learning objectives for social data from multimodal sources

Identify and define three major learning objectives document (this mileston for YR1 and YR2). 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.

Objective 4.3.b // A-2 *Year 1 Progress*	Develop detection and classification methods
No year 1 activities	
Objective 4.3.b // A-3 *Year 1 Progress*	Deep learning applied to multimedia data and related indexing mechanisms
No year 1 activities	
Objective 4.3.b // A-4 *Year 1 Progress*	Verification of learning objectives and methods developed in Activity 2 and 3
No year 1 activities	
Objective 4.3.c // A-1 *Year 1 Progress*	Define key applications for the implementation and testing of the indexing and retrieval mechanisms

Three key applications defined (this mileston for YR1 and YR2). 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.

Objective 4.3.c // A-2 *Year 1 Progress*	Integration with disaster response and other applications defined in Activity 1

No year 1 activities

Objective 4.3.c // A-3 Define ethical and legal perspectives for the use of multimedia data
Year 1 Progress

No year 1 activities

Goal 4.4 (SM4) Auto-annotation of multimedia data

Lead: Milburn **Team Members:** Dagtas, Liao, Zhan., Cothren., Talburt, Ussery, Nachtmann, Rainwater, Karim, Celebi

Objective 4.4.a: Extract and index content describing transportation infrastructure status from social platforms

A-1. Study social platforms to identify types of content that describe transportation infrastructure status
A-2. Develop and implement extraction techniques for identified types of social platform content
A-3. Develop and implement indexing techniques for extracted social platform content

Objective 4.4.b: Fuse data from social platforms describing transportation infrastructure status with other data sources

A-1. Identify other data sources that contain real-time information regarding transportation infrastructure status
A-2. Obtain and index transportation infrastructure data from other data sources
A-3. Develop and implement data fusion techniques to combine data from social platforms and other sources

Objective 4.4.c: Assess credibility of data inputs from Objectives 4.4.a and 4.4.b

A-1. Develop and implement machine learning classifiers to detect quality of information
A-2. Develop and implement schema to map credibility/quality scores for data to probabilistic inputs of transportation infrastructure status

Objective 4.4.d: Develop routing algorithms that use inputs from Objectives 4.4.a-4.4.c to support routing for disaster response

A-1. Identify critical routing problems with application in disaster response
A-2. Develop models of identified disaster response routing problems and assess state of the literature
A-3. Develop and implement routing algorithms for identified routing problem variants
A-4. Implement GIS testbed capable of displaying and analyzing real-time road status and routing algorithm outputs
A-5. Demonstrate models and solution approaches via pilot study of one or more disaster scenarios

Objective 4.4.a // A-1 Study social platforms to identify types of content that describe transportation infrastructure status
Year 1 Progress

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.

Objective 4.4.a // A-2 Develop and implement extraction techniques for identified types of social platform content
Year 1 Progress

No year 1 activities

Objective 4.4.a // A-3 *Year 1 Progress*	Develop and implement indexing techniques for extracted social platform content
No year 1 activities	
Objective 4.4.b // A-1 *Year 1 Progress*	Identify other data sources that contain real-time information regarding transportation infrastructure status

Identify and define content types of interest (e.g., satellite imagery, traffic cameras) from sources other than social platforms. 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.

Objective 4.4.b // A-2 *Year 1 Progress*	Obtain and index transportation infrastructure data from other data sources
No year 1 activities	
Objective 4.4.b // A-3 *Year 1 Progress*	Develop and implement data fusion techniques to combine data from social platforms and other sources
No year 1 activities	
Objective 4.4.c // A-1 *Year 1 Progress*	Develop and implement machine learning classifiers to detect quality of information

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).

Objective 4.4.c // A-2 *Year 1 Progress*	Develop and implement schema to map credibility/quality scores for data to probabilistic inputs of transportation infrastructure status
No year 1 activities	
Objective 4.4.d // A-1 *Year 1 Progress*	Identify critical routing problems with application in disaster response

Select at least two disaster response routing problem variants using Milburn's existing qualitative interview data (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.

Objective 4.4.d // A-2 *Year 1 Progress*	Develop models of identified disaster response routing problems and assess state of the literature

Conduct literature review for identified routing problem variants and publish journal article synthesizing review with qualitative data from 4.4.c.1. 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.

Objective 4.4.d // A-3 *Year 1 Progress*	Develop and implement routing algorithms for identified routing problem variants
No year 1 activities	
Objective 4.4.d // A-4 *Year 1 Progress*	Implement GIS testbed capable of displaying and analyzing real-time road status and routing algorithm outputs

Define GIS system requirements (this mileston for YR1 and YR2). 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.

Objective 4.4.d // A-5 *Year 1 Progress*	Demonstrate models and solution approaches via pilot study of one or more disaster scenarios
No year 1 activities	

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.

Goal 5.1 (LP1)	**Statistical Learning – Random Forests for Recurrent Event Analytics**
Lead: Liu, X. **Team Members:** Chimka	
Objective 5.1.a: Create the Random Forests for Recurrent Event Analytics, which integrates the RF algorithm with classical statistical methods allows dynamic feature information to be incorporated into a tree-based method A-1. Establish a preliminary model, and complete the theoretical investigation A-2. Complete the coding and numerical examples; write, submit, revise paper A-3. Revise paper and research outcomes dissemination through conferences **Objective 5.1.b: Create the Gradient Boosting method for Recurrent Event Analytics, which integrates the boost trees with classical statistical methods allows dynamic feature information** A-1. Establish a preliminary model, and complete the theoretical investigation under Obj 5.2 A-2. Complete the coding and numerical examples; write, submit, revise paper A-3. Revise paper and disseminate research outcomes through conferences **Objective 5.1.c: Perform comparison study between the methodologies above and identify future research directions** A-1. Perform the comparison A-2. Write, submit, and revise paper A-3. Identify future research directions	
Objective 5.1.a // A-1 *Year 1 Progress*	Establish a preliminary model, and complete the theoretical investigation
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.

