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

Strategic Plan

October 5, 2020
Revision 1: January 20, 2022
Revision 2: July 10, 2023

Award Number: OIA-1946391
Jurisdiction: Arkansas
Start Date: July 1, 2020
End Date: June 30, 2025

Contents

List of Acronyms

ACDS	Arkansas Center for Data Science
ADHE	Arkansas Department of Higher Education
AEDC	Arkansas Economic Development Commission
AI	Artificial Intelligence
API	Application Programming Interface
APCD	All-Payer Claims Database
ARCC	Arkansas Research Computing Collaborative
ARGO	The Great Plains Augmented Regional Gateway to the Open Science Grid
ARP	Arkansas Research Platform
ASMSA	Arkansas School for Mathematics, Sciences, and the Arts
ASRI	Arkansas Summer Research Institute
ASU	Arkansas State University
BD2K	Big Data to Knowledge
BIGDATA	NSF-NIH Interagency Initiative: Core Techniques and Technologies for Advancing Big Data Science and Engineering
CCPA	California Consumer Protection Act
CDS&E	NSF program: Computational and Data-Enabled Science and Engineering
CI	Coordinated Cyberinfrastructure Research Theme; also Cyberinfrastructure
CITI	Collaborative Institutional Training Initiative
COVID	Corona Virus Disease
CSTA	Computer Science Teacher Association
CUI	Controlled, Unclassified Data
DART	Data Analytics that and Robust and Trusted: From Smart Curation to Socially Award Decision Making
DC	Data Curation and Life Cycle Research Theme
DG	Data governance
EAB	Employee Advisory Board
EAST	EAST Initiative
ED	Education Research Theme
EPSCoR	NSF program: Established Program to Stimulate Competitive Research
FIS	a financial software company with offices in Arkansas
GDPR	General Data Protection Regulation
GRA	Graduate Research Assistant
HBCU	Historically Black Colleges and Universities
HDFS	Hadoop Distributed File System
HIPAA	Health Insurance Portability and Accountability Act of 1996
HPC	High Performance Computing

IHE	Institute of Higher Education
IUCRC	NSF program: Industry-University Cooperative Research Centers
LP	Learning and Prediction Research Theme
LSAMP	Arkansas Louis Stokes Alliance for Minority Participation
LSTM	long short-term memory
MTPP	marked temporal point process
NASA	National Aeronautics and Space Administration
NASEM	National Academies of Sciences, Engineering, and Medicine
NHPP	Non-Homogeneous Poisson Process
NRT	NSF program: NSF Research Traineeship Program
ML	Machine Learning
MT	Management Team
OURRstore	The Oklahoma University (OU) & Regional & Research Store
PDC	Positive Data Control
PII	Personal Identifying Information
POC	Proof of Concept
RF	Random Forest
SA	Social Awareness Research Theme
SAC	Science Advisory Committee; also known as the Arkansas EPSCoR Steering Committee
SAU	Southern Arkansas University
SBIR	Small Business Innovation Research
SLURM	Simple Linux Utility for Resource Management
SM	Social Media and Networks Research Theme
SSC	Science Steering Committee; also known as the Leadership Team
STC	NSF program: Science and Technology Centers
STEM	Science, Technology, Engineering, and Mathematics
STTR	Small Business Technology Transfer
SURE	Summer Undergraduate Research Experience
SWOT	Strengths, Weaknesses, Opportunities, and Threats
UAF	University of Arkansas, Fayetteville
UALR	University of Arkansas at Little Rock
UAMS	University of Arkansas for Medical Sciences
UAPB	University of Arkansas at Pine Bluff
UCA	University of Central Arkansas
UGRA	Undergraduate Research Assistant
WCOB	Walton College of Business, University of Arkansas, Fayetteville
WD	Workforce Development
XSEDE	NSF program: Extreme Science and Engineering Discovery Environment

0. DART Strategic Plan

0.1. Overview

DART is part of a much larger State effort to establish an educational and research foundation for expanding and supporting a sustainable data science sector within our economy. Data science is targeted in this 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 UAF Data Science B.S. program that serves as a basis for this project's education component was developed because business leaders across the state - including those from the Fortune 500 companies named above - were pleading with universities in the state to help develop the data science savvy workforce necessary for them to remain competitive in fast-changing, data intensive markets. The pleas led to a coordinated plan among academia, industry, and government. At UAF, faculty from three colleges (business, engineering, and arts and sciences) developed and implemented an integrated Data Science B.S. program which has been approved by the Arkansas Department of Higher Education (ADHE) and began enrolling students in Fall 2020. The University also named data science as one of three Signature Research Areas and provided institutional resources (funding and support) for its development on campus. Industrial partners backed up their pleas with action and became active and vocal advisors of the data science program. The trajectory of data science at UAF is a clear result of a coordinated effort between industry who expressed economy-driven needs, the academy which responded with a rigorous and substantial program to meet those needs, and state government that provided additional financial support and other institutional resources.

The necessary pieces are now in place across the state to replicate this trajectory: the 2018 Arkansas Science and Technology Plan identifies Data Science and Analytics as one of three Targeted Priorities in the state; data science programs are growing across the state - UCA offers a BS in Data Science (initiated by SP Addison), ASU, and UALR (SP Tudoreanu) all offer graduate certificates in Data Science, and similar but more focused programs exist at UAMS and UAPB; many senior faculty at participating institutions have robust data science research programs and junior faculty are looking for collaborators. DART, led by the state EPSCoR office and with a state-wide focus, completes the puzzle by providing necessary organizational structure and shared cyberinfrastructure. DART defines a coordinated research plan composed of five topical areas related to data science needs identified by an expanded (beyond the UAF advisory board) number of industrial partners who drive a significant portion of our state's economy. These topical areas are distinct but interrelated and are intentionally composed of researchers from different academic institutions across the state. Using the same relationships that made data science into a substantive, cohesive program at UAF, DART will establish a state-of-the-art data science research ecosystem involving multiple campuses across the state and develop a close interactive relationship between our university researchers and data science dependent industries. This relationship will expand the number of high-paying data science jobs in Arkansas and the resulting expansion of the data science sector of our economy will

depend on a large continuous flow of well-trained data science capable talent. DART will include support for programs degrees and curriculum extending from middle school through graduate school to support this workforce pipeline.

0.2. 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.

0.3. 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.

0.4. Advancing the State of the Knowledge

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.
4. **Data-Skilled Workforce:** As data-driven science and decision making become commonplace, our state and nation will need to rely on a well-educated workforce at almost all levels of responsibility to be aware of the power and pitfalls of using data in decision making. This topic represents a significant addition in year 2. It is a natural and effective way to think about how education and workforce development efforts integrate with research efforts.

These barriers form the integrative research questions on which DART will focus. Activities in each research theme contribute to the integrative questions and the degree of interaction between themes is defined by that joint contribution.

0.5. 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, we will create statistical learning methods in big data environments that are equipped with capabilities for addressing heterogeneity and hidden sub-populations within big datasets. In addition, contributions will be made in mode specification and interpretation through the contribution of efficient variable selection using non-parametric methods. Lastly, we will advance computing in big data environments for traditional statistical modeling through statistical computing performed on distributed/parallelized computing nodes. Holistically, our 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 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.

Social Awareness research theme will greatly 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 Education team will collaborate with colleges and universities across Arkansas to introduce data science and data analytics degree and certificate offerings, designed 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.

The Arkansas Research Platform (ARP) will push the edge of distributed high-performance computing coupled to distributed high performance storage via high bandwidth networks. 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.

0.6. Broader Societal Impact

DART, as a center, is integrating data science research across the State and creating a deep and diverse data-ready workforce, which will pay immediate 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.

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. The proposed 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.

0.7. Overall Project Implementation

Management

A senior management team oversee research activities within and across the topic areas. It will be the responsibility of this team to ensure that collaborative activities are ongoing, productive, and fall within the defined research goals of the project.

ARP will be managed as a unique multi-institutional resource. The co-leads will serve as the Leadership Team for this resource and direct the management of resources at their respective sites. Dr. Cothren will serve as the executive director. The leadership team will define the operational procedures for the ARP combined resource in consultation with a user committee comprised of major users from each campus. A memorandum of understanding among the campuses participating in ARP will define the governance structure and establish operational parameters. This governance and operations model is based on our experience operating the established facilities at UAF, UAMS, and UALR. Direct support to research faculty and students will be provided by existing staff at UAF and UAMS, with additional faculty and students from UALR assisting in the development of testbed solutions.

Our management plan reflects the previous EPSCoR project organizational structure. It includes a state-supported Central Office which provides general oversight for the project and coordinates interactions with state boards and agencies; a management team (MT) comprised of administrators from participating campuses, Table 1, to ensure project implementation on campuses and information flow; a researcher-led Science Steering Committee (SSC) provides oversight for the scientific aspects of the program; and one or more external advisory boards contribute stakeholder perspective and facilitated dissemination of results to other groups.

Table 1: Management team comprised of administrators from campuses receiving subawards.

Name	Role	Institution
Jennifer Fowler	PI, EPSCoR PD	Arkansas Department of Commerce
Jackson Cothren	Co-PI, Science Director	University of Arkansas, Fayetteville
Brian Berry	Administrator, MT	University of Arkansas, Little Rock
Bob Beitle	Administrator, MT	University of Arkansas, Fayetteville
Travis Marsico	Administrator, MT	Arkansas State University
Angela Barlow	Administrator, MT	University of Central Arkansas
Shuk-Mei Ho	Administrator, MT	University of Arkansas for Medical Sciences
Mansour Mortazavi	Administrator, MT	University of Arkansas, Pine Bluff
Anthony Johnson	Administrator, MT	Philander Smith College
Demetrius Gilbert	Administrator, MT	Shorter College
Abdel Bachri	Administrator, MT	Southern Arkansas University

Central Office: Jennifer Fowler Arkansas NSF Program Director serves as the project PI/PD. Fowler is responsible for the overall statewide project management, which includes administration of the central office, overall management of science, outreach, and workforce development efforts, cyberinfrastructure, and evaluation, developing new funding opportunities to leverage or add support and linkage with other federal grants, and providing progress report to and communicating with NSF EPSCoR Office. Cathy Ma, Assistant Director and Program Administrator (PA), is responsible for day-to-day program operations, assisting with financial transactions, organizing events, and managing project staff. Brittany Hillyer, Director of Education, Outreach, and Diversity will expand the education and workforce development capacity of the project and coordinate activities to broadening participation. Hillyer also administers the reporting database (ER Core) and communication activities.

Succession Plan: Jim Hudson, Chief of Staff for Arkansas Department of Commerce, and Chase Rainwater, Associate Professor of Industrial Engineering at the UAF, will take the responsibility of the PI and Co-PI, respectively, in the event that Fowler and Cothren cannot perform their duty.

Roles and responsibilities within the Research Theme

Each faculty and staff participant in DART will be assigned one or more roles in DART. Individual assignments are based on research expertise and interest and designed to foster collaboration and increase communications across the research themes, cyberinfrastructure development, and education and outreach components. Information about roles and responsibilities will be posted on the project website throughout the project.

Research Theme Co-lead: A Co-lead has two important administrative roles: 1) serve as liaison between SSC and the respective research theme faculty groups; 2) lead the respective research theme in maintaining synchronous progress across all activities within the research theme and integrated research themes, including but not limited to: coordinating meetings and distributing communications among research theme faculty members; communicating progress to the SSC; communicating upcoming or existing personnel vacancies to the SSC. The Co-lead

role is primarily an administrative role and does not include any additional resources or support.

Research Theme Sr. Personnel: In each research theme activity, a number of faculty within that theme are assigned as senior personnel. They each have particular expertise and skills needed to accomplish the milestones in that activity. Faculty may be senior personnel in multiple activities.

Research Theme Liaison: Liaisons have the task of disseminating information between research themes. Individuals with this role have subject area expertise that is relevant in multiple themes and across multiple activities. While they are not necessarily responsible for milestones in their cross-theme activity, they should: 1) attend meetings and keep up with progress in that activity, and 2) inform their theme of relevant results and possible collaborations.

Roles and responsibilities in the Project

Science Steering Committee (SSC; also known as: Leadership Team): Comprised of Co-leads 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 includes researchers from peer and aspirant universities or national labs who serve as technical consultants providing recommendations on research progress and strategic and long-term sustainability planning during annual site visits. The EAB serves a critical role in the seed grant program as well as in mentoring and commercialization efforts.

Industry Advisory Board (IAB): The IAB members serve as an intermediary between academia and industry. The IAB includes representatives from Arkansas industry sectors who will be impacted by DART research. One member of the IAB will serve as a member of the EAB during site visits or annual meetings as needed.

Sustainability

DART is an important part of Arkansas' plan to transition to a knowledge-based economy, and the gains made through the NSF EPSCoR Track-1 projects are too precious to be lost for

lack of support. Arkansas has concentrated its priorities for state investments and has built-in methods to provide resources to encourage and maintain the gains from prior investments. DART will utilize a variety of sustainability strategies including leveraging of other existing programs, partnerships with private industry, user fees for facilities, and transitioning of operations to other agencies.

Sustainability of Project Activities:

The Arkansas S&T plan targets research areas where the state has both an academic research strength and an industrial and workforce need. Because the activities and infrastructure in this project are intentionally aligned with the S&T plan, DART will receive funding from future state investments from other state agencies, educational institutions, and organizations as described in the S&T plan. DART is expected to produce many valuable products and services with the potential to improve research capabilities in Arkansas, including improved data science infrastructure (hardware, software, cyber, human), methods, algorithms, scientific knowledge, partnerships, and intellectual property.

Arkansas has created a large number of state incentives and investment programs to support targeted research and the commercialization of its products. As part of the state's commitment to supporting data science in Arkansas the Governor, legislature, and our industrial partners established the Arkansas Center for Data Sciences (ACDS). With funding from the State and its private industry members, ACDS has a long-term mission of promoting and supporting data science research, education and job creation in Arkansas. ACDS and a number of state incentives targeting technology-based research and commercialization will be available to support DART activities that have proven to be effective. We anticipate that DART will become the research and development arm of ACDS, and plans are in motion to secure annual research funding for this goal.

Physical and Cyberinfrastructure Sustainability

As part of its long-term support for data science education, research, and industries, a goal of ACDS is to make statewide access to a shared infrastructure that includes HPC, storage, and data science tools more readily available. The infrastructure planned in DART will help connect not only our large research universities but many of our smaller universities and colleges that have very limited research activities. DART infrastructure and the ARP will be given priority for support by ACDS and the State in their future investments.

Human Infrastructure Sustainability

DART recognizes the importance of maintaining human infrastructure investments including retention of faculty and students, continuing their career development, and maintaining industry relationships. A number of strategies proposed by DART will be sustained directly by ACDS, State, and Federal funding. ACDS plans to continue support for DART student and faculty research fellowships, internships, and stipends for faculty training to expand the number of course offerings available in data science in Arkansas, as described in section 4.4. We also have plans to seek funding to support the Arkansas Summer Research Institute (ASRI) through training and education grants at NSF.

DART will hold regional and virtual workshops to enhance the research competitiveness of the state's faculty in data science and computing. Workshops topics will include science,

grantsmanship, and commercialization. These will be tailored for students and early career faculty and will feature program officers from various agencies, entrepreneurs, and outside scientists. The state also provides small grants to allow young technology-based start-up companies to pay for the cost associated with applying for SBIR funding.

Post RII Track-1 Extramural Funding:

Federal funding priorities are coalescing around data science. A number of Federal programs will be targeted by DART researchers for sustaining extramural activities. A few examples include the following NSF programs: Industry-University Cooperative Research Centers (IUCRC), the Science and Technology Centers (STC): Integrative Partnerships, the Computational and Data-Enabled Science and Engineering (CDS&E) initiatives, and the NSF Research Traineeship Program (NRT). NIH recently released a strategic plan for data science, with the goal to advance NIH data science across the extramural and intramural research communities. DART will investigate NIH programs such as Big Data to Knowledge, as well as interagency programs like the NSF-NIH Interagency Initiative: Core Techniques and Technologies for Advancing Big Data Science and Engineering (BIGDATA), as well as data-focused programs at other agencies like U.S. Department of Energy, U.S. DoD, and NASA.

Emerging Areas and Seed Funding

Over the course of the award period, we anticipate project researchers and advisors will identify emerging or transformative areas of research worthy of support or previously unidentified opportunities. To respond to these opportunities, the Management Team, in consultation with the External Advisory Board and the Industry Advisory Board, will allocate up to $318K in year 1, $380K in years 2, 3, and 4, and $300K in year 5 in seed grants to support emerging areas of research. A request for proposals will be posted online and distributed to all Arkansas institutions, with emphasis on collaboration. Proposals will be reviewed by a panel made of up of the EAB and IAB members. The duration of each project can be from 12 months to 24 months. Areas chosen for support will match the project's scientific focus and support one or more of the project goals. That is, they will be emerging opportunities that strengthen ties to Arkansas business; enhance our talent pool by expanding and/or leveraging research collaborations; or provide unique educational or training opportunities that could strengthen the currently planned programs and lead to new cross-disciplinary efforts not previously proposed. We will target exploratory research projects and unique, newly emerged areas of transdisciplinary education, diversity, or industrial outreach. Included in these funds for researchers and students for allocations on commercial clouds (Azure and Google). The management team, with technical counsel from the Coordinated CI team, will review allocation requests in years 2, 3, and 4 ($50K per year) for computational requirements that are best suited to the resources or burst access when ARP resources are limited.