Objective 5.1.a // A-2 *Year 1 Progress*	Complete the coding and numerical examples; write, submit, revise paper
No year 1 activities	
Objective 5.1.a // A-3 *Year 1 Progress*	Revise paper and research outcomes dissemination through conferences
No year 1 activities	
Objective 5.1.b // A-1 *Year 1 Progress*	Establish a preliminary model, and complete the theoretical investigation under Obj 5.2
No year 1 activities	
Objective 5.1.b // A-2 *Year 1 Progress*	Complete the coding and numerical examples; write, submit, revise paper
No year 1 activities	
Objective 5.1.b // A-3 *Year 1 Progress*	Revise paper and disseminate research outcomes through conferences
No year 1 activities	
Objective 5.1.c // A-1 *Year 1 Progress*	Perform the comparison
No year 1 activities	
Objective 5.1.c // A-2 *Year 1 Progress*	Write, submit, and revise paper
No year 1 activities	
Objective 5.1.c // A-3 *Year 1 Progress*	Identify future research directions
No year 1 activities	

Goal 5.2 (LP2)	Statistical Learning – Marked Temporal Point Process (MTTP) Enhancements via LSTM Networks
Lead: Rainwater **Team Members:** Liu, X.	
Objective 5.2.a: Develop methodology integrating the MTPP with LSTM A-1. Formally define approach integrating intensity function of MTTP into LTSM A-2. Establish proof-of-concept implementation of MTTP/LTSM approach A-3. Perform benchmark of MTTP/LSTM tests on small simulated data sets **Objective 5.2.b: Create scalable implementation of MTTP/LSTM approach applicable to real-world data analysis scenario** A-1. Assess data collected from Activities 1 and 2 of Obj 5.2c to define methodology computation performance requirements A-2. Create version 2 implementation of approach using lessons learned from Activity 2 of Obj 5.2a A-3. Validate performance of scalable implementation on large-scale simulated datasets **Objective 5.2.c: Evaluate and assess MTTP/LSTM approach on real-world discrete data sets** A-1. Acquire healthcare IoT datasets A-2. Acquire civil infrastructure datasets A-3. Establish baseline performance of predictions made by existing approaches applied to datasets from Obj 5.2c Activities 1 and 2 A-4. Assess methodology implementation from Objective 2 on real-world datasets collected in Obj 5.2c Activities 1 and 2	
Objective 5.2.a // A-1 *Year 1 Progress*	Formally define approach integrating intensity function of MTTP into LTSM
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.	
Objective 5.2.a // A-2 *Year 1 Progress*	Establish proof-of-concept implementation of MTTP/LTSM approach
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.	
Objective 5.2.a // A-3 *Year 1 Progress*	Perform benchmark of MTTP/LSTM tests on small simulated data sets
No year 1 activities	
Objective 5.2.b // A-1 *Year 1 Progress*	Assess data collected from Activities 1 and 2 of Obj 5.2c to define methodology computation performance requirements
No year 1 activities	
Objective 5.2.b // A-2 *Year 1 Progress*	Create version 2 implementation of approach using lessons learned from Activity 2 of Obj 5.2a

No year 1 activities	
Objective 5.2.b // A-3 *Year 1 Progress*	Validate performance of scalable implementation on large-scale simulated datasets
No year 1 activities	
Objective 5.2.c // A-1 *Year 1 Progress*	Acquire healthcare IoT datasets
Publish curated data to GitHub. Nothing to report	
Objective 5.2.c // A-2 *Year 1 Progress*	Acquire civil infrastructure datasets
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.	
Objective 5.2.c // A-3 *Year 1 Progress*	Establish baseline performance of predictions made by existing approaches applied to datasets from Obj 5.2c Activities 1 and 2
No year 1 activities	
Objective 5.2.c // A-4 *Year 1 Progress*	Assess methodology implementation from Objective 2 on real-world datasets collected in Obj 5.2c Activities 1 and 2
No year 1 activities	

Goal 5.3 (LP3) Deep Learning – Novel Approaches

Lead: Karim **Team Members:** Celebi, Luu, Schrader, Kim, Kursun, Cheng, Alroobi.

Objective 5.3.a: Extract explanatory features from Deep Network

- A-1. Development of novel self-supervised and flow-based deep learning approaches
- A-2. Developing a library of classifiers for benchmarking
- A-3. Application of the developed methods on real-world datasets

Objective 5.3.b: Address high dimensionality issues in Deep Reinforcement Learning (DRL) using algebraic and topological methods

- A-1. Investigate group theoretical, and topological properties of generalized neural network architectures
- A-2. Apply group and topological structures of neural networks in the context of DRL

Objective 5.3.c: Designing a novel rewarding model, and addressing interpretability issues in DRL

- A-1. Design an improved reward process for DRL
- A-2. Explore PH based filtering to optimize scenario space
- A-3. Explore DRL interpretability

Objective 5.3.a // A-1 *Year 1 Progress*	Development of novel self-supervised and flow-based deep learning approaches

Development of the first unsupervised convolutional area and the first flow-based deep learning approach. 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.

Objective 5.3.a // A-2 *Year 1 Progress*	Developing a library of classifiers for benchmarking

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.