Partnerships and Collaborations:

DART will utilize a variety of partnerships and collaborative activities to maintain synergy and relevance in all aspects of the project. Some partners have been identified and we will continue to pursue new partnerships and collaborations during the project.

Research Partnerships and Existing Collaborative Opportunities

DART will maintain an ongoing collaboration with the PiLog Group through its academic outreach organization, the PiLog Academy. PiLog funded two doctoral-level research assistantships (RAs) at UALR last year and is currently funding another RA position. In addition, they are providing the research team with unlimited access to a remote Hadoop Distributed File System (HDFS) Cloudera stack to support research in entity resolution and master data management.

The Sam. M. Walton College of Business (WCOB) at UAF maintains access to Consumer Panel Data from Nielsen at Kilts Center for Marketing and has established relationships allowing access to visits and sales data from Sam's Club and Dillard's. WCOB also provides academic access to the Acxiom Infobase Demographical Database including the PersonicX Clusters – Classic, LifeStage Groups, Insurance Groups, Financial Groups, and Digital—as well as population density, ethnic groups and other supporting data used to create the clusters. Researchers also have access to the All-Payer Claims Database (APCD) through the Arkansas Center for Health Improvement and funded by the Arkansas Biosciences Institute.

The J.B. Hunt Innovation Center of Excellence, established in May 2017, is an industry-funded research center at UAF driven through a collaboration between the College of Engineering, WCOB, and J.B. Hunt professionals. The center is funded through a 5-year; $2.75 million grant provided by J.B. Hunt. Research includes working with J.B. Hunt to overcome the challenges associated with developing methodologies that can integrate information of different types and sources to make improved strategic and operational decisions. Successful research projects are measured by their financial impact to J.B. Hunt and the degree in which they disrupt the transportation logistics industry.

Industry Partnerships and Collaborations:

The ACDS is a nonprofit public-private partnership organization that aims to develop, engage, and retain homegrown top talent in data analytics and computing. The board of ACDS consists of the Blue-Ribbon Committee members, including C-level executives from First Orion, Walmart, EZ Mart, Tyson Foods, Murphy USA, J.B. Hunt, Stephens Inc., Inuvo, AT&T, and others. DART will collaborate closely with ACDS and will receive feedback and advice from the board on numerous project aspects. Additional advisors will be recruited from companies like Acxiom, FIS, Bank OZK, and other large, medium and small businesses in the data sector.

0.8. Overall Risk Management Plan

Challenges and Risks

A project of this size and scope faces several challenges and risks including those that were knowable as the project was developed as well as additional risks and challenges arising from the current global situation. But Covid-19 creates unique and significant challenges to the project. Aside from the added friction of videoconferencing (friction that has lessened to some degree as we've learned from our experiences) is the toll taken on faculty and students in time and emotional state. Preparation for and delivery of both online and hybrid classes is far more time consuming and exhausting than all online or all face-to-face delivery. Mostly importantly, faculty working from home face a complex balance of family and work with children learning

from home and spouses working from home all sharing bandwidth and computer resources. Evidence is already mounting that women are more effected by work from home conditions. Schedules and work assignments of all faculty and staff supporting this project have changed since proposal develop – spring, summer, and fall 2020 course delivery moved to remote and hybrid learning, and many project personnel were instructed to work remotely effective March 2020. Out of state travel continues to be suspended from a halt date of March 2020. After an adjustment period, the team effectively shifted to remote working, including successfully executing regular meetings.

Coordinating cyberinfrastructure across seven research campus is primarily a policy and organizational challenge more than a technical challenge. In particular, this project must overcome several organizational issues.

- The lack of a federated identity agreement among participating institutions slows and complicates access to a federated system.
- The need to include campus IT departments who manage the campus enterprise and research networks requires more and larger meetings to agree on approaches that conform to disparate security and enterprise computing policies at each campus.
- Personnel changes at UALR require a shift in visualization strategies to better reflect the current expertise of the team.
- Covid-19 pandemic mitigation measures could challenge our ability to install new equipment at campuses on a timely basis and conduct productive training.

The research goals defined in the project will require the creation of research teams that have not collaborated before or teams led by excellent, but early career, principal investigators. This poses potential challenges regarding team management and efficient alignment of skills to challenges. While the approaches to development of methodologies dominate a majority of the research in the Learning and Prediction them are understood by participating researchers, data sets to test and validate new methodologies will need to be defined in collaboration with the other themed areas contributed to DART. In addition, this large commitment to research into such a wide variety of complex techniques will require a steady stream of student researchers with quantitative and computational skills. This increased need for graduate students with this expertise may require targeted recruitment over the course of the project.

We see these particular challenges and risk:

- Education themes do not have a strong history of close collaboration with the Research themes.
- Data Science capability and infrastructure vary widely within the collaborating institutions.
- Enabling technologies are rapidly advancing so participants will need to stay abreast of current trends.
- We are dependent on current industry to provide access to relevant data and other resources to support the DART program.
- Challenges of Academia and Industry working together on real-world data, real-world problems for class use and training students.

- The long timeframe for program approval for public programs at the state level (ADHE) and the various different agencies that accredit private institutions.
- Working with committees at different campuses on course and program approvals.
- The lack of human capital to teach courses in data science and computer science at some collaborating campuses.
- The lack of technology infrastructure, both hardware and software, to support the teaching of data science and computer science courses at some collaborating campuses
- Some collaborating institutions have been more involved than others on the ongoing work that has taken place to build the infrastructure in preparation for this project.
- Limited support of international undergraduate students (inability to compensate).
- The COVID-19 pandemic limits the opportunity for face-to-face meetings and has led to travel restrictions.
- Support for new and modified programs on collaborating campuses may be adversely affected by the COVID-19 pandemic.
- Participation in workshops and colloquia designed to broaden participation and to develop a common curricular framework and shared pedagogical strategies may be negatively impacted by the COVID-19 pandemic.
- Recruiting for Educational Activities and Programs may be impacted by the COVID-19 pandemic.
- Attracting Graduate Research Assistants to the program due to the COVID-19 pandemic.

Mitigation Plans:

To address these many challenges and risks we propose the following mitigation plans:

Campus IT coordination: The first step is to form joint working group that includes together IT leadership, relevant networking staff, and DevOps staff from UAF and UAMS. Regular monthly meetings, more frequent e-mail exchanges, and shared document repositories will facilitate smooth operation of the core of the CI infrastructure. UALR equivalents will be added to this working group in the second half of year 1 to complete the core team. This core team will develop template CI Plans to share with DART institutions who join the regularly scheduled or special meetings of the working group on an as-needed basis. Getting IT staff involved early in the process and involving as first-class members of the project has proven effective in solving both technical and administrative issue.

Development/integration effort of visualization into thrust areas: Personnel changes at UALR after submission of the EPSCoR DART proposal will require a redefinition of the visualization-support approach in the project. UALR will setup a post-doc position mainly responsible with project-related research and development. This position will be anchored in the EAC and supported/guided by four faculty in the computer-science department with experiences and expertise from a broad spectrum of backgrounds. The faculty together with the post-doc will analyze and discuss current/existing visualization approaches and strategies or the lack thereof with each of the thrust areas as well as with stakeholders involved in the shared testbed development. This will be used to start investigating how existing tools in the thrust areas are

already capable of fulfilling needs as well as starting custom developments. Our expectation is that patterns will emerge that should allow to define and create more integrated solutions, which are usable in multiple thrust areas as well as take advantage of the ARCC infrastructure work. The post-doc position will be further augmented by several graduate and undergraduate student positions. UALR is committed in pursuing additional instruments to provide the necessary talent and resources in this area.

Collaboration (esp. in the Covid-19 era): The DART research community needs simultaneous access to a system (Git, Globus, Jupyter notebooks, and others) as well as a shared and familiar video-conferencing platform. Several entities in the project, including AEDC, UAMS and UA, have enterprise licenses to Zoom which are available for DART. Many educational programs have moved on-line or to a hybrid model using Blackboard. ARCC will concentrate initial ARP service roll-out based on highest need across the research community. As noted above, Covid-19 mitigation plans have placed an unusually high load on faculty and, while online meetings and seminars reduce the amount of travel, they can also reduce the effectiveness of the meetings and limit sidebar conversations that tend to be very productive.

Education and Outreach (esp. in the Covid19-era): In past efforts, the education and workforce components were developed with little or no cooperation between those responsible for the separate components. In this project collaboration and cooperation was built in from the beginning. The expanding talent pool of highly trained data scientists that are expected to result from the program expansions supported by this project are critical to the ongoing expansion of data science efforts across the state.

The ACDS is also working on workforce and education across the state. The co-leads of the education effort have been working closely with ACDS Director Bill Yoder developing links between industry and education and links between institutions of higher education. Since the Track-1 proposal was submitted, this team has hosted numerous meetings related to the development of a cohesive statewide data science effort. This has given us a greater insight into the challenges we face and gives a better footing on which to build solutions. These efforts give us confidence that we can meet the challenge. As an example, when institutes of higher education (IHEs) talk about shared programs and courses, the participants always bring up obstacles and describe why such collaborations can never happen. As part of our effort to assure the naysayers, we have already discussed these issues with the ADHE, and we are confident that ADHE will be a contributing partner in our efforts to embed data science into the educational ecosystem of Arkansas. We have similarly engaged industrial partners and we will continue in these efforts. We do not believe that achieving our goals will be easy, but we understand the obstacles and we are committed to overcoming them.

The risks and challenges related to the current global situation were unexpected, and they could make progress more difficult. At this stage there are both known and unknown unknowns. The key factor is clearly the duration of the current global pandemic and whether or not an effective vaccine is developed. Anything that has ongoing negative impacts on IHE enrollments will impact our efforts. However, the emergence of data science provides new opportunities and can provide IHEs expansion possibilities in an area where the global pandemic is not expected to have a negative impact on opportunity. It is clear that an ongoing pandemic will reduce the number of international students.

Communication and Dissemination: Maintaining the campus communication team could be a challenge, it has not been done previously and campuses tend to want to publish research results and new award information as soon as possible. In the past it has been difficult to make sure the grant is cited properly. Our plan to mitigate this is to maintain frequent communication with the comms team and ensure quick turnaround of any needed quotes and citation materials.

1. Coordinated Cyber Infrastructure

The proposed research will be supported by a data science cyberinfrastructure (CI) platform capable of providing secure, distributed, agile, scalable, and on-demand services. We propose to architect and build a shared high performance computing environment, the Arkansas Research Platform (ARP) (Figure 5) and integrate it with existing high-performance computing and petabyte scale storage resources. In combination, these will provide 1) libraries of pre-configured containers designed to support a variety of well-known and novel workflows in machine and statistical learning, graph theory, bioinformatics, and geoinformatics, 2) containers configured for parallel computation and distributed memory on HPC resources for analysis of very large datasets, 3) the ability for researchers to create and share new containers and share, and 4) explore requirements necessary to stream data to visualization environments both proximate and distant from the computing resources to aid in analysis and meta-analysis of experiments.

1.1. Advancing the State of the Knowledge

A special CI advisory board, chaired by James Deaton, executive director of the Great Plains Network, and composed of the CoPIs of the NSF funded CyberTeam award #1925681, has been formed and will advise in refinement and management of the Arkansas Research Platform described below. This special advisory board will be useful not only in providing external experience in building data science computing platforms, but in coordinating the connection of ARP to the Great Plains Network Research Platform, the Great Plains Augmented Regional Gateway to the Open Science Grid, and on to nationally organized compute and storage resources, complimenting existing connections through XSEDE

A unifying function of the CI is support for the development, optimization and management of analysis pipelines from each of the research themes. Our preliminary experience with this approach has been quite positive with existing containerized pipelines for image curation, genomics analysis, and machine learning.

1.2. Project Implementation

Goal	Goal Name	Lead(s)	Team Members
CI1	Establish the Arkansas Research Platform as a shared data science resource across the jurisdiction	Cothren, Prior	Chaffin, Springer, Tarbox
CI2	Visualization for complex data in diverse data-analytics application domains	Springer	Cothren, Prior

In May of 2020, UAF and the UAMS entered a partnership to consolidate the management of each campus' high-performance computing centers into an entity called the Arkansas Research Computing Collaborative (ARCC). ARCC will manage ARP and all activities related to it. ARCC will be implemented in the first year of the grant and expanded to include resources available at the Emerging Analytics Center at the UALR. These three institutions will act as resource providers and consumers while the other institutions in DART will consume these resources with direct access to big data through Globus, code sharing through a dedicated DART GitHub, and computing through interactive and batch sessions on the ARP HPC and private cloud computing resources.

The platform will be configured and managed in accordance with lessons learned from existing similar effort like the Pacific Research Platform and, in particular, the Great Plains Research Platform; and with input from an advisory board chaired by the executive director of the Great Plains Network, James Deaton. We will use best practices learned from these projects, with the goal to enable sharing resources among topic-integrated, but geographically diverse, research teams. Remote storage and processing of larger compute jobs will be hosted at XSEDE national computing sites and commercial providers (e.g., Google, Azure). Connectivity across the state will be provided by ARE-ON and serve the seven research campuses and two education-focused HBCU campuses, Philander Smith College and Shorter College.

Authorization and authentication to ARP will be a challenging problem. Authorization to Pinnacle and Grace and their connected storage arrays will be managed by existing HPC staff at the respective campuses. They will maintain authorized user accounts which are not shared across the systems. However, authentication will be managed through two channels. First, the University of Arkansas System (UAS) will install use Cirrus Bridge to connect to the UAS Azure tenant, giving access to all UAS institutions (including UAF, UAMS, UALR, UAPC, and seven 2-year colleges). This is expected to authenticate most researchers across the state. Non-UAS institutions (ASU, SAU, UCA) will be added individually to the UAS Azure tenant. UAF will also operate a separate Azure tenant. UAMS will also enable the Grace portal to work with a Keycloak SAML configuration.

ARP will use Globus Data services for big data sharing across the seven research campuses. Endpoints will be set up at main storage sites proximate to the clusters at UAF and UAMS, and to the visualization resources at UALR. ASU, UCA, SAU, and UAPB may also have endpoints but researchers at these institutions will more likely use endpoints at their workstations.. Data will be backed up to the NSF-funded regional tape resource OURRstore at the University Oklahoma. ARP will support a variety of research computing platforms: traditional bare-metal HPC jobs, Singularity containers, and kernel virtual machines (KVMs) on the existing Pinnacle and Grace clusters as well as the new data science cluster funded as part of this proposal. The Sample Linux Utility for Resource Management (SLURM) scheduler will be used to provision all three types of jobs. Singularity containers are a widely accepted, secure standard in a multi-user HPC environment where access to the data of other users must be restricted. The container jobs will require a user to either download a container from external repositories like NVidia NGC (in native Singularity format) or Docker Hub (easily convertible to Singularity format) or use containers stored on shared local storage on Pinnacle or Grace. Once the container is in place, a single-line command in the SLURM job script will bind the user's input data directory

to the container and run the executable inside the container on the input data. The job terminates when either the executable in the container finishes processing the input, or the job exceeds the requested wall time. A variety of big data management resources such as HDFS and Apache Spark will be enabled using the MapR Sandbox and internet-facing graphical interfaces will be provided to run services such as Jupyter Notebook, RStudio Server, and the HPC scheduler, Open OnDemand, at both UAF and UAMS.

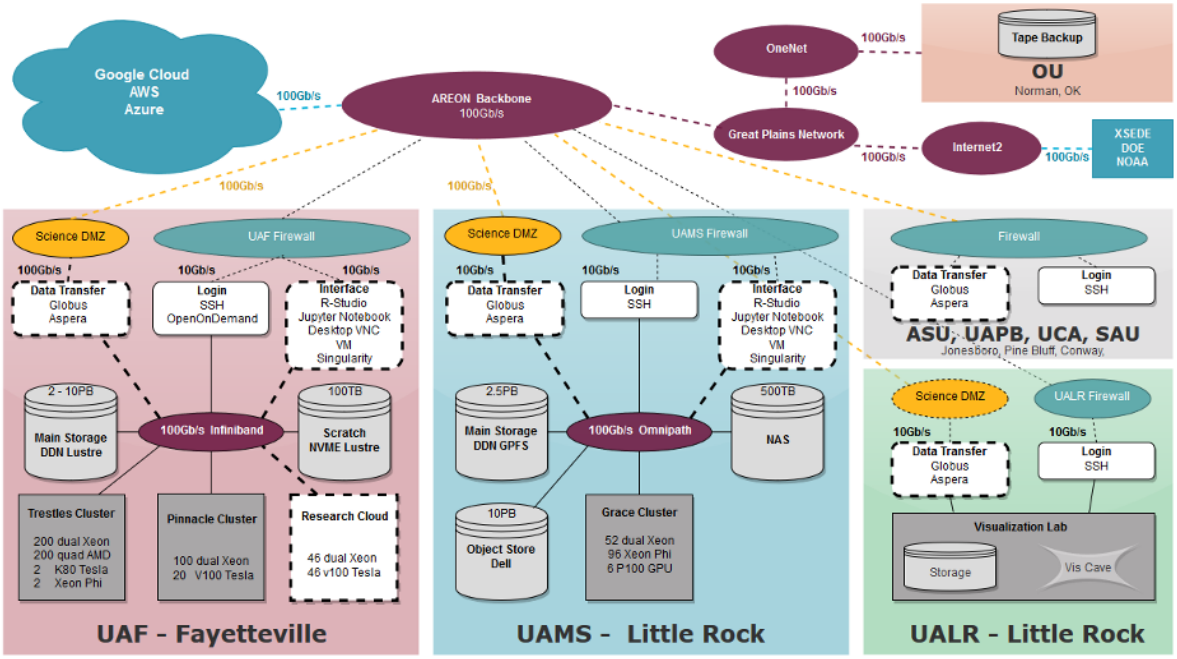

Figure 1: The computational backbone of ARP consisting of existing and new, grant-supported equipment.

To support the proposed research, a dedicated data science-oriented cluster will be purchased and physically joined to Pinnacle at UAF. It will consist of approximately 46 nodes each with dual-Xeon Cascade Lade 20-core processors, 768 GB of memory, 480 GB local solid-state storage, one Nvidia Tesla V100 GPU. Nodes will be connected via EDR or HDR InfiniBand. Additional storage (500TB) will be added at UAMS and 48 nodes from the existing Grace cluster re-tasked to contribute to ARP. To be investigated is the dynamic allocation of resources between ARP and traditional local HPC based on demand. Currently, only UAF has 100Gb/s capability to connect to the ARE-ON backbone. Proposed funds will add 100Gb/s capability to the UAMS computing center in year 2, finally realizing a high-bandwidth connection between the two major computing clusters in the state via links provided by the statewide ARE-ON backbone.

ARP will be connected with the Great Plains Network Research Platform to leverage regional resources and to potentially increase the compute infrastructure available to DART and other researchers in Arkansas. ARP plans to connect Arkansas-based resources to the Open Science Grid (OSG) using gateway nodes, leveraging work being done for the CC* Compute award #2018766: GP-ARGO: The Great Plains Augmented Regional Gateway to the Open Science Grid. The connection will give DART and other researchers in Arkansas access to the vast resources for High Throughput Computing available on the Open Science Grid, while

providing resources to the Open Science Grid that are not being used for local projects. ARP will partner with The Carpentries to deliver high quality data science-oriented training in scientific software development, data management, and code management. UA is currently a Silver Member and will seek to offer at least 5 online training sessions per year in various locations.

Summary of changes in 2023 strategic plan revision

Milestones related to the advisory board expansion under 1.1.a were removed. Additionally, since the CI team made the decision to utilize a GitHub organization repository instead of GitLab, references to GitLab have been updated. Similar changes were incorporated to reflect flexibility around Hadoop and containerization efforts under 1.1.b. More significant revisions were made to the activities and milestones under Goal 1.2, due to personnel turnover, the inability to recruit and hire a qualified postdoctoral candidate, and a number of administrative challenges. Many of these issues have now been resolved during Year 3, but since the project has entered Year 4 revisions were made to use the rest of the project time as productively as possible.

Goal 1.1 (CI1)	Establish the Arkansas Research Platform as a shared data science resource across the jurisdiction
Objective 1.1.a: Establish the Arkansas Research Computing Collaborative (ARCC) **Objective 1.1.b:** Upgrade cluster for data science research activity and integrate with existing resources **Objective 1.1.c:** Establish a science DMZ in Little Rock (UAMS, UALR) and high-speed connection with UAMS **Objective 1.1.d:** Establish a data and code sharing environment (GitHub and Globus) **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)	
Goal 1.1 Output Metrics **Hardware and software infrastructure improvements (5):** -- Install, configure, and make available data science nodes on Pinnacle Portal -- ScienceDMZ at UAF and UAMS/UALR -- 100Gb connection between ScienceDMZ -- Establish a dedicated DART GitLab repository -- Setup Globus data management services to point at DART storage arrays **Documentation and user guides (4):** -- Create one (1) ARCC technical management document -- Amend existing MOU for ARCC expansion -- Create two (2) CI Plans (1 x UAF, 1 x UAMS) -- Create one (1) GitLab user guidelines reference document **Workshops, demonstrations, and trainings (38):** -- Host two (2) online workshops per year for onboarding to ARP resources in YR2-5 (8 total) -- Host five (5) online software carpentry workshops per year in YR2-5 (20 total) -- Train and certify two (2) new software carpentry instructors per year in YR2-5 (10 total) **Applications and platforms (5):** -- Create one (1) distributed computing testbed for HDFS, Apache Spark, others (DC) --Create four (4) spatiotemporal testbeds for (CI/DC/SM/LP)	

Objective 1.1.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Create ARCC advisory board with regional partners (GPN)		Advisor board is formed with establishes roles and responsibilities consistent with MOU			
Activity 2: Establish ARCC governance, operations, and staff between UA and UAMS		Document defining organizational structure, roles, and responsibilities of ARCC			
Objective 1.1.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Specify and purchase data science cluster based on document from 1.1.a	Issue UA purchase order for additional equipment	Receive data science nodes for Pinnacle (anticipated)			
Activity 2: Test and deploy hardware elements for Pinnacle expansion for DART		Install, configure, and make available data science nodes on Pinnacle			
Activity 3: Install and configure data science cluster to work with existing resources at UA, UAMS, UALR resources		Collect testbed specifications and software/platform needs	Create containerized Hadoop-based testbed for DC	Create testbeds for SM	
Objective 1.1.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Specify and purchase 100Gb switch		Issue UAMS purchase order for 100 Gb switch	Receive 100 Gb switch (anticipated)		

Activity 2: Install 100Gb switch			Install and configure new 100 Gb switch		
Activity 3: Establish ScienceDMZ at UAMS	Create UAMS CI Plan	Specify and acquire additional DMZ components	Establish and validate 100Gb link to UAF and integrated DMZ		
Objective 1.1.d	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Create/identify federated identify or other authentication mechanism for all sites that provides access to core ARP resources			Establish federated access for all project participants		
Activity 2: Engage other research themes to develop research-specific training modules in e.g. Python, R, Git, HPC, Singularity		-- Host 5 online software carpentry workshops -- Train 2 software carpentry instructors	-- Host 5 online software carpentry workshops -- Train 2 software carpentry instructors	-- Host 5 online software carpentry workshops -- Train 2 software carpentry instructors	-- Host 5 online software carpentry workshops -- Train 2 software carpentry instructors
Activity 3: Develop and deploy training materials for code sharing, large data transfer protocols		Host 2 online ARP-specific training sessions	Host 2 online ARP-specific training sessions	Host 2 online ARP-specific training sessions	Host 2 online ARP-specific training sessions
Objective 1.1.e	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Identify the number and type (HIPAA, proprietary economic, CUI, etc.) of private and secure data sources that will be need to be accessed by DART researchers.	Collect research theme needs				
Activity 2: Setup capacity for storing and managing CUI and HIPAA data at UAF		Deploy restricted access storage	Draft user guidelines and policy		

Goal 1.2 (CI2)	**Visualization for complex data in diverse data-analytics application domains**				
Objective 1.2.a: Define domain-specific integration of visualization solutions					
Goal 1.2 Output Metrics **Publications, presentations, and reports (3):** -- Three (3) presentations, reports, or other publications: 1 in YR1 and 2 in YR2 **Workshops, demonstrations, and trainings (4):** -- One (1) online workshops per year for advanced visualization in YR2-5 (4 total) **Applications and platforms (2):** -- Two (2) augmented reality or virtual reality based visualization applications					
Objective 1.2.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Develop and deploy visualization infrastructure software				VR neuroimaging and 3D arterial applications developed	Applications disseminated and training conducted
Activity 2: Engage other research themes to develop research-specific advanced visualization training				Host 1 online advanced visualization workshops	Host 1 online advanced visualization workshops

2. Data Life Cycle and Curation

The overall goal of the Data Life Cycle and Curation Theme is to create unsupervised and scalable methods that significantly increase the level of automation in the data curation process from acquisition to disposal. The primary curation processes targeted for automation in DART research are data quality management, data integration, and data governance. While there are many tools already available for these processes, they are all dependent upon human supervision to work effectively. One of the most common complaints of data scientists is they only spend 20% of their time on modeling and problem solving and the other 80% on data preparation. The same is true in industry where most of the effort in data operations is consumed in cleansing, standardizing, and integrating data prior to its actual application in information products. As data volume continues to grow at a rapid rate, the curation process has become a significant bottleneck for data operations resulting in long delays before data are available for analytics and other data-driven operations.

The Arkansas Science and Technology Plan which has three primary objectives.

1. To identify opportunities for academic and industrial collaboration
2. To align future investments in university research competencies with industry areas of technology focus
3. To stimulate improvement in technology skills and talent development.

Because the lack of automation in data curation is a problem for both industry and academic research, the theme of automating data curation fits well with the first and second objectives. Collaborating with industry through testing real-world datasets will be an essential component of the research. As the first university research to focus on this research theme, the researchers, including student research assistants at participating schools, will develop high-level skills in data analytics, data governance, and machine learning.

Within Arkansas, the UAMS, UAF, and UALR campuses will be the primary drivers for the research. UAMS will bring into the research the special needs of data curation for biomedical informatics, UAF expertise in AI and machine learning, and UALR the industry perspective.

In addition to the internal partners, the preliminary research has already attracted interest from a number of external academic and industry collaborators including the MIT Chief Data Officer and Information Quality program, PiLog Group, and Noetic Partners. As the research matures it is likely to attract more collaborators and potentially result in the development new open source and commercial products and generate new business opportunities.

2.1. Advancing the State of the Knowledge:

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.

If such machines can be built, the benefits are enormous. They will revolutionize data operations in research, industry, and government. When data cleaning, data integration, and data governance become unsupervised, automated processes, then much more data can be

ingested and analyzed, and greater advances in data analytics can be made is less time. At the same time, the improvements in data governance will make enterprise data assets more secure while making them more available and discoverable for authorized users.

2.2. Project Implementation

Goal	Goal Name	Lead(s)	Team Members
DC1	Automate heterogeneous data curation	Talburt	Talburt, Cothren, Liao, Rainwater, Tudoreanu, Ussery, Wang, Xu, Yang
DC2	Explore secure and private distributed data management	Talburt	Talburt, Wang, Tudoreanu, Pierce, Liu, Rainwater
DC3	Harmonize multi-organizational and siloed data	Ussery	Ussery, Byrum, Jun, Yang, Rainwater

The initial research will focus on three important use cases from industry and academia. The first use case is "multiple sources of the same information." Solving this use case is Objective 2.1 of the Activity Matrix. In *Journey to Data Quality* (Lee, *et al.*, MIT Press, 2006), having multiple sources is at the top of the list of the 10 most common root causes of data quality problems. It is a process repeated over and over every day in organizations across the globe. They receive large files of information about customers, patients, products, events, and other entities from multiple sources in different formats. The current approach is for a data analyst to profile each source and design and ETL process to transform it into a standard layout. After each source is prepared the records then go through a linking process designed by a data integration analyst so that records describing the same entity can be identified, linked together, and eventually integrated into an information product. The goal of this use case is to create an unsupervised process to produce equivalent results. There are four concurrent streams of work within this first use case (DC Goal 1). These are:

- Automate Data Quality Assessment, Years 1 - 5
- Automate Data Cleansing, Years 1 - 5
- Automate Data Integration, Years 1 - 5
- Implement Data Cleansing and Data Integration Models in HDFS, Years 2 - 5

The second use case is around the design of an enterprise positive data control (PDC) system. As more and more organizations recognize data and information as a key asset, they are adopting the discipline of data governance. One of the key objectives of data governance (DG) is to always know exactly what data is present in the system, exactly where it is located, and what it represents. Just as with data curation, DG requires a great deal effort by a large number of persons in the role of "data stewards" making it difficult to keep up with the growing data volume and variety. At the same time, an organization must comply with a growing number of new data regulations such at the General Data Protection Regulation (GDPR) from the European Union, and the California Consumer Protection Act (CCPA). Both regulations allow a consumer the "right to be forgotten," *i.e.* having their personal identifying information (PII) deleted from all systems. Without effective DG, organizations are at risk of non-compliance carrying substantial fines and penalties. The goal of this use case is to create a controlled data

operations environment, in which all data access is controlled and every data action is automatically tracked and recorded. There are three concurrent streams of work within this second use case (DC Goal 2). These are:

1. Build a POC for Positive Data Control (PDC), Years 1 - 3
2. Extend the Functionality of the POC, Years 3 - 4
3. Implement Data Portability and Exchange Functionality

Figure 2: data washing machine

The third use case focuses on the curation of genomic and proteomic data, which is now growing at the incredible rate of petabytes of new data per week. The goal of this use case is to design scaffolding systems that capture and disseminate both data and processing steps, and that can function as a democratizing tool for data in a similar manner as blockchain functions for currency. This effort intends to allow sharing of data processing steps that can be beneficial to (a) valid integration of data from multiple silos, (b) multi-enterprise, distributed data governance and curation efforts, (c) supporting an organization that has multiple or lax data governance processes, and (d) upfront automated data curation by machine learning algorithms. There are three concurrent streams of work within this second use case (DC Goal 3). These are:

- Standardize pipelines for genome and proteome storage, retrieval, and visualization, Years 1 - 2
- Automate quality scores for biological sequence data, Years 2 - 3
- Apply machine learning methods to systems biology, Years 3 – 5

Summary of changes in 2023 strategic plan revision

Milestones under Objective 2.1.a for Years 3 and 4 were moved to Years 4 and 5 to allow some flexibility for the team to complete the tasks considering minor delays due to personnel loss and student hiring. Activity 3 under the same objective was revised to incorporate EAB suggestions and needs of other DART participants. Revisions were also made to activities and milestones under Objective 2.2 to incorporate EAB and RSV feedback.