Objective 5.3.a // A-3 *Year 1 Progress*	Application of the developed methods on real-world datasets

Application of the developed methods with applications on natural images and textures, and classification of a malware dataset. 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 explore 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.

Objective 5.3.b // A-1 *Year 1 Progress*	Investigate group theoretical, and topological properties of generalized neural network architectures

Investigate group theoretical approaches to generalized NN architecture design within the context of interpretability for Objectives 5.3a (Activity 1) and 5.3c (Activity 1). 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.

Objective 5.3.b // A-2 *Year 1 Progress*	Apply group and topological structures of neural networks in the context of DRL
No year 1 activities	
Objective 5.3.c // A-1 *Year 1 Progress*	Design an improved reward process for DRL

Development of a generalized model of reward function in DRL addressing the issues with both sparse and dense feedback. 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.

Objective 5.3.c // A-2 *Year 1 Progress*	Explore PH based filtering to optimize scenario space
No year 1 activities	
Objective 5.3.c // A-3 *Year 1 Progress*	Explore DRL interpretability
No year 1 activities	

Goal 5.4 (LP4)	Deep Learning – Efficiency and Specification

Lead: Luu **Team Members:** Le, Rainwater

Objective 5.4.a: Create Novel Deep Learning Networks Executable with Reduced Computational Resources and Assess Performance

A-1. Develop and demonstrate new low-cost deep neural network algorithms
A-2. Develop new objective loss functions in deep neural networks
A-3. Develop low-cost deep learning methods for high-dimensional data

Objective 5.4.b: Address Low-cost Deep Learning Algorithmic Analysis and Challenges

A-1. Mathematically analyze the proposed deep learning methods
A-2. Improve the computational time and accuracy performance
A-3. Analyze the complexity

Objective 5.4.c: Explore Low-cost Deep Learning Applications in Natural Images and Medical Images

A-1. The developed deep learning algorithms will be optimized and implemented in two applications, including natural images and medical imaging
A-2. The developed deep learning algorithms will be further optimized and deployed in high dimension data, such as: videos and medical MRI volumetric data

Objective 5.4.a // A-1 *Year 1 Progress*	Develop and demonstrate new low-cost deep neural network algorithms

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 Light-weight Deep Learning Algorithms. Nothing to report

Objective 5.4.a // A-2 *Year 1 Progress*	Develop new objective loss functions in deep neural networks

No year 1 activities

Objective 5.4.a // A-3 *Year 1 Progress*	Develop low-cost deep learning methods for high-dimensional data

No year 1 activities

Objective 5.4.b // A-1 *Year 1 Progress*	Mathematically analyze the proposed deep learning methods

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.

Objective 5.4.b // A-2 *Year 1 Progress*	Improve the computational time and accuracy performance

No year 1 activities

Objective 5.4.b // A-3 *Year 1 Progress*	Analyze the complexity
No year 1 activities	
Objective 5.4.c // A-1 *Year 1 Progress*	The developed deep learning algorithms will be optimized and implemented in two applications, including natural images and medical imaging

Develop Low-cost Deep Learning Approaches in Image Classification. Nothing to report.

Objective 5.4.b // A-2 *Year 1 Progress*	The developed deep learning algorithms will be further optimized and deployed in high dimension data, such as: videos and medical MRI volumetric data
No year 1 activities	

Goal 5.5 (LP5) Harnessing Transaction Data through Feature Engineering

Lead: Zhang, S. **Team Members:** Nachtmann

Objective 5.5.a: Design advanced feature engineering techniques for high-dimensional temporal data

A-1. Extract and process APCD data
A-2. Extract and engineer features from the high-dimensional temporal data
A-3. Explore and test automation of feature engineering in transaction data

Objective 5.5.b: Create an improved prediction and decision-making framework incorporating feature engineering with health transaction data

A-1. Develop deep learning prediction models and algorithms with feature engineering
A-2. Incorporate representation learning in prediction with engineered features
A-3. Compare and validate prediction with existing models

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

A-1. Extract and process business transaction data
A-2. Employ feature engineering and prediction in the business transaction data
A-3. Validate the developed framework with the business transaction data

Objective 5.5.a // A-1 *Year 1 Progress*	Extract and process APCD data

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.

Objective 5.5.a // A-2 *Year 1 Progress*	Extract and engineer features from the high-dimensional temporal data
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.	
Objective 5.5.a // A-3 *Year 1 Progress*	Explore and test automation of feature engineering in transaction data
No year 1 activities	
Objective 5.5.b // A-1 *Year 1 Progress*	Develop deep learning prediction models and algorithms with feature engineering
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.	
Objective 5.5.b // A-2 *Year 1 Progress*	Incorporate representation learning in prediction with engineered features
No year 1 activities	
Objective 5.5.b // A-3 *Year 1 Progress*	Compare and validate prediction with existing models
No year 1 activities	
Objective 5.5.c // A-1 *Year 1 Progress*	Extract and process business transaction data
No year 1 activities	
Objective 5.5.c // A-2 *Year 1 Progress*	Employ feature engineering and prediction in the business transaction data
No year 1 activities	
Objective 5.5.c // A-3 *Year 1 Progress*	Validate the developed framework with the business transaction data
No year 1 activities	

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 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 educational 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.

Goal 6 (ED1)	**Contributing to the Data Science Educational Ecosystem by developing a combination of model programs, degrees, pedagogy, and curriculum including a 9-week middle school coding block; a technical certificate, certificate of proficiency, and associate of science in data science; and a Bachelor of Science in data science with minors or concentrations.**

Lead: Shubert, Addison **Team Members:** Chowdhury, Scott, Siddique, Swaid, Fowler

Objective 6.1.a: Create a full 9-week curriculum for the middle school coding block to help struggling K12 teachers meet state coding requirement and provide rich training to K12 students

- A-1. Hold a two-day workshop to include K20 computer science educators to outline the curriculum and pedagogy and establish the project timeline, roles, and deliverables.
- A-2. Develop curriculum
- A-3. Make curriculum publicly available through ADE and educational service co-ops
- A-4. Pilot curriculum in at least 6 schools.