Goal 2.1 (DC1)	Automate heterogeneous data curation
Objective 2.1.a: Automate Reference Clustering / Automate Data Quality Assessment **Objective 2.1.b:** Automate Data Cleansing **Objective 2.1.c:** Automate Data Integration	
Goal 2.1 Output Metrics **Publications, presentations, and reports (46):** -- 14 research publications describing new methods and processes -- 11 journal and conference publications -- 16 presentations -- Five (5) PhD dissertations **Patents and start-ups (2):** -- Two (2) potential patents and/or business incubation **Workshops, demonstrations, and trainings (5):** -- Five (5) conference workshops **Datasets and algorithms (18):** -- Nine (9) Mathematical formulations and/or algorithms for optimal collaborative data collection and cleansing. -- Seven (7) Mathematical formulations, algorithms, and/or statistical methods for need- and prediction-based data collection, and scalable decision making	

Objective 2.1.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Define metrics for data quality to measure impact of unsupervised data cleansing on data standardization and reference clustering	Define at least one metric for completeness, standardization, and clustering quality of unstandardized reference data; Design and implement an unsupervised algorithm for each metric	Design and implement an algorithm using ML or Graph techniques for **one** metric		Design and implement a scalable algorithm in HDFS for **one** metric	Design and implement a scalable algorithm in HDFS for **all** metrics

Activity 2: Set baseline data quality for initial test datasets used in prior research and acquire additional test datasets	-- Establish baseline quality using supervised methods for existing datasets -- Compare results of unsupervised quality metrics developed in Activity 1 to supervised results	-- Add 5 new person and 5 new business reference datasets for testing, at least 2 real-world -- Add 5 new product reference datasets		Add 3 new person and 2 new business reference datasets with more than 1 million records for testing HDFS code	Add 5 new product reference datasets with more than 1 million records for testing HDFS code
Activity 3: Curate test datasets and generate synthetic data for other researchers	Establish a repository for the reference datasets and make available to other researchers			Investigation of synthetic occupancy generator, ideal parameters explored	Development of 2 synthetic datasets
Activity 4: Develop a framework for collaborative data collection and cleansing for knowledge discovery	Formulate a hierarchical and as-needed data collection and cleansing strategy	-- Refine the formulation by including various practical constraints and test on small-scale problems -- Formulate a collaborative data collection strategy involving multiple teams		Solve large-scale problems by considering a tree-based tool for data clustering and cleansing	Solve large-scale problems considering hierarchical and unsynchronized data collection, and test on inland waterway datasets
Activity 5: Develop a need- and prediction-based feedback mechanism for future data collection and making scalable decisions	-- Formulate a framework for sequential data collection on an as-needed basis -- Refine the formulation by including various practical constraints and test on small-scale problems	-- Formulate a Bayesian framework for sequential data collection based on predictive models -- Investigate analytical approaches for using large datasets for different levels of decision making	Refine the Bayesian framework for sequential data collection based on predictive models and medium-sized data	Study an optimum data guided approach for sequential learning and dynamic data collection	Solve large-scale sequential data analysis and multi-level decision making problems and test on inland waterway datasets

Objective 2.1.b	Specific Milestones				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Improve the unsupervised frequency-based data cleansing method used in prior POC; Explore and test alternative methods and models for unsupervised data cleansing including ML, AI, and graph approaches	-- Document and train team on data cleansing methods developed in prior research -- Design and implement in Python or Java improvements to the prior frequency-based approach	-- Design and test an ML or Graph implementation to the prior frequency-based approach -- Design and test new techniques for unsupervised data cleansing in Python, Java, ML, or Graph	Continue to design and test new techniques for unsupervised data cleansing in Python, Java, ML, or Graph		
Activity 2: Migrate successful data cleansing models into scalable processes		-- Refactor and migrate prior frequency-based approach into a scalable process	Refactor and migrate most successful of new data cleansing techniques into scalable processes		

Objective 2.1.c	Specific Milestones				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Improve the unsupervised frequency-based data integration method used in prior POC and explore and test alternative methods and models for unsupervised data integration including ML, AI, and graph approaches	-- Document and train team on reference clustering method developed in prior research -- Design and implement in Python or Java improvements to the prior frequency-based approach	-- Design and test an ML or Graph implementation to the prior frequency-based approach -- Design and test new techniques for unsupervised data cleansing in Python, Java, ML, or Graph	Continue to design and test new techniques for unsupervised data cleansing in Python, Java, ML, or Graph		
Activity 2: Migrate successful reference clustering models into a scalable HDFS processes		Refactor and migrate **prior** frequency-based approach into a scalable HDFS process	Refactor and migrate **most successful** of new reference clustering techniques into a scalable HDFS processes		

Goal 2.2 (DC2)	**Explore secure and private distributed data management**
Objective 2.2: Build a POC and demo for Positive Data Control (PDC)	
Goal 2.2 Output Metrics **Publications, presentations, and reports (7):** -- Three (3) conference presentations -- Three (3) research papers -- One (1) dissertation	

Workshops, demonstrations, and trainings (5):

-- Five (5) workshops on deep learning and natural language processing

Applications and platforms (2):

-- One (1) working prototype of a positive data control system

-- One (1) novel design of a system with multiple prototypes and simulated data governance rules

Objective 2.2.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 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		-- Setup a test platform with at least one processing function (e.g Hive), metadata function (e.g Atlas), and security function (e.g Ranger); -- Build POC with a simple PDC layer where Hive user is forced to go through PDC layer for all read/write operations	Modify POC to synchronize Hive operations with metadata layer (Atlas) and security permissions (Ranger)	Add APIs to PDC layer to control an additional data transform and one data movement tools	Create and give demo of typical use case where PDC controls all user actions
Activity 2: Implement Data Catalog and Data Governance for DART				Build catalog of DC team datasets and also document policy and procedures for updates	Share model & process to all groups catalog of DC team datasets and also document policy and procedures for updates
Activity 3: Collaborate with Learning & Prediction to improve automated Data Curation				Use ML to optimize parameters of data washing machine.	
Goal 2.3 (D3)	**Harmonize multi-organizational and siloed data**				

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

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

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

Goal 2.3 Output Metrics

Publications, presentations, and reports (17):

-- Five (5) papers published on standardized pipelines for genomics and proteomics.

-- Five (5) conference presentations

-- Five (5) Journal publications

-- Two (2) PhD dissertations

Workshops, demonstrations, and trainings (1):

-- Host one (1) workshop on how to use these standardized pipelines

Datasets and algorithms (2):

-- At least one (1) standardized database with genomics and proteomics quality scores shared with DART researchers

-- One (1) algorithm (code with associated training and testing data)

Objective 2.3.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Define and download datasets to be curated	Build genomics database, including quality scores, and gene/ protein annotation	Extend to proteomics database - all for fast characterization of proteins (links to SwissProt)	Update databases (every 6 months)		
Activity 2: Optimize data storage and retrieval	Use Elastic Cloud Storage for fast retrieval	Develop integrated database for proteomics & genomics, including annotations	Update databases (every 6 months)		
Activity 3: Develop visualization methods	Prototype of R-BioTools for visualizing genomes	Publish one (1) R-BioTools paper for visualizing genomes	Develop visualization methods for very large trees	Visualization of phylogeny for Ark. Pathogens	

Objective 2.3.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Develop pan-genome and Pan-proteome databases	Develop architecture / structure for rapid storage/retrieval of taxa-specific pan- and core-genomes		Develop rapid storage/retrieval for core- and pan-proteomes		
Activity 2: Develop taxonomy links to downloaded genomes/proteomes	Compare duplicate, known type strain genomes using ANI, Mash, 16S rRNA	Use Mash and other methods to assign nearest neighbors in phylogenetic space.	Update taxonomy (every 3 to 6 months)		
Activity 3: Develop a genomic database for Arkansas genomic pathogen surveillance of antimicrobial resistance			Build Arkansas pathogen database, with links to known pathogens; develop GPU-based methods for fast calculation of genomic distances	Finish development of GPU-based methods for fast calculation of genomic distances	Publish one (1) paper

Objective 2.3.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Define training sets to be used for ML	Identify key datasets and problems for ML	Develop ML models for known toxins	Develop ML models for antibiotic resistance	Apply ML models to Ark. Pathogens	
Activity 2: Integrate multi-omic models for ML	Integrate genomic / microbiome / taxonomy datasets (petabytes)	Integrate genomic, transcriptomic, proteomic, and metabolomic datasets (petabytes)	Integrate model organisms (mouse, rat, human, yeast, etc.) as well as microbial	Publish one (1) Integration paper	

3. Social Awareness

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 large amounts of many different kinds of 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. As such, social awareness should be an integral part of research and training in the area of data analytics. In this theme, we are focused on developing cutting-edge socially aware data analytics to address social concerns and meet laws and regulations in national-priority applications, thus better enabling big data analytics to promote social good and prevent social harm.

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

3.1. Advancing the State of the Knowledge:

Our developed technology can achieve meaningful and rigorous privacy protection when mining private data or collecting sensitive data from individuals; ensure non-discrimination, due process, and understandability in decision-making; achieve safe adoption, and robustness of machine learning and big data analytics techniques, especially in adversarial settings; and help incorporate social awareness in domain- or application-specific projects. Our research projects in this theme will advance the state of the knowledge in the following perspectives.

SA1: Privacy-preserving and attack resilient deep learning

- Vulnerabilities of deep learning algorithms under existing and new attacking models
- A universal threat- and privacy-aware deep learning framework to achieve meaningful and rigorous differential privacy protection and resilience against various attacks
- Better understanding of the applicability of threat- and privacy-aware deep learning models in real world applications

SA2: Socially aware crowdsourcing

- Improve crowdsourcing data quality with considerations of uncertainty
- Enhance available inference and learning models with novel algorithms for improved effectiveness and efficiency
- Verify and validate the robustness and trustworthiness of information from crowdsourcing data

SA3: User-centric data sharing in cyberspaces

- Understand personal identifying information and their privacy issues
- New appropriate multimodal deep learning techniques to identify discriminative and stigmatizing information
- A user-centric privacy monitoring and protection framework

SA4: Deep learning for preventing cross-media discrimination

- Deep learning-based techniques to detect cross-media discrimination
- New generative adversarial models to remove cross-media discrimination
- 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

SA7: Cryptography-assisted secure and privacy-preserving learning

- Develop cryptography-aware privacy preserving machine learning methods, and develop privacy protection for classification input data in machine learning applications

3.2. Project Implementation

Goal	Goal Name	Lead(s)	Team Members
SA1	Privacy-Preserving and Attack Resilient Deep Learning	X. Wu	Q. Li, Zajiciek
SA2	Socially Aware Crowdsourcing	Hu	N. Wu, X. Wu
SA3	User-centric Data Sharing in Cyberspaces	N. Wu	Q. Li, Hu
SA4	Deep Learning for Preventing Cross-Media Discrimination	L. Zhang	X. Wu, Zajicek
SA7	Cryptography-Assisted Secure and Privacy-Preserving Learning	Q. Li	Zajicek, N. Wu, Luu
SA8	Causality-based Fairness in Social Networks	L. Zhang	Gauch

The research goals of the Social Awareness research theme are to develop novel techniques to provide privacy preservation, fairness, safety, and robustness to a variety of data analytics and learning algorithms including automated data curation, social media and network analysis, and deep learning, and ensure the adoption of the developed techniques meet regulations, laws and user expectations. The Social Awareness Activity Matrix presents details of objectives, activities and planned milestones over the five-year period for each of the seven projects.

Summary of changes in 2023 strategic plan revision

Goal 3.5, Marketing Strategy Design with Fairness, was removed due to the loss of personnel Zenghui Sha. The team could not find a replacement with similar expertise. This also impacted some milestones and activities planned under 3.4, Deep Learning for Preventing Hate Speech. As a result, objective 3.4 has been refined and milestones have been consolidated. Due to the loss of Xiuzhen Huang, the planned activities and objectives for SA6 and SA7 have also been revised. Milestones were consolidated, and some new activities were incorporated to address EAB feedback regarding collaboration between the SA and SM teams.

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

Objective 3.1.a: Identify potential vulnerabilities of deep learning algorithms
Objective 3.1.b: Develop a universal threat- and privacy-aware deep learning framework
Objective 3.1.c: Conduct comprehensive evaluations of the proposed framework and models

Goal 3.1 Output Metrics

Publications, presentations, and reports (5):

-- Two (2) conference papers
-- One (1) journal paper
-- One (1) thesis
-- One (1) proposal

Workshops, demonstrations, and trainings (1):

-- One (1) tutorial given at major AI conference

Objective 3.1.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Research existing attacks including model inversion attacks and data poisoning attacks and capture mechanisms behind the threat models	Document literature research of attack models and mechanisms behind attacks				
Activity 2: Study the potential risks due to correlations among input data features, parameters, output, target victims, and latent feature space in deep learning algorithms	Initiate theoretical investigation on the risks of deep learning algorithms	Disseminate the findings of both theoretical and empirical studies on risks of deep learning algorithms			
Activity 3: Study the sensitivity and impact of input data features, parameters, and the objective functions on the model output and identify appropriate differential privacy preserving mechanisms for different computational components in a variety of deep learning models	Initiate theoretical investigation of privacy preserving mechanisms.	Disseminate the findings of both theoretical and empirical studies on privacy preserving mechanisms used for deep learning algorithms			

Objective 3.1.b	Specific Milestones				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Investigate the tradeoff of achieving privacy, resilience to adversarial attacks, and utility		Research the tradeoff of privacy, resilience, and utility	Complete both theoretical and empirical studies on privacy, resilience, utility tradeoff in the deep learning setting		
Activity 2: Study the mechanisms of redistributing injected noise across input data features, model parameters, and coefficients of objective functions based on their vulnerability and impact on the model output		Examine the noise redistribution mechanism	Complete both theoretical and empirical studies on the noise redistribution mechanism in the deep learning setting		
Activity 3: Develop and implement threat- and privacy-aware deep learning models		Design algorithms of threat- and privacy-aware deep learning models	Complete initial implementation	Complete the model improvement and extension	

Objective 3.1.c	Specific Milestones				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Evaluate the developed framework and models against baselines using benchmark datasets				Complete the data collection of benchmark datasets and implementation of baselines	Complete empirical evaluation of developed models against baselines on benchmark datasets
Activity 2: Evaluation and validation with participating companies				Complete the data collection (including preprocessing) of real-world datasets	Complete empirical evaluation of developed models against baselines on collected real world datasets

Goal 3.2 (SA2) | Socially Aware Crowdsourcing

Objective 3.2.a: Improve crowdsourcing data quality with considerations of uncertainty
Objective 3.2.b: Enhance available inference and learning models with novel algorithms for improved effectiveness and efficiency
Objective 3.2.c: Verify and validate the robustness and trustworthiness of information from crowdsourcing data

Goal 3.2 Output Metrics

Publications, presentations, and reports (5):

-- Two (2) conference papers

-- One (1) journal paper -- One (1) thesis -- One (1) proposal					
Objective 3.2.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Allow uncertain labels in crowdsourcing data collection	Selected the approaches through literature review	Implemented and tested			
Activity 2: Aggregate raw labels after label collection	Computational schemes are identified	Implemented and tested			
Activity 3: Filter out possible noises to further improve data quality	Identified possible sources of noises	Filtering algorithms designed	Algorithms implemented and tested	Publish obtained results, and investigate the impacts of human factors	
Objective 3.2.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Build theoretic foundations	Specified mathematical requirements				
Activity 2: Develop learning models and inference algorithms		Algorithms designed to meet specification	Algorithms implemented and tested		
Activity 3: Test and apply these learning models and algorithms		Testing dataset selected	Initial tests completed	Refined learning models and algorithms	
Objective 3.2.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Establish additional evaluation metrics			Quality metrics established		
Activity 2: Develop algorithms to calculate the metrics			Quality metrics are quantified	Algorithms implemented and tested	Disseminate the findings of the algorithms and refinements

Activity 3: Verify and validate computational results				Completed system integration and testing	Performance evaluated and compared

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

Objective 3.3.a: Investigate on personal identifying information and their privacy issues
Objective 3.3.b: Investigate appropriate multimodal deep learning techniques to identify discriminative and stigmatizing information
Objective 3.3.c: Develop a user-centric privacy monitoring and protection framework

Goal 3.3 Output Metrics

Publications, presentations, and reports (5):

-- Two (2) conference papers
-- One (1) journal paper
-- One (1) thesis
-- One (1) proposal

Objective 3.3.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Investigate personally identifying information (PII) and privacy issues				Disseminate findings on sensitivity of PII attributes	
Activity 2: Develop appropriate text analysis techniques to identify sensitive information from unstructured data	Research appropriate text analysis techniques to identify sensitive information from unstructured data	Develop appropriate techniques for identifying sensitive information from unstructured data			

Objective 3.3.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Research state-of-art multimodal deep learning techniques for identifying private sensitive information	Study and document state of art multimodal deep learning techniques	Document and disseminate the findings on the determination of appropriate multimodal techniques for detecting sensitive information			

Activity 2: Investigate appropriate techniques for identifying discriminating and stigmatizing information	Document and disseminate the findings of state-of-art techniques for identifying discrimination information	Document and disseminate the findings on the determination and development of appropriate techniques for identifying stigmatizing information		

Activity 3: Develop appropriate deep learning text analysis techniques to accurately remove discriminating and stigmatizing information	Design deep learning techniques for removing discriminating and stigmatizing information Implement deep learning techniques for removing discriminating and stigmatizing information	Test and document the efficiency of the techniques and improve them when necessary		
Objective 3.3.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5

Activity 1: Develop a risk assessment method for possible privacy breach given the amount of personal identifying information disclosed/published			Research the relationship among PII attributes and how the release of one attribute would affect the overall privacy	Develop a risk assessment framework for assessing the possible risk of privacy breach given the amount of PII has been released	Document the findings from the testing and refining of the framework
Activity 2: Develop appropriate techniques for safeguarding sensitive information by helping end users monitor and proactively control the release of their personal information				Develop appropriate techniques for safeguarding sensitive information based on the research from activity 1 and 2	Document the findings from the resting and refining of the framework

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

Objective 3.4.a: Explore deep learning-based techniques to detect cross-media discrimination
Objective 3.4.b: Develop deep learning models and a causality-based deep learning framework for robust hate speech detection

Goal 3.4 Output Metrics

Publications, presentations, and reports (5):

-- Two (2) conference papers
-- One (1) journal paper
-- One (1) thesis
-- One (1) proposal

Workshops, demonstrations, and trainings (1):

-- One (1) tutorial given at major AI conference

Datasets and algorithms (1):

-- One (1) algorithm (code with associated training and testing data)

Objective 3.4.a	**Specific Milestones**					
	Year 1	Year 2		Year 3	Year 4	Year 5
Activity 1: Use deep convolutional neural networks (CNN) to recognize discrimination-sensitive objects from images	Initiate theoretical investigation on using CNN to recognize discriminatory objects	Complete design and implementation of the CNN-based model				
Activity 2: Adopt long short-term memory (LSTM) network to model the text	Initiate theoretical investigation on using LSTM to model discriminatory text	Complete design and implementation of the LSTM-based model				
Activity 3: Utilize bilinear model to capture the implicit relationship between the detected discrimination-related objects and the text		Initiate the theoretical investigation on the implicit relationship between the detected discrimination-related objects and the text				
Objective 3.4.b	**Specific Milestones**					
	Year 1	Year 2	Year 3	Year 4	Year 5	
Activity 1: Design and implementation of contextual embedding-based model for coded hate speech detection				Disseminate comparison of multimodal hateful image/text detection models	Publish developed model(s) and framework	

Goal 3.7 (SA7)	**Cryptography-Assisted Secure and Privacy-Preserving Learning**
Objective 3.7.a: Develop cryptography-aware privacy-preserving machine learning methods, and develop privacy protection for classification input data in machine learning applications.	
Goal 3.7 Output Metrics **Publications, presentations, and reports (5):** -- Two (2) conference papers -- One (1) journal paper -- One (1) thesis -- One (1) proposal **Datasets and algorithms (1):** -- One (1) algorithm (code with associated training and testing data)	

Objective 3.7.a	Specific Milestones				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Research the hybrid use of existing cryptography techniques and differential privacy in federated machine learning	A survey of existing cryptography techniques and their applications in differentially private federated learning				
Activity 2: Develop new applied cryptography techniques to use in combination with differential privacy for federated machine learning	Design of preliminary new cryptography techniques used for differentially private federated learning	Design of new cryptography techniques used for differentially private federated learning			
Activity 3: Assessment of privacy-preserving machine learning				Empirical and user study of privacy-preserving learning solutions	Analytical assessment of privacy-preserving learning solutions

Goal 3.8 (SA8)	**Objective 3.8.a:** Develop a deep-learning model for causality-based fair node classification in social networks.