Objective 6.1.b: Create a set of postsecondary programs of core courses with options for electives for a consistent set of Data Science Undergraduate degrees (e.g., Assoc. Degrees, 2+2 and "2, then 2"), Concentrations, and certificates

- A-1. Create the 5-year Plan to meet the Objective
- A-2. Identify the level of involvement and timing by academic institutions within the State
- A-3. Review UAF and UCA Data Science Programs with the Teams
- A-4. Convene workshops annually of engaged academic and government institutions to establish baseline
- A-5. Define Data Science Objectives and Outcomes base for defined degrees and certificates
- A-6. Define Data Science Courses Objectives, Learning Outcomes, and applicability to the defined degrees and certificates
- A-7. Dissemination of developed program details with collaborating institutions, government, and industry partners

A-8. Ensure defined programs are in line with appropriate accrediting bodies
A-9. Prepare and submit program proposals of each type at each level for appropriate approval
A-10. Evaluate progress and iteratively improve for future programs as appropriate
A-11. When accreditation is available, propose first-pass consultative reviews by those bodies for ready institutions
A-12. Prepare those institutions which do not have accredited programs in closely aligned areas for the pre-accreditation visit year
A-13. Identify appropriate accreditation by program and provide visibility to academic institution administrations
A-14. Create and maintain clearing house for course materials
A-15. Connect students, courses, problems, data, etc., with the Research Themes

Objective 6.1.a // A-1 *Year 1 Progress*	Hold a two-day workshop to include K20 computer science educators to outline the curriculum and pedagogy and establish the project timeline, roles, and deliverables

Workshop completed, plan finalized and disseminated to stakeholders. Due to the prolonged pandemic, we are postponing the face-to-face workshop until fall 2021. 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.

Objective 6.1.a // A-2 *Year 1 Progress*	Develop curriculum

No year 1 activities

Objective 6.1.a // A-3 *Year 1 Progress*	Make curriculum publicly available through ADE and educational service co-ops

No year 1 activities

Objective 6.1.a // A-4 *Year 1 Progress*	Pilot curriculum in at least 6 schools

No year 1 activities

Objective 6.1.b // A-1 *Year 1 Progress*	Create the 5-year Plan to meet the Objective

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.

Objective 6.1.b // A-2 *Year 1 Progress*	Identify the level of involvement and timing by academic institutions within the State

Cohorts identified; all collaborators assigned. We plan to complete this by the end of Year 1. 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 this spring to collect the additional responses needed and fulfill this milestone.

Objective 6.1.b // A-3 *Year 1 Progress*	Review UAF and UCA Data Science Programs with the Teams

1 meeting complete. 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. These efforts were spearheaded by DART personnel Stephen Addison and Emre Celebi. Dr. Celebi is the chair of UCA's Computer Science department.

Objective 6.1.b // A-4 *Year 1 Progress*	Convene workshops annually of engaged academic and government institutions to establish baseline

3 Workshops completed. One workshop has been convened to date, a second one is scheduled in April 2021, and the third one will be scheduled for June 2021. 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.

Objective 6.1.b // A-5 *Year 1 Progress*	Define Data Science Objectives and Outcomes base for defined degrees and certificates

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, is on the agenda to be presented and 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.

Objective 6.1.b // A-6 *Year 1 Progress*	Define Data Science Courses Objectives, Learning Outcomes, and applicability to the defined degrees and certificates

No year 1 activities

Objective 6.1.b // A-7 *Year 1 Progress*	Dissemination of developed program details with collaborating institutions, government, and industry partners

Info disseminated to stakeholders. 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.

Objective 6.1.b // A-8 *Year 1 Progress*	Ensure defined programs are in line with appropriate accrediting bodies

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.

Objective 6.1.b // A-9 *Year 1 Progress*	Prepare and submit program proposals of each type at each level for appropriate approval

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 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, which should take place in April 2021. Shorter also partnered with IBM and 4 faculty completed the data science and artificial intelligence badge programs.

Objective 6.1.b // A-10 *Year 1 Progress*	Evaluate progress and iteratively improve for future programs as appropriate
No year 1 activities	
Objective 6.1.b // A-11 *Year 1 Progress*	When accreditation is available, propose first-pass consultative reviews by those bodies for ready institutions
No year 1 activities	
Objective 6.1.b // A-12 *Year 1 Progress*	Prepare those institutions which do not have accredited programs in closely aligned areas for the pre-accreditation visit year
No year 1 activities	
Objective 6.1.b // A-13 *Year 1 Progress*	Identify appropriate accreditation by program and provide visibility to academic institution administrations
No year 1 activities	

Objective 6.1.b // A-14 *Year 1 Progress*	Create and maintain clearing house for course materials

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.

Objective 6.1.b // A-15 *Year 1 Progress*	Connect students, courses, problems, data, etc., with the Research Themes

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.

6.1. Major Accomplishments

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 focused for the first series of "spokes" at three HBCUs and will continue to focus on them as we bring the successive cohorts and waves into the program. The focus is on showing the cohorts and waves, by example and guidance, how to do what is necessary to instantiate the courses, programs, and 2+2 agreements and to leverage each other where appropriate. This includes sharing curricula and specific course syllabi as well as providing input into the information technology infrastructure needed to support the curriculum and providing up to date information

about industry-standard software packages needed to support state-of-the-art academic programs. The early cohorts and waves will serve as exemplars and to help us to initiate and develop programs across the State with each year thereby increasing the number of teachers for the "teach the teachers" and "teach the administrators" available and accelerating the number of institutions participating and, as a result, the number of instructors capable, and number of students engaged.

Integrated into Year 1 activities has been a focus on active participation and contributions by graduate research assistants and undergraduate research assistants. They are integrated into all activities and interface with all faculty and staff as contributors in research, planning, and execution. This introduces them to the processes of developing large architectures and systems and gives them exposure to, and experience with working with, senior administrators, faculty at all levels, and staff from all academic institutions, the AEDC, the ADHE, and the ACDS.

6.2. Challenges

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. 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.

Goal 7.1 (WD1)	**Provide K20 teacher and faculty opportunities for professional development spanning multiple disciplines.**

Lead: Fowler, Stanley **Team Members:** NA

Objective 7.1.a: Enable K12 teachers to integrate new Computer Science/Data Science technologies into their classrooms, empower students to hone their leadership skills, and provide

- A-1. Host one training session annually, issue technology kits to teacher participants
- A-2. Host two support/training webinars annually
- A-3. Establish platform for teachers to disseminate resources and troubleshoot
- A-4. Participate (booth or breakout) in EAST Initiative annual conference
- A-5. Invite participating teachers to annual All-Hands

Objective 7.1.b: Provide opportunities to support education and broadening participation related activities in the fields of Computer Science/Information Science/Data Science at K20 levels and both formal and informal settings

- A-1. Fund seed mini-grants annually at $5,000 each
- A-2. Recipients attend Annual All-Hands

Objective 7.1.c: Provide funding for faculty at collaborating institutions to learn new skills and tools in data science so they can effectively teach the new curriculum.

- A-1. Identify faculty to be trained and assign to 5 cohorts
- A-2. Fund faculty annually at $5000 each for training

Objective 7.1.d: Host workshops to enhance the research competitiveness of the state's faculty in data science and computing, particularly in grantsmanship and entrepreneurship technology for their classrooms.

- A-1. Host annual workshops on a variety of grantsmanship and entrepreneurship topics

Objective 7.1.a // A-1 *Year 1 Progress*	Host one training session annually, issue technology kits to teacher participants
No year 1 activities	
Objective 7.1.a // A-2 *Year 1 Progress*	Host two support/training webinars annually
No year 1 activities	
Objective 7.1.a // A-3 *Year 1 Progress*	Establish platform for teachers to disseminate resources and troubleshoot
No year 1 activities	
Objective 7.1.a // A-4 *Year 1 Progress*	Participate (booth or breakout) in EAST Initiative annual conference
1 Conference completed. 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.	
Objective 7.1.a // A-5 *Year 1 Progress*	Invite participating teachers to annual All-Hands
No year 1 activities	
Objective 7.1.b // A-1 *Year 1 Progress*	Fund seed mini-grants annually at $5,000 each
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.	
Objective 7.1.b // A-2 *Year 1 Progress*	Recipients attend Annual All-Hands
1 awardee presentation at All Hands meeting. One of the Year 1 awardees will be invited to present at the first DART face-to-face meeting in September 2021.	

Objective 7.1.c // A-1 *Year 1 Progress*	Identify faculty to be trained and assign to 5 cohorts

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.

Objective 7.1.c // A-2 *Year 1 Progress*	Fund faculty annually at $5000 each for training

10 faculty trained. This activity will be completed by the end of the year.

Objective 7.1.d // A-1 *Year 1 Progress*	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.

Goal 7.2 (WD2) Provide educational training opportunities inside and outside the classroom for students.

Lead: Fowler, Stanley **Team Members:** Schubert, Addison

Objective 7.2.a: Student Support at Participating Institutions: Support undergraduate research assistants during the fall and spring semesters at each participating primarily undergraduate institution, and graduate research assistantships at the academic research institutions

- A-1. Provide undergraduate research assistantships annually
- A-2. Host Provide graduate research assistantships annually (previously labeled as Activity 3)

Objective 7.2.b: Summer Internships: Facilitate industry internships for student participants at companies in relevant sectors and research centers.

- A-1. Identify internship opportunities for students at relevant companies
- A-2. Follow up with hosting companies for feedback and evaluation

Objective 7.2.c: Connect with other research thrusts to develop relevant research-based capstone projects

- A-1. Develop capstone projects annually
- A-2. Disseminate projects to all collaborating institutions
- A-3. Invite capstone students to present at annual All-Hands

Objective 7.2.d: ASRI- intensive data science and computing summer camps for undergraduates

- A-1. Host ASRI Annually and invite all DART undergrads
- A-2. Evaluate and revise programming based on student and presenter feedback

Objective 7.2.a // A-1 *Year 1 Progress*	Provide undergraduate research assistantships annually
15 UG supported (Complete). Please see participant funding table.	
Objective 7.2.a // A-2 *Year 1 Progress*	Host Provide graduate research assistantships annually (previously labeled as Activity 3)
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.	
Objective 7.2.b // A-1 *Year 1 Progress*	Identify internship opportunities for students at relevant companies
No year 1 activities	
Objective 7.2.b // A-2 *Year 1 Progress*	Follow up with hosting companies for feedback and evaluation
No year 1 activities	
Objective 7.2.c // A-1 *Year 1 Progress*	Develop capstone projects annually
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.	
Objective 7.2.c // A-2 *Year 1 Progress*	Disseminate projects to all collaborating institutions
3+ capstones published. This activity will be completed by August 2021.	
Objective 7.2.c // A-3 *Year 1 Progress*	Invite capstone students to present at annual All-Hands
No year 1 activities	
Objective 7.2.d // A-1 *Year 1 Progress*	Host ASRI Annually and invite all DART undergrads
1 ASRI Complete. 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.