Goal 3.8 Output Metrics

Publications, presentations, and reports (5):

-- Two (2) conference papers
-- One (1) journal paper
-- One (1) proposal

Datasets and algorithms (1):

-- One (1) algorithm (code with associated training and testing data)

Objective 3.8.a	Specific Milestones				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Derive causality-based fairness notions for social networks				Social network data collected	
Activity 2: Develop techniques for causal inference on networked data					Disseminate developed techniques
Activity 3: Develop and evaluate a deep learning model for fair node classification					Publish model and evaluation results

4. 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 the following scientific challenges: 1) difficult to analytically assess collective impact of social media and networking on societal polarization and other social phenomenon due to fragmentation of debates and discussions in the existing social media and networking platforms; 2) lack of a central social media and networking platform for debate and discussions on important issues at national and international levels; 3) hard to detect mis/dis-information, how it disseminates, and assess its impact; 4) mining and using social media/networking data for logistics planning for disaster response; 5) content-based indexing of unstructured and multimedia data on social media platforms and towards an integration with decision-making systems through deep learning methods; and 6) arduous to visualize large social network data.

4.1. Advancing the State of the Knowledge:

Below is the list of envisioned advancements to the state of the knowledge by each sub-theme (SM1 through SM4).

- SM1 will advance the state of the knowledge in argumentation polarization modeling by developing innovative quantitative opinion polarization techniques to model the formation and evolution of opinion polarization in large-scale cyber argumentation and deliberation with social networks.
- SM1 will elevate methods and techniques to predict individual or collective opinions on single or multiple solutions of issues using collaborative filtering and machine learning-based techniques. SM1 will improve social network methodology and social network sampling and data generation.
- The proposed research in SM2 advances our understanding of the role of Information and Communication Technology-mediated communications in the formation of emergent organizations with implications to business, marketing (explaining viral behaviors), and many other settings. SM2 is of particular interest to information scientists exploring the influence of social systems on user behaviors; studying ties between people, technology, and institutions; examining organizational structures, roles, and crowd processes; investigating the notions of individual and collective identities in a variety of information systems supporting crowdsourcing, citizen participation, eGovernance, crisis and disaster management, and several other manifestations of emergent organizations. SM2 would contribute to the theory of collective action to model the dynamics of deviant cyber behaviors, borrow from the literature on collective identity formation to explain the motivation needed to sustain such coordinated acts, assimilate factors pertaining to collective failures/success, and leverage notions of hypergraph to model complex (multidimensional and supra-dyadic) relations commonplace among members of deviant groups.
- The proposed research in SM2 advances the literature on cyber-collective actions and study the role of social media in organization and coordination of cyber social

movements from individual, community, inter-organizational, and transnational perspectives.

- The proposed research in SM2 develops socio-computational predictive models that are efficient, reliable, scalable, explainable, reproducible, and theoretically grounded to help understand behaviors from social media platforms. SM3 will allow decision making processes to utilize multi-source, heterogenous and multimodal data towards better performance, as well as expand the scope of learning and artificial intelligence techniques to multimedia data. SM3 will afford applications such as disaster recovery to take advantage of vast, increasingly popular multimedia data that is often unstructured and insufficiently indexed.
- SM4 provides methods for indexing and fusing transportation infrastructure status data from a variety of sources (e.g., social media, satellite imagery, traffic camera videos).
- SM4 applies credibility detection methods for transportation infrastructure status data on social media.
- SM4 enables the use of near-real-time transportation infrastructure status data in logistics planning methods to support disaster response. SM4 also creates new vehicle routing models and solution approaches for complex real-time logistics planning problems in disaster response.

4.2. Project Implementation

Goal	Goal Name	Lead(s)	Team Members
SM1	Mining cyber argumentation data for collective opinions and their evolution	Gauch	Gauch, Adams, S. Yang
SM2	Socio-computational models for safer social media	Agarwal	Agarwal, Milburn
SM3	Auto-annotation of multimedia data	Milburn	Dagtas, Milburn, Cothren
SM4	Informing disaster response with social media	Milburn	Milburn, Dagtas, Liao, Cothren, Talburt, Ussery, Nachtmann, Rainwater, Karim, Celebi

Summary of changes in 2023 strategic plan revision

There are several activities in SM4 where we had planned to develop approaches and test them on **two** disaster scenarios. We are suggesting modifying this deliverable to one disaster scenario instead of two. We have had much attrition on SM4 since the project began (first with Frank Liu, then Justin Zhan, Christopher Angel, and most recently Xiao Huang, the latter of whom has been very instrumental in moving SM4 forward). With the decrease in resources, decreasing the number of disaster scenarios we test will enable us to retain the planned methodology development tasks The activity under 4.4 related to GIS mapping of the disaster scenarios has been removed due to the loss of SP Angel and no comparable replacement with similar expertise.

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

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

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

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

Goal 4.1 Output Metrics

Publications, presentations, and reports (5):

-- Five (5) peer-reviewed journal and/or conference papers (articles)

Assessments, questionnaires, and surveys (1):

-- One (1) IRB-approved questionnaire for collecting discourse data

Software (1):

-- One (1) cyber discourse social network platform

Datasets and algorithms (2):

-- Two (2) advanced natural language algorithms (code with associated training and testing data)

Objective 4.1.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Brainstorming the features needed for the cyber discourse social network platform	Key features for the platform are determined				
Activity 2: Software design for the platform	Software design document is finalized				
Activity 3: Implement the platform		Platform is implemented			
Activity 4: Test the platform			The platform is tested		
Activity 4: Deploy and disseminate the platform				The platform is deployed and disseminated	

Objective 4.1.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Design the "hot button" questionnaire items	Develop questionnaire for collecting discourse data				
Activity 2: Develop Individual and network question measures and submit IRB consent form	The question measures are determined, and IRB protocol is approved				
Activity 3: Data collection by engaging students in the "hot button" issue discussions			The discussion data are collected, and two refereed articles are submitted/published		
Objective 4.1.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Develop advanced natural language processing algorithms	The advanced natural language processing algorithms are developed				
Activity 2: Test the natural language processing algorithms using the existing data		The algorithms are tested using existing data			
Activity 3: Validate the natural language processing algorithms using the data collected by the developed platform			The algorithms are validated using the data collected from the platform and three articles are submitted.	The algorithms are adapted for publicly available social media data. Three articles are submitted.	New algorithms are developed for publicly available social media data. Three articles are submitted.

Goal 4.2 (SM2)	**Socio-computational models for safer social media**				
Objective 4.2.a: Characterize online information environment (OIE) **Objective 4.2.b:** Develop socio-computational models to identify key actors and key groups of actors **Objective 4.2.c:** Study tactics, techniques, and procedures (TTPs) of deviant cyber campaigns **Objective 4.2.d:** Develop socio-computational models to measure power of a cyber campaign					
Goal 4.2 Output Metrics **Publications, presentations, and reports (4):** -- One (1) taxonomy -- One (1) journal paper -- Two (2) conference presentations **Applications and platforms (1):** -- Socio-computational models for OIE and TTP implemented in a web-based application **Datasets and algorithms (2):** -- Two (2) socio-economic models and associated datasets					
Objective 4.2.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Study social media spaces and cyber campaigns to identify characteristics and features	Social media platforms identified; Cyber campaigns identified; Characteristics and features identified				
Activity 2: Create a taxonomy of dimensions to characterize social media spaces	Taxonomy developed				
Activity 3: Revisit and adjust taxonomy as social media space evolves		Revised taxonomy developed and published based on new social media, campaigns, features, and characteristics			

Objective 4.2.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Review cyber campaigns and social media data	Data sources identified; Data acquisition procedures established Database setup	Data reviewed and modifications incorporated; Data collected and shared with DART teams	Cyber campaign data reviewed, additional data collected if needed; Data published according to NSF's data sharing policies and social media platforms' terms and agreements		
Activity 2: Identify behavioral traits for key actors and key groups by leveraging OIE characterization		Key actors and key groups identified empirically	Behavioral traits of key actors and key groups identified		
Activity 3: Develop computational model(s) for key actor and key group discovery		Model(s) developed	Model(s) refined; Published in peer reviewed forums; Model transitioned to usable web-based application		
Activity 4: Evaluate model(s)			Model(s) evaluated, and refinements proposed		
Objective 4.2.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Review campaigns, social media platforms, and involved actors and groups			Campaigns identified for further review	Review of campaigns concluded with study methodology and findings published as case studies in peer reviewed forums	

Activity 2: Identify and document tactics, techniques, and procedures (TTPs) (e.g., platform orchestration, botnets, inorganic behaviors, stalking, pacing, leading, threadjacking, hashtag latching, boosting, echo chambers)			TTPs identified and documented/published		
Activity 3: Examine tactics, techniques, and procedures (TTPs) vis-a-vis OIE characterization				TTPs categorized based on the OIE taxonomy developed	TTP categorization revised based on behavior evolution
Objective 4.2.d	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Review OIE characterization and TTPs to identify campaign attributes				Review of TTPs and OIE categorization completed	Campaign attributes identified to help assess its power
Activity 2: Develop computational model to measure power of a campaign by integrating attributes, key actors, key groups, collective action theory (and other theoretical constructs)				Model(s) developed	Model(s) refined Published in peer reviewed forums Model transitioned to usable web-based application
Activity 3: Evaluate model(s)					Model(s) evaluated, and refinements proposed

Goal 4.3 (SM3)	**Auto-annotation of multimedia data**

Objective 4.3.a: Develop multimedia indexing methods for social media data
Objective 4.3.b: Design and implement deep learning methods for multimedia data
Objective 4.3.c: Build Integrated smart applications based on unstructured multimedia data

Goal 4.3 Output Metrics

Publications, Presentations, and Reports (3):

-- Two (2) journal articles
-- One (1) conference paper

Datasets and Algorithms (3):

-- One (1) algorithm for Indexing
-- One (1) algorithm for deep learning for multimedia data
-- One (1) algorithm for integrated smart applications

Objective 4.3.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Define priorities and characteristics for multimedia data on social platforms	Key characteristics defined				
Activity 2: Design and build algorithms for efficient retrieval of nontraditional data		Image, video, and 3D retrieval methods defined and tested			
Activity 3: Advanced querying methods and implementation of interfaces with multimedia capabilities				Querying tools for multimedia data	

Objective 4.3.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Define learning objectives for social data from multimodal sources	Identify and define three major learning objectives document				
Activity 2: Develop detection and classification methods		Object and event detection methods implemented			
Activity 3: Deep learning applied to multimedia data and related indexing mechanisms			Two learning methods developed and tested		
Activity 4: Verification of learning objectives and methods developed in Activity 2 and 3					Testing and reporting
Objective 4.3.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Define key applications for the implementation and testing of the indexing and retrieval mechanisms	Three key applications defined				
Activity 2: Integration with disaster response and other applications defined in Activity 1			Integration methods developed		
Activity 3: Define ethical and legal perspectives for the use of multimedia data		Use and access-based ethics principles defined for multimedia social data			

Goal 4.4 (SM4)	**Informing disaster response with social media**

Objective 4.4.a: Extract and index content describing transportation infrastructure status from social platforms
Objective 4.4.b: Fuse data from social platforms describing transportation infrastructure status with other data sources
Objective 4.4.c: Assess credibility of data inputs from Objectives 4.4.a and 4.4.b
Objective 4.4.d: Develop routing algorithms that use inputs from Objectives 4.4.a-4.4.c to support routing for disaster response

Goal 4.4 Output Metrics

Publications, presentations, reports (8):

-- Three (4) journal articles
-- Two (2) conference papers
-- One case study of a natural disaster scenario

Applications and Platforms (1):

-- GIS routing platform informed with social media feeds

Datasets and Algorithms (5):

-- One (1) schema for mapping each datum to a probability describing its credibility
-- Four (4) algorithms

Objective 4.4.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Study social platforms to identify types of content that describe transportation infrastructure status	Identify and define social platform content types of interest (e.g., image, video, text, etc.)				
Activity 2: Develop and implement extraction techniques for identified types of social platform content		Develop social platform extraction techniques for content types of interest and pilot test on at least two disaster scenarios			
Activity 3: Develop and implement indexing techniques for extracted social platform content		Develop and implement indexing techniques for extracted social platform content and pilot test on at least two disaster scenarios			

Objective 4.4.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Identify other data sources that contain real-time information regarding transportation infrastructure status	Identify and define content types of interest (e.g., satellite imagery, traffic cameras) from sources other than social platforms				
Activity 2: Obtain and index transportation infrastructure data from other data sources		Obtain and index identified content types for at least two disaster scenarios			
Activity 3: Develop and implement data fusion techniques to combine data from social platforms and other sources			Fuse data from social platforms and other sources for at least two disaster scenarios		
Objective 4.4.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Develop and implement machine learning classifiers to detect quality of information	Obtain testing data from social platforms	Develop machine learning classifiers to detect false or low-quality information			
Activity 2: Develop and implement schema to map credibility/quality scores for data to probabilistic inputs of transportation infrastructure status			Develop schema for mapping each datum to a probability describing its credibility		

Objective 4.4.d	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Identify critical routing problems with application in disaster response	Select one disaster response routing problem variants using Milburn's existing qualitative interview data				
Activity 2: Develop models of identified disaster response routing problems and assess state of the literature	Conduct literature review for identified routing problem variants and publish journal article synthesizing review with qualitative data from 4.4.c.1				
Activity 3: Develop and implement routing algorithms for identified routing problem variants		For routing problem variant, develop, validate and test at least one solution algorithm each on randomly generated test networks		Publish journal article	
Activity 4: Demonstrate models and solution approaches via pilot study of one or more disaster scenarios		Select one disaster scenario for pilot	Obtain test data for selected disaster scenario	Test algorithm on real disaster scenario pilot	Publish journal article

5. Learning and Prediction

Research in this topical area will focus on various techniques in prediction interpretation for large-scale, deep learning using multi-source integrated data sets. Goal 5.1 Statistical Learning – Random Forests (RF) for Recurrent Event Analytics integrates the RF algorithm with classical statistical methods including the non-parametric MCF estimator and parametric NHPP, which will allow dynamic and unstructured covariate information to be incorporated into a tree-based method. Goal 5.2 Statistical Learning – Marked Temporal Point Process Enhancements via Long Short-Term Memory Networks combines the marked temporal point process with LSTM to model the dynamic event data without assuming any parametric forms. Goal 5.3 Deep Learning – Novel Approaches uses unsupervised learning to build a deep network from a series of shallow networks, each having simpler and more interpretable objective functions to cope with the stochastic nature real-life events, address the high dimensionality of data and action spaces, explore topological and group compression approaches, and investigate design and interpretability issues in deep reinforcement learning. Goal 5.4 Deep Learning – Efficiency and Specification develops new objective loss functions incorporated in any deep networks to solve million-scale problems to automatically and efficiently cluster easy- and hard- samples to optimize deep learning models and better model the distributions of deep features extracted from hard samples in more accurate ways. Goal 5.5 Harnessing Transaction Data through Feature Engineering investigates a framework of utilization transaction data for improved prediction and decision making in medicine and business with a particular goal to improve the identification of opioid use disorder and intervention timing by enhancing prediction and utilizing valuable information from transaction data.

5.1. Advancing the State of the Knowledge:

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 will focus on applying statistical learning techniques alongside more advanced deep learning techniques to address three major challenges.

1. Violation of fundamental statistics principles: In big data environments where sample size is large (in some extreme cases, every single individual in a population can be observed), many basic principles in statistics need to be revised, such as the classical relationship between population and sample, the assumption of independent and identically distributed (iid) random variables, and so on. This challenge requires statistical learning methods in big data environments to be equipped with capabilities for addressing heterogeneity and hidden sub-populations within big datasets.
2. Mode specification and interpretation: With high-dimension, dynamic, unstructured covariate information which has been made possible by big data technologies, it is no longer realistic to specify parametric assumptions that adequately describe the

complicated nonlinear dynamics between responses and covariate information. In addition, a certain amount of covariate information is almost always redundant in the statistical modeling or domain knowledge perspectives. Efficient variable selection using non-parametric methods has become one of the most sought-after capabilities in big data analytics.

3. Computing in big data environments: It is already a mainstream practice in today's industry to store massive data on distributed big data platforms. This trend requires the statistical computing to be performed on distributed/parallelized computing nodes—an extremely critical issue but is often ignored in traditional statistical modeling.

We will investigate these challenges surrounding high-dimensional, dynamic, and unstructured data sets and explore solutions in the domains of genomics, transaction scenarios in eCommerce, and supply chain logistics.

5.2. Project Implementation

Goal	Goal Name	Lead(s)	Team Members
LP1	Statistical Learning – Random Forests for Recurrent Event Analytics	Vazquez	Chimka
LP2	Statistical Learning – Marked Temporal Point Process Enhancements via Long Short-Term Memory Networks	Rainwater	Vazquez
LP3	Deep Learning – Novel Approaches	Karim	Celebi, Luu, Kim, Cheng, Alroobi, Stine, Al-Shami
LP4	Deep Learning – Efficiency and Specification	Luu	Le, Rainwater
LP5	Harnessing Transaction Data through Feature Engineering	S. Zhang	Nachtmann

The Learning and Prediction research theme will be implemented through the achievement of five major goals and their associated objectives and supporting activities.

Summary of changes in 2023 strategic plan revision

LP1 was revised due to the loss of Liu and replacement with Vazquez, who has different expertise. Some new activities and milestones have been introduced to incorporate feedback from EAB and other teams. Some activities and milestones under 5.3.b were updated based on the expertise and capabilities of Stine and Al-Shami, after the loss of personnel Schrader and Kursun.

Goal 5.1 (LP1)	**Statistical Learning – Random Forests for Recurrent Event Analytics**

Objective 5.1.a: Create the Random Forests for Recurrent Event Analytics, which integrates the RF algorithm with classical statistical methods allows dynamic feature information to be incorporated into a tree-based method.

Objective 5.1.b: Create the Gradient Boosting method for Recurrent Event Analytics, which integrates the boost trees with classical statistical methods allows dynamic feature information.

Objective 5.1.c: Perform comparison study between the methodologies above and identify future research directions

Goal 5.1 Output Metrics

Publications, presentations, reports (7):

-- Two (2) manuscripts submitted for publication
-- Two (2) student theses or dissertations proposed
-- One (1) student theses or dissertations defended
-- One (1) submitted research proposals

Objective 5.1.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Establish a preliminary model, and complete the theoretical investigation	Complete the preliminary theoretical investigation on the proposed modeling approach				
Activity 2: Complete the coding and numerical examples; write, submit, revise paper		Complete the numerical studies, and submit a research paper			
Activity 3: Revise paper and research outcomes dissemination through conferences			Complete the model improvement and paper revision		

Objective 5.1.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Establish a preliminary model, and complete the theoretical investigation under Obj 5.2			Complete the preliminary theoretical investigation on the proposed modeling approach		
Activity 2: Complete the coding and numerical examples; write, submit, revise paper				Complete the numerical studies, and submit a research paper	
Objective 5.1.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Perform the numerical comparison study and identify future directions					Publish paper
Objective 5.1.d	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Evaluate the capabilities of the available subsampling methods for building RFs					Publish paper
Activity 2: Develop effective methods to select attractive subsamples for RF to provide accurate predictions over the full big data set and future data					Prepare a manuscript

Goal 5.2 (LP2)	**Statistical Learning – Marked Temporal Point Process (MTTP) Enhancements via LSTM Networks**

Objective 5.2.a: Develop methodology integrating the marked temporal point process (MTPP) with long short-term memory networks (LSTM)

Objective 5.2.b: Develop unsupervised and dynamic degradation labeling strategy for remaining life modeling

Objective 5.2.c: Evaluate and assess approach on real-world discrete data sets

Goal 5.2 Output Metrics

Publications, presentations, reports (6):

-- One (1) conference paper
-- Two (2) conference presentations
-- Two (2) journal publications
-- One (1) case study

Workshops, demonstrations, and trainings (2):

-- One (1) graduate seminar
-- One (1) industry workshop

Objective 5.2.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Formally define approach integrating intensity function of MTTP into LTSM	Submit conference paper with initial model				
Activity 2: Establish proof-of-concept implementation of MTTP/LTSM approach	Present conference paper with preliminary results of implementation	Submit journal article with conceptual findings and initial implementation of approach			
Activity 3: Perform benchmark of MTTP/LSTM tests on small simulated data sets		Publish white paper and GitHub repository with benchmark tests/results			

Objective 5.2.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Assess data collected from Activities 1 and 2 of Obj 5.2c to define methodology computation performance requirements		Produce system requirements document for V2 implementation			
Activity 2: Develop unsupervised and dynamic degradation		Submitted conference paper	Alpha version of approach benchmarked against NASA dataset	Conference paper presentation	
Activity 3: Validate performance of scalable implementation on large-scale simulated datasets					Publish paper
Objective 5.2.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Acquire healthcare IoT datasets	Publish curated data to GitHub				
Activity 2: Acquire civil infrastructure datasets	Publish curated data to GitHub				
Activity 3: Establish baseline performance of predictions made by existing approaches applied to datasets from Obj 5.2c Activities 1 and 2		Publish white paper and GitHub repository with benchmark predictions	Make conference presentation on baseline performance of predictions compared against existing approach		
Activity 4: Assess methodology implementation from Objective 2 on real-world datasets collected in Obj 5.2c Activities 1 and 2				Workshop with industry stakeholders providing data in Activities 1 and 2	Prepared manuscript journal article covering practice-based computational efforts

Goal 5.3 (LP3)	**Deep Learning – Novel Approaches**

Objective 5.3.a: Extract explanatory features from Deep Network
Objective 5.3.b: Address high dimensionality issues in Deep Reinforcement Learning (DRL) using algebraic and topological methods
Objective 5.3.c: Designing a novel rewarding model, and addressing interpretability issues in DRL

Goal 5.3 Output Metrics

Publications, presentations, reports (10):

-- Six (6) conference papers
-- Two (2) Journal publication
-- Two (2) Master's theses;

Workshops, demonstrations, and trainings (3):

-- One (1) workshop
-- One (1) teaching module
-- One (1) Special topics undergraduate class offering

Datasets and algorithms (1):

-- One (1) dataset

Objective 5.3.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Development of novel self-supervised and flow-based deep learning approaches	Development of the first unsupervised convolutional area and the first flow-based deep learning approach	Adaptation of the novel methods for application to the tactical agility dataset	Development of the areal postprocessing of the self-supervised model and the Flow Autoencoder as deep graphical model with dual objectives	Building of the deep stacked self-supervised model	Adaptation of the developed novel deep networks to one of the datasets of the DC thrust
Activity 2: Developing a library of classifiers for benchmarking	Development of linear dimensionality reduction methods	Development of the standard autoencoder method	Development of signal feature extraction tools	Development of the clustering/segmentation tools	Development of the baseline classifier for the DC dataset

Activity 3: Application of the developed methods on real-world datasets	Application of the developed methods with applications on natural images and textures, and classification of a malware dataset	Comparisons of the developed methods/libraries on the tactical agility dataset and malware dataset	Exploratory studies with the datasets of the researchers in the DC thrust of DART. Evaluate various deep learning models on the malware dataset	Demonstration of the quality of the features extracted by the self-supervised deep network and flow-based autoencoder methods on the tactical agility dataset	Comparisons of the developed methods/ libraries on the dataset of the DC thrust. Develop the deployment platforms for the developed classification systems
Objective 5.3.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Investigation of PH in the context of Deep Learning	Investigate group theoretical approaches to generalized NN architecture design within the context of interpretability for Objectives 5.3a (Activity 1) and 5.3c (Activity 1)	Exploration of internal topologies in generalized NN structures using TDA and PH to identify architectures which address high dimensionality and enhance the developments in Objectives 5.3a (Activities 1 and 2) and Objective 5.3c (Activity 2)	Study of topological representation of input space and objects in the context of Deep Learning		
Activity 2: Incorporation of PH-enhanced deep learning models in DRL				PH representation of environments/ scenarios in DRL	Fine-tuning PH-representation based on the feedback from 5.3.c activity
Objective 5.3.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Design an improved reward process for DRL	Development of a generalized model of reward function in DRL addressing the issues with both sparse and dense feedback	Development of use cases of the developed reward model			

Activity 2: Explore PH based filtering to optimize scenario space			Development of a PH based scenario-space optimization algorithm	Development of DRL models exploiting optimized scenario-space	
Activity 3: Explore DRL interpretability					Causal inference and do-calculus modeling for DRL

Goal 5.4 (LP4)	**Deep Learning – Efficiency and Specification**

Objective 5.4.a: Create Novel Deep Learning Networks Executable with Reduced Computational Resources and Assess Performance
Objective 5.4.b: Address Low-cost Deep Learning Algorithmic Analysis and Challenges
Objective 5.4.c: Explore Low-cost Deep Learning Applications in Natural Images and Medical Images

Goal 5.4 Output Metrics

Publications, presentations, reports (5):

-- One (1) journal article
-- Two (2) conference papers
-- One (1) invited presentation at an Arkansas Institution
-- One (1) invited presentation elsewhere

Workshops, demonstrations, and trainings (3):

-- One (1) conference workshop
-- One (1) tutorial
-- One (1) teaching module

Objective 5.4.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Develop and demonstrate new low-cost deep neural network algorithms	Develop Teacher - Student Distillation Deep Learning Algorithms; Develop Light-weight Deep Learning Algorithms	Develop Deep Network Compression Algorithms; Develop Deep Network Pruning Algorithms			
Activity 2: Develop new objective loss functions in deep neural networks			Develop auto and semi-auto deep network searching algorithms to discover optimal deep neural networks given a particular application and data training sets.	Deliver new novel deep network loss functions in combination with the deep networks to improve the performance	
Activity 3: Develop low-cost deep learning methods for high-dimensional data				Develop Depth-wise Separable Deep Network for Volumetric Data	Optimize and implement the deep networks on low-cost computers; Fine-tuned and evaluated in on-the-edge devices.
Objective 5.4.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Mathematically analyze the proposed deep learning methods	Develop analytic approaches to the proposed methods in Activities 1.1	Develop analytic approaches to the proposed methods in Activities 1.2			
Activity 2: Improve the computational time and accuracy performance			Improve the computational time and accuracy performance on the standard databases and challenges	Improve the performance for the loss functions in deep learning	
Activity 3: Analyze the complexity					Analyze the complexity of the presented low-cost

					deep learning methods when deploying on the dataset from the other thrusts in this project
Objective 5.4.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: The developed deep learning algorithms will be optimized and implemented in two applications, including natural images and medical imaging	Develop Low-cost Deep Learning Approaches in Image Classification	Develop Low-cost Deep Learning Approaches in MRI Segmentation			
Activity 2: The developed deep learning algorithms will be further optimized and deployed in high dimension data, such as: videos and medical MRI volumetric data			Develop Low-cost Deep Learning Approaches in Automatic Human Activity Recognition in videos.	Develop Memory-Effective Deep Network for Volumetric Data	

Goal 5.5 (LP5)	**Harnessing Transaction Data through Feature Engineering**				
Objective 5.5.a: Design advanced feature engineering techniques for high-dimensional temporal data **Objective 5.5.b:** Create an improved prediction and decision-making framework incorporating feature engineering with health transaction data **Objective 5.5.c:** Employ and validate the new framework for prediction and decision making with business transaction data					
Goal 5.5 Output Metrics **Publications, presentations, reports (7):** -- Three (3) journal publications -- Three (3) conference presentations -- One (1) doctoral dissertation					
Objective 5.5.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Extract and process APCD data	Obtain and prepare cleaned data for research				
Activity 2: Extract and engineer features from the high-dimensional temporal data	Acquire features that are highly representative				
Activity 3: Explore and test automation of feature engineering in transaction data		Achieve automated feature engineering	Optimize feature engineering with transaction data	Improve robustness of automated feature engineering	
Objective 5.5.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Develop deep learning prediction models and algorithms with feature engineering	Complete selection and testing of deep learning models	Improve the predictive models			
Activity 2: Incorporate representation learning in prediction with engineered features		Implement and test autoencoders	Compare and test representation learning methods		

Activity 3: Compare and validate prediction with existing models				Identify the best and robust predictive models	
Objective 5.5.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Extract and process business transaction data			Obtain and prepare business transaction data		
Activity 2: Employ feature engineering and prediction in the business transaction data			Employ and test various models for prediction	Establish the link between prediction and decision making	
Activity 3: Validate the developed framework with the business transaction data					Finalize and promote a robust framework for transaction data

6. Education

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.

DART will foster the establishment of a Statewide Data Science Educational Ecosystem by:

- Developing a combination of model programs, degrees, pedagogy, and curriculum including certificates, an associate of science in data science; and a Bachelor of Science in data science with minors or concentrations.
- Providing resources and training for educators including $5,000 Seed Grants for project-related Education & Broadening Participation; Career Development Workshops for project participants and educators; and K12 teacher professional development on data science topics.
- Providing educational opportunities inside and outside the classroom for students. Undergraduate and graduate research assistantships in DART labs will be funded along with intensive data science and computing summer camps for undergraduates and research-based capstone projects and internships with industry partners.
- Ensuring broad participation to impact the pipeline of data science skilled workers through Summer Undergrad Research Experiences in DART labs for underserved students, scholarships for underserved students to the ASRI; and by connecting students to opportunities through the ACDS.

6.1. Advancing the State of the Knowledge:

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. When the project is complete data science programs will be widespread across the state and will be available to all interested Arkansans.

6.2. Broader Societal Impacts:

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 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.

DART will include a data science summer institute for undergraduates, summer internships and research experiences, increased data science educational opportunities, and revamped curriculum to include relevant data science topics and capstone projects. Developments in data cyberinfrastructure will increase sharing information among educational institutions, research institutions, and industry.

6.3. Project Implementation

Goal	Goal Name	Lead(s)	Team Members
ED1	Contributing to the Data Science Educational Ecosystem by developing a combination of model programs, degrees, pedagogy, and curriculum including certificates; an associate of science in data science; and a Bachelor of Science in data science with minors or concentrations.	Schubert and Addison	Chowdhury (UAPB), Scott (Shorter), Shoultz, Zhang (PSC), Hillyer, Blanchett (Shorter), Karim (SAU), Berry (NAC), Qualls (A-State), Zeng (ATU)

The project team will create an informed, scalable, and replicable set of model programs, degrees, pedagogy, and curriculum. This provides an excellent opportunity for the State's colleges and universities to work together, identify key areas of concentration based on regional skills, needs, and interests, and optimize investments in people, facilities, and resources across academia, government, NGOs, and industry. The following activities and initiatives are designed to populate the front of the education pipeline with a rich, diverse set of students.

Establishing Data Literacy and Responsibility

Educating and developing a data-literate and data responsible workforce requires more than modifying existing courses. Companies, government agencies, and NGOs need a workforce whose education, training, and experience includes domain of application, addressing ethical issues and social implications, and considering a hands-on practicum or capstone integrative experience. Pre-existing courses typically cover essential material but do so in the context of those majors necessitating new courses in an overall design that ensures program core curriculum continuity across the courses, internships, research, and years. Developing the base rigorous data science major, with concentrations, provides the basis from which associate degrees, minors, tracks, additional concentrations, technical certificates, etc., can be developed.

Courses developed for a rigorous data science major begin at an introductory level and progress through a series of new topics of increasing depth and complexity while remaining connected to previous courses, removing the scaffolding along the way. Partnering with institutions at all levels during the curriculum development process ensures smooth transitions for students from the 2- to 4-year colleges while simultaneously ensuring outcome and learning continuity. Specialization is likely to begin in the third and fourth years of a 4-year suggested plan of study which can be implemented through domain-focused concentrations. More advanced and/or domain specific programs (M.S., Ph.D.) can also be developed leveraging the research programs and rigorous data science foundations at the undergraduate level. Designing, developing, and pilot implementing such a system program is proposed next.

Model Postsecondary Programs

UAF recently developed a rigorous Bachelor of Science in Data Science with multiple concentrations as its foundation. This program, cooperatively developed by the colleges of engineering, business, and arts and sciences, is designed with a core set of data science courses for all students ("hub") and a set of concentrations for domain expertise ("spokes") providing a means for adding concentrations (domains, "spokes") as appropriate. The project team will use this program as the basis, reference point, and template for expanding postsecondary data science education throughout the State. In order to provide rigorous Data Science degrees, the focus on developing and expanding the program will be based on the employer workforce needs and a focus on creating a process and pipeline through the State's higher education institutions that is rigorous, replicable, scalable, and achievable.