Objective 7.2.d // A-2 *Year 1 Progress*	Evaluate and revise programming based on student and presenter feedback
No year 1 activities	

Goal 7.3 (WD3) Ensuring broad participation to impact the pipeline of data science skilled workers

Lead: Fowler, Stanley **Team Members:** Schubert, Addison

Objective 7.3.a: Summer Undergraduate Research Experiences for underserved students: Fund summer undergraduate research experiences (URE), for underserved students

A-1. Provide summer UREs to URM students annually

A-2. Students participate in annual All-Hands meeting poster competition

Objective 7.3.b: Scholarships for underserved students to the ASRI

A-1. Provide scholarships/recruit students annually

Objective 7.3.c: Connecting students to opportunities through the ACDS

A-1. Co-host statewide workshops on Data Science topics

A-2. Collaborate on Data Science apprenticeship programs- recruiting partners and developing curriculum

Objective 7.3.a // A-1 *Year 1 Progress*	Provide summer UREs to URM students annually

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.

Objective 7.3.a // A-2 *Year 1 Progress*	Students participate in annual All-Hands meeting poster competition
No year 1 activities	
Objective 7.3.b // A-1 *Year 1 Progress*	Provide scholarships/recruit students annually

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.

Objective 7.3.c // A-1 *Year 1 Progress*	Co-host statewide workshops on Data Science topics

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.

Objective 7.3.c // A-2 *Year 1 Progress*	Collaborate on Data Science apprenticeship programs- recruiting partners and developing curriculum

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.

8. Communication and Dissemination

How do we grow DART and gain public interest? Great consideration is given to how project communicaitons 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.

Goal 8.1 (CD1)	**Provide K20 teacher and faculty opportunities for professional development spanning multiple disciplines.**

Lead: Fowler, Stanley **Team Members:** Ford, Cothren

Objective 8.1.a: Day to Day Communication: Daily project-related communication will take place mostly via email and GitLab. If during the first year the project faces challenges with these two platforms, other platforms like Slack will be explored

A-1. Establish platform to maintain daily communication

Objective 8.1.b: Monthly Webinars & Component Meetings: Monthly webinar meetings will be held to share updates, events, and other news between faculty, students, administrators, and industry partners. Components will also meet monthly to manage project milestones and activities.

A-1. Host DART topical webinars monthly

A-2. Host monthly component team meetings

Objective 8.1.c: Face-to-Face Meetings: Two project-wide face-to-face meetings per year will be hosted. The Annual All-hands Meeting and Poster Competition will be attended by all project faculty, students, industry partners, administrative committee members, evaluators, and external advisory board members. The Annual Retreat will be for faculty and graduate student participants. These meetings will facilitate team building and foster a sense of collaboration among the group.

A-1. Host annual All-Hands meeting & poster competition

A-2. Host annual retreat for faculty and grad students

Objective 8.1.a // A-1 Establish platform to maintain daily communication

Year 1 Progress

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 re-established.

Objective 8.1.b // A-1 *Year 1 Progress*	Host DART topical webinars monthly

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 CoLeads
- January 2021: First Steps toward a Data Washing Machine, presented by Data Curation and Life Cycle CoLeads
- February 2021: Socially Aware Data Analytics, presented by Social Awareness CoLeads
- March 2021: Social Media and Networks, presented by Social Media and Networks CoLeads
- April 2021 (planned): Learning and Prediction, presented by Learning and Predication CoLeads
- June 2021 (planned): Education and Outreach, presented by Education and Outreach CoLeads

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.

Objective 8.1.b // A-2 *Year 1 Progress*	Host monthly component team meetings

11 meetings per component (6 components).

Objective 8.1.c // A-1 *Year 1 Progress*	Host annual All-Hands meeting & poster competition

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.

Objective 8.1.c // A-2 *Year 1 Progress*	Host annual retreat for faculty and grad students

1 Retreat. Year 1 event cancelled due to prolonged pandemic

Goal 8.2 (CD2) Educate the public about DART accomplishments

Lead: Fowler, Stanley **Team Members:** Ford, Cothren

Objective 8.1.a: Maintain public-facing communication outlets to inform public about DART

- A-1. Establish project website
- A-2. Maintain project website and refresh content at least quarterly
- A-3. Publish quarterly blog posts about DART on AEDC blog
- A-4. Maintain @arepscor Facebook, Twitter, and YouTube channels and refresh DART content frequently

Objective 8.1.b: Campus Communications: A project-wide communications team composed of communications staff from each participating institution, including AEDC, will be created. The communications team will use uniform citations and branding for all project-related releases. AEDC will issue press releases and blog posts related to overall project success, special events, and seed grant opportunities. The communications team will work together to release other pertinent information like new grant awards, patents, publications, and other highlights from each campus.

- A-1. Establish listserv and group of communications reps from each participating campus
- A-2. Hold annual check-in meetings to ensure proper citation of project and related messaging and disseminate project updates

Objective 8.1.c: Project Data: Project data will be submitted by participants to the project's internal reporting system, ER Core. Mandatory NSF reporting data will be collected, as well as additional information like startup companies and other major accomplishments.