During the development of the UAF program, an employer needs survey was conducted with the help of an industry advisory committee. Employer input on workforce requirements and needs were strongly stated: the existing two-year, certificate, and master's degree-level programs are not producing students who have the skills they need. While these graduates have the vocabulary and ability to use basic tools, they lack the rigorous foundations, training, and experience in critical thinking, analysis, synthesis, domain knowledge, and communication skills these companies need to be successful. And, there was no consistency across programs which compounded the lack of rigor. By designing an overall program at the state level, a key challenge in matching 2-year and 4-year curricula can be addressed. The program will align with recent National Academies of Sciences, Engineering, and Medicine (NASEM) recommendations to "evolve a range of educational pathways to prepare students for an array of data science roles in the workplace."

A team of faculty from multiple colleges, departments, and campuses will use the UAF Data Science program as a base to develop a full range model postsecondary curriculum. The curriculum will consist of a set of core courses with options for electives that could vary by campus meeting the requirements for student outcomes, pedagogy, and core curriculum continuity. The pilot institutions include two- and four-year campuses around the state that will implement iterations of the larger model. Two-year campuses can choose to implement an associate of science degree or a technical certificate, while four-year institutions can choose any combination of technical certificate, associate degree, bachelor's degree, or minor concentration. This method will ensure consistency in quality, availability, and transferability for students statewide while allowing institutions to maintain competitive advantages and autonomy.

UCA has domain specific concentrations in data science in Business, Computer Science, and Mathematics. Many of the aspects of the UAF B.S. program are already implemented in these distinct concentrations and faculty at UCA are currently working on packaging these courses into a standalone B.S. degree. Thus, early in the project there will be two four-year degree programs that will facilitate the development of two- and four-year programs across the state.

Integration with Research Projects in the Proposal

A key opportunity in the design and development of the degree program will be to leverage the topics in the research program as examples for the courses, directly and indirectly, and to integrate these into the curriculum, practicums, and senior design/capstone courses. Such topics

include automated data curation, curating heterogenous data, antidiscrimination and preserving privacy in data.

Institutional Partnerships and Active Participation: 11 institutions have agreed to participate in implementing a technical certificate or associate of science in data science. Of the 11, two are HBCUs, and 10 are two-year campuses. Three other institutions, including an additional HBCU and two academic research institutions, will implement a B.S. in data science based on our model curriculum.

Summary of changes in 2023 strategic plan revision

The middle school coding block activity was removed. The project champion, Daniel Moix, left the state for another opportunity and we could not identify a replacement candidate. After discussing the issue with other educators and the Arkansas Department of Education, we learned that this is no longer a critical need. We will replace this with a series of events and workshops to raise interest and awareness of data science and data analytics across the K20 pipeline and throughout the state.

The following activities and milestones regarding program accreditation were removed, due to the timing of when programs can pursue accreditation. A certain number of students must matriculate from a degree program before an ABET accreditation application can be submitted, and the programs being established under DART will not matriculate students before the end of the project. We have identified the appropriate accreditation bodies and have prepared the curriculum for accreditation by those bodies.

- Activity 11: When accreditation is available, propose first-pass consultative reviews by those bodies for ready institutions.
- Activity 12: Prepare those institutions which do not have accredited programs in closely aligned areas for the pre-accreditation visit year.
- Activity 13: Identify appropriate accreditation by program and provide visibility to academic institution administrations

Goal 6 (ED 1)	**Developing a combination of model programs, degrees, pedagogy, and curriculum including a 9-week middle school coding block; a technical certificate, certificate of proficiency, and associate of science in data science; and a Bachelor of Science in data science with minors or concentrations.**				
Objective 6.1.a: Raise awareness, interest, and enrollment in data science courses through targeted workshops, special events, career fairs, and related activities at campuses throughout Arkansas. **Objective 6.1.b:** Create a set of postsecondary programs of core courses with options for electives for a consistent set of Data Science Undergraduate degrees (e.g., Assoc. Degrees, 2+2 and "2, then 2"), Concentrations, and certificates					
Objective 6.1.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Host "Data Science for Arkansas" events at collaborating campuses				At least one event per campus complete	At least one event per campus complete
Objective 6.1.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Create the 5-year Plan to meet the Objective	Plan disseminated to stakeholders	Review 5-yr plan & update as needed	Review 5-yr plan & update as needed	Review 5-yr plan & update as needed	Complete 5-yr plan
Activity 2: Identify the level of involvement and timing by academic institutions within the State	Cohorts identified; all collaborators assigned	Begin Cohort 1	Begin Cohort 2 & Complete Cohort 1	Begin Cohort 3 & Complete Cohort 2	Complete Cohort 3
Activity 3: Review UAF and UCA Data Science Programs with the Teams	1 meeting complete		Update meeting complete		Update meeting complete
Activity 4: Convene workshops annually of engaged academic and government institutions to establish baseline	3 Workshops completed	3 Workshops completed	3 Workshops completed	3 Workshops completed	3 Workshops completed
Activity 5: Define Data Science Objectives and Outcomes base for defined degrees and certificates	Info disseminated to stakeholders		Updated Info disseminated to stakeholders		Updated Info disseminated to stakeholders

Activity 6: Define Data Science Courses Objectives, Learning Outcomes, and applicability to the defined degrees and certificates		Info disseminated to stakeholders		Updated Info disseminated to stakeholders	
Activity 7: Dissemination of developed program details with collaborating institutions, government, and industry partners	Info disseminated to stakeholders	Updated Info disseminated to stakeholders		Updated Info disseminated to stakeholders	
Activity 8: Ensure defined programs are in line with appropriate accrediting bodies	Identify "Wave 1" of accreditation candidates		Review "Wave 1" for accreditation readiness	Identify "Wave 2" of accreditation candidates	Review "Wave 2" for accreditation readiness
Activity 9: Prepare and submit program proposals of each type at each level for appropriate approval	Begin "Cohort 1" Proposal Preparation	Submit "Cohort 1" Proposals	Begin "Cohort 2" Proposal Preparation	Submit "Cohort 2" Proposals	Begin "Cohort 3" Proposal Preparation
Activity 10: Evaluate progress and iteratively improve for future programs as appropriate			Evaluation report disseminated to stakeholders		Evaluation report disseminated to stakeholders
Activity 14: Create and maintain clearing house for course materials	Create shared resources with UAF UCA existing materials and establish cataloging methodology	Add Cohort 1 developed materials	Update contributed materials	Add Cohort 2 developed materials	Add Cohort 3 developed materials

7. Workforce Development and Broadening Participation

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. DART will provide a number of programs to address this challenge, including professional development opportunities for K20 educators, undergraduate and graduate students.

7.1. Project Implementation

Goal	Goal Name	Lead(s)	Team Members
WD1	Provide K20 teacher and faculty opportunities for professional development spanning multiple disciplines	Fowler, Hillyer	-
WD2	Provide educational training opportunities inside and outside the classroom for students	Fowler, Hillyer	Schubert, Addison
WD3	Ensuring broad participation to impact the pipeline of data science skilled workers	Fowler, Hillyer	Schubert, Addison

Goal 7.1 Provide K20 teacher and faculty opportunities for professional development spanning multiple disciplines.

Objective 7.1.a. K-12 Teacher Professional Development: In partnership with the EAST Initiative, DART will train K12 teachers throughout the state focusing on two areas: student leadership and emerging technology integration. Teachers will learn methods to integrate these new technologies into their classrooms, empower students to hone their leadership skills, and will receive technology for their classrooms. Each topical area will consist of two (2) one-day training sessions. In addition, a series of webinars will be developed to share best practices, disseminate updated resources, and provide support for troubleshooting technology.

Objective 7.1.b. Education & Broadening Participation Seed Grants: DART will solicit proposals for project related mini-seed grants for education, outreach, and broadening participation. Eligible entities will include school districts, post-secondary institutions, educational service co-ops, non-profits, or other entities supporting data science and computer science education and outreach activities in Arkansas. The central office will manage the solicitations which will be posted on the AEDC website, DART website, and emailed through higher education channels. The proposals will be reviewed by the central office, and outside experts as needed.

Objective 7.1.c. Career Development Workshops: DART will hold regional and virtual workshops to enhance the research competitiveness of the state's faculty in data science and computing. Workshop topics will include science, grantsmanship, and commercialization. These will be tailored for students and early career faculty and will feature program officers from various agencies, entrepreneurs, and outside scientists. Three virtual workshops will be planned per year with possibility of face-to-face later in the project. Workshops will be free for

attendees and open to the Arkansas higher education community. The workshops will also be recorded, and recordings will be made available to all DART participants.

Goal 7.2: Provide educational training opportunities inside and outside the classroom for students.

Objective 7.2.a. Student Support at Participating Institutions: DART will 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. This will provide students with the opportunity to work on real research problems provided by industry and will increase retention and broaden participation in the data science pipeline. All student participants will be invited to DART meetings as outlined in the communication plan below. In addition, DART will host student forums every other month to provide updates and receive feedback from the students regarding their participation and reinforce their success in the project. All student participants will also be invited to present their research in the annual Poster Competition, which will be judged by faculty, EAB, and IAB members. The winning students will receive travel awards to major national and regional conferences.

Objective 7.2.b: Summer Internships: DART will work with industry partners, including the industry advisory committee, to facilitate at least 5 internships annually during Years 2-5 for student participants at companies in relevant sectors and research centers. DART will work with the host companies and students to evaluate and iteratively improve the recruitment and hosting process. Dr. Addison and Dr. Schubert will lead this initiative. An application process will be developed with the industry partners and students will be recruited from all DART participating institutions.

Objective 7.2.d. Arkansas Summer Research Institute: The ASRI will be hosted in partnership with the Arkansas School for Mathematics, Sciences, and the Arts (ASMSA). The ASRI is an intensive professional development experience for STEM students (seniors in high school up to senior undergrads). This 1-week event will be offered in 2 sessions each year and will provide workshops, panel discussions, and research training activities. The main goals of ASRI are to a) build a diverse support network of peers for undergrads in Arkansas and b) provide professional development to STEM undergraduates. The ASRI will build on previous successful iterations but with a modified focus on Dart related topics. Each year's program will be evaluated by the external evaluator. Participants will be added to ASRI alumni Facebook and LinkedIn groups to facilitate longitudinal tracking and communications.

Goal 7.3 Ensuring broad participation to impact the pipeline of data science skilled workers.

Objective 7.3.a. Summer Undergraduate Research Experiences for underserved students: DART will fund summer undergraduate research experiences (URE), for underserved students. Students will be recruited through established and successful campus-based programs like McNair and Arkansas Louis Stokes Alliance for Minority Participation (LSAMP), and by faculty at participating institutions. Faculty will apply for funds to host these students for 8 weeks, with a limit of $8,000 per award. Funds will support student stipends, housing, student-specific supplies, and in-state travel. Faculty are encouraged to host students who have completed the

ASRI and to send recruited SURE students to the ASRI. The applications will be reviewed and awarded by the central office.

Objective 7.3.b. Scholarships to ASRI: DART will provide scholarships and targeted outreach to recruit URM students to attend the ASRI. Students will be recruited at all Arkansas higher education campuses, on social media, through email, and hopefully in-person recruiting events. DART will also leverage relationships with LSAMP, the McNair Achievement Program, and other organizations that are connected with underserved students. Transportation stipends will be provided to these students as needed, as well as any technology or necessary supplies.

Objective 7.3.c. ACDS: DART will connect URM students to opportunities for internships, apprenticeships, jobs, and professional development through ACDS. Dr. Schubert and Dr. Addison will establish monthly meetings with Bill Yoder of ACDS where opportunities will be discussed. DART participants will be notified of opportunities in an email newsletter.

Summary of changes in 2023 strategic plan revision

The milestone related to inviting students to present their capstone projects at the annual meeting was removed. These students are invited to present at the monthly webinar series instead. Similarly, the milestone for inviting educators who complete the professional development workshops in partnership with EAST Initiative was removed. Finally, the activity and milestones for faculty training grants was removed. This was not being utilized and due to the delays in implementing new degree programs at two-year campuses, there was no faculty interest in participating in this activity.

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

Objective 7.1.a: Enable K12 teachers to integrate new Computer Science/Data Science technologies into their classrooms.
Objective 7.1.b: Education and broadening participation mini grants
Objective 7.1.c: Career development workshops

Objective 7.1.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Host one training session annually, issue technology kits to teacher participants		1 Training session complete, kits issued	1 Training session complete, kits issued	1 Training session complete, kits issued	1 Training session complete, kits issued
Activity 2: Host two support/training webinars annually		Two webinars completed	Two webinars completed	Two webinars completed	Two webinars completed
Activity 3: Establish platform for teachers to disseminate resources and troubleshoot			Platform implemented; past participants onboarded	New participants onboarded	New participants onboarded
Activity 4: Participate (booth or breakout) in EAST Initiative annual conference	1 Conference completed	1 Conference completed	1 Conference completed	1 Conference completed	1 Conference completed

Objective 7.1.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Fund seed mini-grants annually at $5,000 each	10 seed grants awarded	10 seed grants awarded	10 seed grants awarded	10 seed grants awarded	10 seed grants awarded
Activity 2: Recipients attend Annual All-Hands	1 awardee presentation at All Hands complete	1 awardee presentation at All Hands complete	1 awardee presentation at All Hands complete	1 awardee presentation at All Hands complete	1 awardee presentation at All Hands complete
Objective 7.1.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Host annual workshops on a variety of grantsmanship and entrepreneurship topics	3 Workshops completed	3 Workshops completed	3 Workshops completed	3 Workshops completed	3 Workshops completed

Goal 7.2 (WD)	**Provide educational training opportunities inside and outside the classroom for students.**				
Objective 7.2.a: Student Support at Participating Institutions: Support undergraduate research assistants during the fall and spring semesters at each participating primarily undergraduate institution, and graduate research assistantships at the academic research institutions. **Objective 7.2.b:** Summer Internships: Facilitate industry internships for student participants at companies in relevant sectors and research centers. **Objective 7.2.c:** Connect with other research thrusts to develop relevant research-based capstone projects **Objective 7.2.d:** ASRI- intensive data science and computing summer camps for undergraduates					
Objective 7.2.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Provide undergraduate research assistantships annually	15 UG supported	15 UG supported	15 UG supported	15 UG supported	15 UG supported
Activity 2: Provide graduate research assistantships annually	40 GA supported	40 GA supported	40 GA supported	40 GA supported	40 GA supported
Objective 7.2.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Identify internship opportunities for students at relevant companies		5 internships placed	5 internships placed	5 internships placed	5 internships placed
Activity 2: Follow up with hosting companies for feedback and evaluation		Develop intern and hosting company feedback and evaluation methodologies and instruments	Year 2 feedback and evaluation; iterative improvement for Year 3	Year 3 feedback and evaluation; iterative improvement for Year 4	Year 4 feedback and evaluation; iterative improvement for Year 5; Year 5 feedback and evaluation

Objective 7.2.d	Specific Milestones				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Host ASRI Annually and invite all DART undergrads	1 ASRI Complete	1 ASRI Complete	1 ASRI Complete	1 ASRI Complete	1 ASRI Complete
Activity 2: Evaluate and revise programming based on student and presenter feedback		Evaluation report disseminated to stakeholders	Evaluation report disseminated to stakeholders	Evaluation report disseminated to stakeholders	Evaluation report disseminated to stakeholders

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

Objective 7.3.a: Summer Undergraduate Research Experiences for underserved students: Fund summer undergraduate research experiences (URE), for underserved students
Objective 7.3.b: Scholarships for underserved students to the ASRI
Objective 7.3.c: Connecting students to opportunities through the ACDS

Objective 7.3.a	Specific Milestones				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Provide summer UREs to URM students annually	10 UG supported	10 UG supported	10 UG supported	10 UG supported	10 UG supported
Activity 2: Students participate in annual All-Hands meeting poster competition		1 poster competition complete	1 poster competition complete	1 poster competition complete	1 poster competition complete
Objective 7.3.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Provide scholarships/recruit students annually	20+ scholarships provided	20+ scholarships provided	20+ scholarships provided	20+ scholarships provided	20+ scholarships provided

Objective 7.3.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Co-host statewide workshops on Data Science topics	1+ workshop completed	1+ workshop completed	1+ workshop completed	1+ workshop completed	1+ workshop completed
Activity 2: Collaborate on Data Science apprenticeship programs- recruiting partners and developing curriculum	Develop apprentice and hosting company feedback and evaluation methodologies and instruments	Year 1 feedback and evaluation; iterative improvement for Year 2	Year 2 feedback and evaluation; iterative improvement for Year 3	Year 3 feedback and evaluation; iterative improvement for Year 4	Year 4 feedback and evaluation; iterative improvement for Year 5; Year 5 feedback and evaluation

8. Communication and Dissemination

The grand challenge of the communication and dissemination effort is to ensure that all DART participants are aware of their role and responsibilities to the project, and to make the public aware of DART and its success.