- A-1. Establish DART ER Core Site
- A-2. Maintain DART ER Core site and provide annual training to participants

Objective 8.1.d: Technical dissemination channels: Project faculty will submit journal articles to scientific publications associated with data science and computing. Funds for travel stipends will be reserved to send students and faculty to national meetings related to data science and computer science research and education. Impacts and significant findings of research activities will be presented at these meetings. Relevant meetings include national meetings for professional societies and industry meetings.

- A-1. Presenting at national conferences / professional societies
- A-2. Publications
- A-3. Statewide Workshops for Cohorts and Waves

Objective 8.2.e: Science Journalism Challenge: Beginning in year 2, the project will host a statewide science reporting challenge. Journalism students (undergraduate and graduate level) from Arkansas institutions may apply to win first, second, or third place awards. Applicants will be paired with a project participant student and faculty member to develop a story about relevant research. Submissions will be evaluated by a committee of Arkansas public relations professionals, including representatives from major advertising agencies and news outlets. The first-place award will include publication in a major Arkansas news outlet. Award recipients will also be invited to present their stories at the annual all-hands meeting.

- A-1. Create committee responsible for implementing SJC
- A-2. Develop plan for SJC
- A-3. Host annual SJC
- A-4. Invite SJC winners to present at annual All-Hands

Objective 8.2.a // A-1 *Year 1 Progress*	Establish project website

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.

Objective 8.2.a // A-2 *Year 1 Progress*	Maintain project website and refresh content at least quarterly

No year 1 activities

Objective 8.2.a // A-3 *Year 1 Progress*	Publish quarterly blog posts about DART on 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.

Objective 8.2.a // A-4 *Year 1 Progress*	Maintain @arepscor Facebook, Twitter, and YouTube channels and refresh DART content frequently

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.

Objective 8.2.b // A-1 *Year 1 Progress*	Establish listserv and group of communications reps from each participating campus

Committee formed; host first meeting. This activity will be completed by the end of Year 1. Initial contact has been made with communications offices at most of the participating campuses. An initial meeting will be 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.

Objective 8.2.b // A-2 *Year 1 Progress*	Hold annual check-in meetings to ensure proper citation of project and related messaging and disseminate project updates

No year 1 activities

Objective 8.2.c // A-1 *Year 1 Progress*	Establish DART ER Core Site

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.

Objective 8.2.c // A-2 *Year 1 Progress*	Maintain DART ER Core site and provide annual training to participants

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.

Objective 8.2.d // A-1 *Year 1 Progress*	Presenting at national conferences / professional societies

No year one activites were defined; however, no travel took place during Year 1; some virtual conferences were attended.

Objective 8.2.d // A-2 *Year 1 Progress*	Publications

No year one activites were defined; however, 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.

Objective 8.2.d // A-3 *Year 1 Progress*	Statewide Workshops for Cohorts and Waves

2 Workshops complete.

Objective 8.2.e // A-1 *Year 1 Progress*	Create committee responsible for implementing SJC

Committee formed; host first meeting. This activity will be completed by the end of Year 1.

Objective 8.2.e // A-2 *Year 1 Progress*	Develop plan for SJC

Plan disseminated to stakeholders. This activity will be completed by the end of Year 1.

Objective 8.2.e // A-3 *Year 1 Progress*	Host annual SJC
No year 1 activities	
Objective 8.2.e // A-4 *Year 1 Progress*	Invite SJC winners to present at annual All-Hands
No year 1 activities	

9. Solicitation-Specific Project Elements

Describe your progress and achievements with respect to each of the additional project elements identified in the solicitation under which your award was made, such as Workforce Development, Diversity, Partnerships and Collaborations, etc. As with the Research and Education section above, provide quantitative information when available and appropriate, describe problems and opportunities encountered and your response, and summarize products. Evidence of linkages, coordination, and collaboration with other NSF-funded programs should be provided where appropriate. Refrain from re-descriptions of the end objectives and focus on specific accomplishments. Identify the principal individuals and institutions responsible for each major activity/accomplishment, as well as significant collaborations. A reasonable number of figures may be included in this section, as needed to assist in reporting.

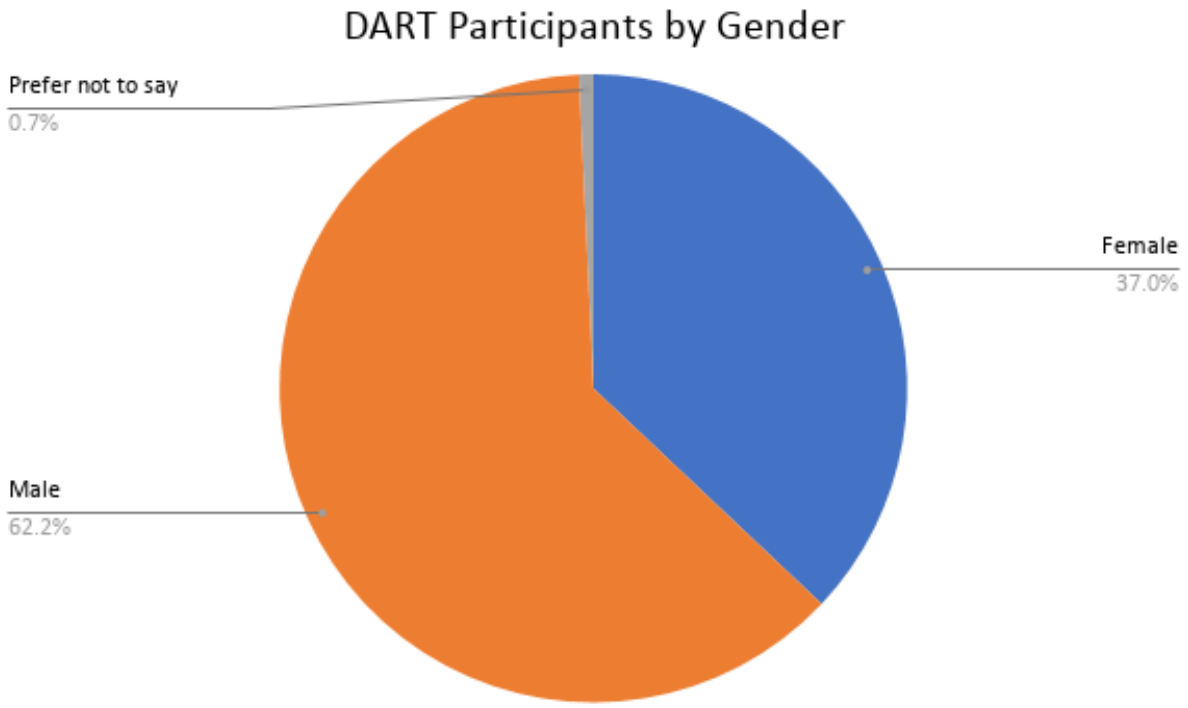
10. Broadening Participation

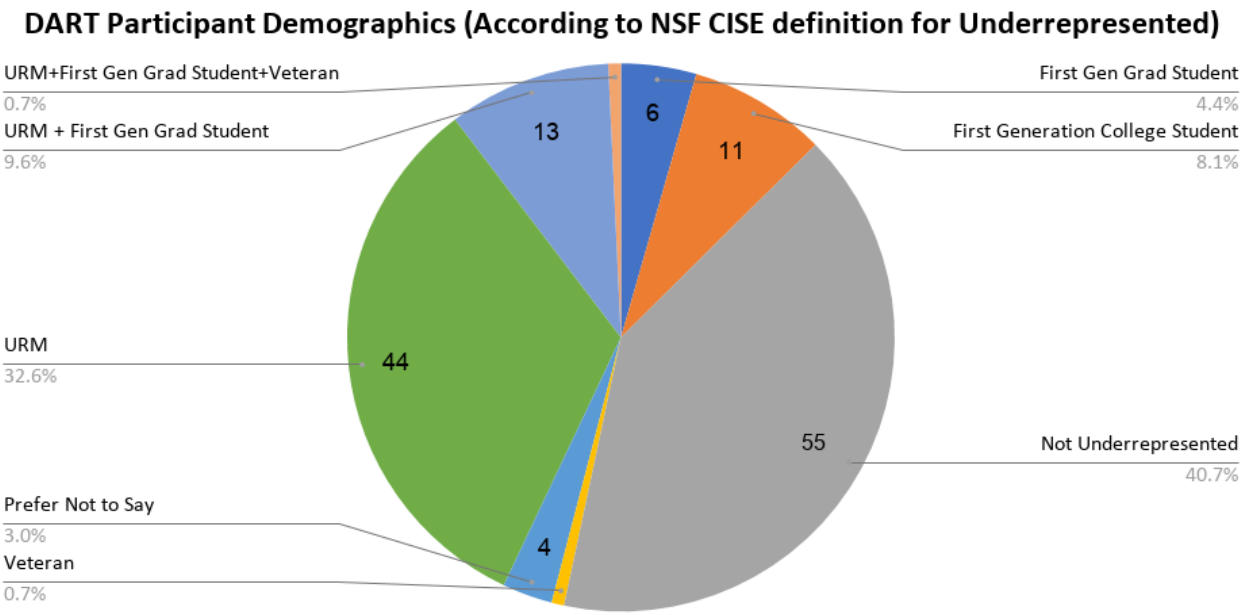
The numbers of participants and Federally required demographics can be found in Table B included as an attachment to this report. Below we have described our participants in more inclusive and representative terms than Table B allows. The gender and ethnic diversity of the project will be a challenge 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 project during Year 1 has 135 participants and 16 confirmed advisory board members at the time of this report. Additional advisory board members are still being recruited.

Table 4. 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


DART Participants by Gender
Prefer not to say
0.7%
Female
37.0%
Male
62.2%


DART Participant Demographics (According to NSF CISE definition for Underrepresented)
URM+First Gen Grad Student+Veteran
0.7%
First Gen Grad Student
4.4%
URM + First Gen Grad Student
9.6%
First Generation College Student
8.1%
URM
32.6%
Not Underrepresented
40.7%
Prefer Not to Say
3.0%
Veteran
0.7%
13
6
11
44
55
4

11. Expenditures and Unobligated Funds

As required by the Programmatic Terms and Conditions, reports should include an update on project spending and specifically an estimate of the funds expected to remain unobligated at the end of the current support period. If that estimate is greater than 20% of the current year award amount, the PI also must provide a plan and timeline for expenditure of those funds.

If more than 20% of the current year award amount continues to remain unobligated by the yearly anniversary date of the award, approval to carry forward that amount must be granted by NSF EPSCoR. The awardee's Sponsored Projects Office should prepare the request, which must include a plan and timeline for expenditure of the funds, and submit the request via email to the managing NSF PO.

12. Special Conditions

Include in the report specific information relating to: any outstanding actions taken or planned during the current reporting period in response to Jurisdiction-Specific Programmatic Terms and Conditions placed on the project at the time of the award; recommendations made through the Reverse Site Visit and Site Visit processes (if applicable); and any other actions required by NSF EPSCoR. Your external evaluation report or plan should be separately provided (C.1, below), but any significant changes to project activities implemented in response to external evaluator or advisory board recommendations may also be described in this section.

13. Tabular/Graphic representation of progress to date (Attachment)