8.1. Implementation Plan

Goal	Goal Name	Lead(s)	Team Members
CD1	Maintain interproject communication to accomplish milestones and relay updates	Fowler, Hillyer	Ford, Cothren
CD2	Educate the public about DART accomplishments	Fowler, Hillyer	Ford, Cothren

Goal 8.1: Maintain interproject communication to accomplish milestones and relay updates.

Objective 8.1.a: Day to Day Communication: Daily project-related communication will take place mostly via email and GitLab. If during the first year the project faces challenges with these two platforms, other platforms like Slack will be explored. DART will also hold virtual office hours monthly or as needed where participants can drop in and ask any project related questions. The office hours will be staffed by representatives from the SSC and central office.

Objective 8.1.b.- Monthly Webinars & Component Meetings: DART will hold monthly webinars beginning in September 2020. Each webinar will have a short presentation followed by open discussion and questions. DART monthly webinars will be on the 3rd Wednesday of each month, in the same time period as the SSC monthly meeting. The topics will rotate between project management items like reporting and overviews of DART as a project, as well as specific research topics, important research results, and more. Each research component will also meet as a team monthly to communicate needs and progress. These meetings will mainly occur online via Zoom or Microsoft Teams and similar platforms.

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. The central office is responsible for the logistical planning of these two events. The retreat will take place in a central location and the All-Hands meeting location will be rotated among the regions of the state. Online participation will be facilitated for both meetings for participants who cannot join in person.

Goal 8.2: Educate the public about DART accomplishments.

Objective 8.2.a.- Maintain public-facing communication outlets to inform public about DART: DART will establish a public project website that contains general information about the project, important research results, contact people, a list of all participants, links to faculty websites, and other content that will be refreshed at least quarterly. The central office will also maintain the existing @arepscor Facebook, Twitter, and YouTube pages with DART related content. Blogs will be posted with important updates related to DART on the AEDC website.

Objective 8.2.b.- Campus Communications: A project-wide communications team composed of communications staff from each participating institution, including AEDC, will be created. The communications team will use uniform citations and branding for all project-related releases. AEDC will issue press releases and blog posts related to overall project success, special events, and seed grant opportunities. The communications team will work together to release other pertinent information like new grant awards, patents, publications, and other highlights from each campus.

Objective 8.2.c.- Project Data: Project data will be submitted by participants to the project's internal reporting system, ER Core. Mandatory NSF reporting data will be collected, as well as additional information like startup companies and other major accomplishments. Participants will be encouraged to enter data throughout the year into ER Core, and there will be a cutoff date before annual reporting each year for data submission for the report. The central office will establish and maintain this site.

Objective 8.2.d.- Technical dissemination channels: Project faculty will submit journal articles to scientific publications associated with data science and computing. Funds for travel stipends will be reserved to send students and faculty to national meetings related to data science and computer science research and education. Impacts and significant findings of research activities will be presented at these meetings. Relevant meetings include national meetings for professional societies and industry meetings. All outputs related to this will be reported in ER Core. See individual research component matrices for specific dissemination output goals.

Summary of changes in 2023 strategic plan revision

The activities related to the science journalism challenge have been removed. We could not get traction with this activity early in the project and made several attempts to salvage it, with no success. Considering the relationships established with communications teams on each campus and the progress made via social media, the website, and other outlets, it was determined that the time and labor required to launch the journalism challenge would not be productive. Also, the frequency of blog posts was removed to provide more flexibility and bandwidth for the central office team to carry out the other activities.

Goal 8.1	**Maintain interproject communication to accomplish milestones and relay updates**				
Objective 8.1.a: Day to Day Communication **Objective 8.1.b:** Monthly Webinars & Component Meetings **Objective 8.1.c:** Face-to-Face Meetings					
Objective 8.1.a	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Establish platform to maintain daily communication	Platform established and participants onboarded				
Objective 8.1.b	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Host DART topical webinars monthly	11 webinars complete	11 webinars complete	11 webinars complete	11 webinars complete	11 webinars complete
Activity 2: Host monthly component team meetings	11 meetings per component (6 components)	11 meetings per component (6 components)	11 meetings per component (6 components)	11 meetings per component (6 components)	11 meetings per component (6 components)
Objective 8.1.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Host annual All-Hands meeting & poster competition	1 All Hands & Poster Competition complete	1 All Hands & Poster Competition complete	1 All Hands & Poster Competition complete	1 All Hands & Poster Competition complete	1 All Hands & Poster Competition complete
Activity 2: Host annual retreat for faculty and grad students	1 Retreat completed	1 Retreat completed	1 Retreat completed	1 Retreat completed	1 Retreat completed

Goal 8.2	Educate the public about DART accomplishments

Objective 8.2.a: Maintain public-facing communication outlets to inform public about DART
Objective 8.2.b: Campus Communications
Objective 8.2.c: Project Data
Objective 8.2.d: Technical dissemination channels

Objective 8.2.a	Specific Milestones				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Establish project website	Project website published				
Activity 2: Maintain project website and refresh content		Content posted	Content posted	Content posted	Content posted
Activity 3: Maintain @arepscor Facebook, Twitter, and YouTube channels and refresh DART content frequently	Following increased by 10%	Following increased by 10%	Following increased by 10%	Following increased by 10%	Following increased by 10%

Objective 8.2.b	Specific Milestones				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Establish listserv and group of communications reps from each participating campus	Committee formed; first meeting complete				
Activity 2: Hold annual check-in meetings to ensure proper citation of project and related messaging and disseminate project updates		1 meeting complete	1 meeting complete		
Objective 8.2.c	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Establish DART ER Core Site	ER Core Site published & accessible				

Activity 2: Maintain DART ER Core site and provide annual training to participants	Participants onboarded; 3 training webinars complete	3 training webinars complete	3 training webinars complete	3 training webinars complete	3 training webinars complete
Objective 8.2.d	**Specific Milestones**				
	Year 1	Year 2	Year 3	Year 4	Year 5
Activity 1: Presenting at national conferences / professional societies		2 presentations complete	4 presentations complete	4 presentations complete	5 presentations complete
Activity 2: Publications		1 publication complete	2 publications complete	3 publications complete	4 presentations complete
Activity 3: Statewide Workshops for Cohorts and Waves	2 Workshops complete	2 Workshops complete	2 Workshops complete	1 Workshops complete	1 Workshops complete

9. Appendix A: Project SWOT Table

Likelihood of Occurrence	Impact	Challenge/Risk	Mitigation Steps
High	High	Faculty overwhelmed by new course delivery and less time for research	Reduced meeting and internal reporting requirements in year 1 to allow focus on research; AEDC follow-up with schools to ensure adequate release time
High	High	Enabling technologies are advancing rapidly so teams will need to stay ahead of current trends	Stay abreast of industry advancements (and keep an eye on academic work)
High	High	Lack of human infrastructure to teach Data Science courses at collaborating campuses	Use the 'Oxford Model' of tutorials - one-on-one instructors, remotely working with students
High	Medium	Personnel turnover, both faculty and students	Avoid single points of failure, always have at least two people up-to-speed on each critical task. Need good reports and documentation of everyone's work and progress. Work on keeping people happy, and feeling engaged. Make all feel like they are part of something larger.
High	Medium	Matching human perception with machine learning and auto-annotation techniques remain a huge challenge (e.g. face recognition vs. mood recognition).	Use a reduced data set along with proper data compression and management techniques
Medium	High	Difficulty in recruiting international, and possibly full-time domestic, GRA students	Work with graduate school committees to ensure appropriate outcomes
Medium	High	Hiring freezes	Work with school administrations to give priority openings supporting DART
Medium	High	Lack of funded management roles in DART	Work with campuses to identify persons will to help in-kind as management resources
Medium	High	Lack of diversity	Strive to hire more diversity when openings need to be filled
Medium	High	Some research teams have not worked together before which may lead to team management challenges	Regular meetings, joint team assignments, collaborations

Likelihood of Occurrence	Impact	Challenge/Risk	Mitigation Steps
Medium	High	Rapidly changing and very competitive subfields of privacy, security, and fairness in AI	Stay on and ahead of current trends and adapt research activities. Attend virtual conferences early, followed by in-person attendance where possible. Collaborate with researchers outside the state.
Low	High	SM4 requires broad collaboration and use of outputs from SM3, Research Theme DC, and Research Theme LP in order to be successful. Specifically, the novel vehicle routing models and solution approaches require novel inputs from the various big data methodologies outlined in the overall project plan.	SM4 lead will proactively communicate with PIs from SM3, DC and LP beginning in Year 1 and throughout the project duration. Establish liaison relationships across these Goals.
Medium	Medium	Lack of federated identity across participating institutions. Possible with UA system but not others. Creates friction trying to data/code resources across Globus and GitLab	Regular meetings with the 'right people' across the Institutes to make sure all know what's going on.
Medium	Medium	Tradeoff of conducting cutting edge research on social awareness and implementing known algorithms, software, platform to address social concerns	Discuss with project team to balance intellectual merit and broader impacts
Medium	Medium	In SM, minimizing impacts of fake and misleading information dissemination is challenging.	To reduce the impact, a hybrid model of assessing collective argument credibility will be developed.
Medium	Medium	Improving the quality of discourse is challenging.	Mining unstructured data from our cyber argumentation and social networking platform to analyze its various characterizations, such as user participation, community structure, user's influence, people-pleasers or devil's advocates, and diversity aware social connection recommendation.
Medium	Medium	Social media data collection is often challenged by API rate limits. This may impact our current data collection methodology.	Data acquisition cost has been budgeted in the SM research theme to obtain social media data. Also, modifying the frequency of data collection and using a parallel/distributed data collection architecture will help mitigate any API rate limit related issues.

Likelihood of Occurrence	Impact	Challenge/Risk	Mitigation Steps
Medium	Medium	Maintaining privacy of the online identities collected from social media data.	Before conducting the research, PI will submit the study protocol to university's Institutional Review Board (IRB) for ethics and safety approval. PI has received approval from his university's IRB for other studies involving similar data collection methodologies from social media platforms. PI has completed the CITI program courses on Social and Behavioral Responsible Conduct of Research, Human Research, and Social and Behavioral Research Investigators and Key Personnel that expire on May 8, 2023.
Medium	Medium	Indexing and annotation mechanisms may not match the level and nature of those required by advanced learning techniques.	Multi-modal annotation mechanisms that combine existing text-based data will augment the techniques developed to improve the accuracy and performance.
Medium	Medium	Obtaining transportation infrastructure data from a variety of data sources (e.g., social media, satellite imagery, traffic camera videos) for the same disaster event, so that the data can be fused to predict infrastructure status.	Purchase of social media data is part of the project plan and budget for Research Theme SM; will pursue the procurement of other data types from transportation and emergency management officials.
Medium	Low	Considering that the project is ambitious (it strives for a high level of quality, verifiability, science and productivity for a short period of realization) while depending on many external factors, it is possible to deviate from the schedule.	Elaboration of a risk assessment for the delay in the schedule, followed by plans for their suppression. Establishing a project implementation monitoring system on a weekly basis. Risk assessment will continue during project's course and mitigation strategies will be developed.
Medium	Low	Computation and storage requirements of the techniques and data sets may go beyond the capabilities of the systems and equipment used.	Learning and indexing mechanisms will initially use a hybrid approach to adapt algorithms to user expectations and ground truth gradually.

10. Appendix B: Project and Theme-specific Logic Models

Project-wide, long-term Outcomes (Impacts):

-- Research contributions will improve the learning and prediction of data in a spectrum of applications including commerce, cybersecurity, disaster and emergency management, energy, environment, healthcare, retail, and transportation

-- Research outputs will generate interest in data science and help engage, encourage, and recruit a broad spectrum of learners as well as researchers

-- Arkansas should see a growth in research and education initiatives in data science creating a large number of diverse pipelines for data scientists, engineers, and technicians

-- DART will 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

2023 Revision: Upon acceptance of these requested revisions to the strategic plan, the project's external evaluator will update the logic model and evaluation plan accordingly.

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
Research Theme 1: Cyber Infrastructure	**Goal 1 (CI1)** **Staff**: Cothren, Prior, Springer, Chaffin, Tarbox, Deaton, DuRousseau, Pummill, Merrifield **Partnerships**: Great Plains Network	Objective 1.1.a: Establish the Arkansas Research Computing Collaborative (ARCC) Objective 1.1.b: Upgrade cluster for data science research activity and integrate with existing resources	**Hardware and Software Infrastructure:** -- Install, configure, and make available data science nodes on Pinnacle Portal -- ScienceDMZ at UA and UAMS/UALR -- 100Gb connection between ScienceDMZs. -- Establish dedicated DART GitLab repository -- Setup Globus data management services to point at DART storage arrays **Documentation and User Guides:** -- Create a technical management document defining organizational	-- 100% increase in active accounts on Pinnacle and Grace -- 100% increase in overall use measured in gigflops -- Code developed by DART researchers is shared via GitLab repository linked to public GitHub -- Sharing of large data sets among individuals, institutions, and HPC clusters.	-- enhanced academic collaboration across Arkansas campuses measured by increased publications using CI resources -- enhanced collaboration between industry and academia measured by the number of such projects that use CI resources -- trusted data sharing between collaborating partners measured by the number of recorded data transfers and the total number of TB transferred

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		Objective 1.1.c: Establish a Little Rock (UAMS, UALR) ScienceDMZ and high-speed connection with UAMS	structure, roles, and responsibilities of ARCC for personnel at participating campuses -- Amend existing MOU for ARCC expansion -- UAF and UAMS will create CI Plans to support DART (1 x UAF, 1 x UAMS); these CI Plans will serve as templates for other Institutions -- Create and publish document outlining GitLab user guidelines and minimum standard for code repository **Workshops, demonstrations, and trainings:** -- Two (2) online workshops per year for onboarding to ARP resources in YR2-5 -- Five (5) online software carpentry workshops per year in YR2-5 focusing on developing and sharing code and data; data		
		Objective 1.1.d: Establish a data and code sharing environment (GitLab and Globus)			

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		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)	science programming; and data management -- Train and certify two (2) new software carpentry instructors (across the jurisdiction) per year in YR2-5 **Applications and platforms:** -- Create one (1) distributed computing testbed for HDFS, Apache Spark, others (DC) - -Create four (4) spatiotemporal testbeds for (CI/DC/SM/LP)		
	Goal 2 (CI2) **Staff:** Springer, Conde, Huff, Milanova **Equipment:** As determined by need and existing capability	Objective 1.2.a: Investigate state-of-the-art visualization solutions	**Workshops, demonstrations, and trainings:** -- One (1) online workshops per year for advanced visualization in YR2-5		

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		Objective 1.2.b: Define domain-specific integration of visualization solutions Objective 1.2.c: Introduce/integrate visualization for shared test beds	**Publications, presentations, and reports:** -- Three (3) presentations, reports, or other publications: 1 in YR1 and 2 in YR 2 **Applications and platforms:** -- Develop one (1) visualization solution for each research theme, including CI (5 total) -- Integrate one (1) visualization into existing testbed for each research theme (4 total)		
Research Theme 2: Data Life Cycle and Curation	**Goal 2.1 (DC1)** **Staff:** Talburt, Cothren, Liao, Liu, Rainwater, Tudoreanu, Ussery, Wang, Xu, Yang **Partnerships:** UAMS, UAF. UALR **Facilities:** Campus offices, labs, IT Services and Networking	Objective 2.1.a: Automate Reference Clustering / Automate Data Quality Assessment Objective 2.1.b: Automate Data Cleansing	**Workshops, demonstrations, and trainings:** -- Five (5) conference workshops **Publications, presentations, and reports:** -- 14 research publications describing	-- Citations of papers and reports -- DART Git repository accesses, downloads, and branches show acceptance and use of algorithms within DART -- Public Git repository accesses, downloads, and branches show adoption of algorithms in the data	-- Use of HDFS as a platform for data curation increases in industry -- The balance between manual and automated data curation, as evidenced by industry and academic articles, tilts toward unsupervised automation in the data

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
	Equipment: Local Resources **Infrastructure:** Globus File Sharing, GitHub	Objective 2.1.c: Automate Data Integration	new methods and processes -- 11 journal and conference publications -- 16 presentation -- 5 PhD dissertations -- 2 potential patents and business incubation **Datasets and algorithms:** -- Nine (9) Mathematical formulations and algorithms for optimal collaborative data collection and cleansing. -- Seven (7) Mathematical formulations, algorithms and statistical methods for need- and prediction-based data collection, and scalable decision making	science community -- Data science programs in the state adopt some new methodologies and algorithms in coursework	curation process -- Genome and proteome quality scores become commonly accepted standards as evidenced in common repositories of that data. -- Good practice, as evidenced in government and industry standards, prohibit sending data without encapsulating a complete explanation of it syntax, semantics, and data quality requirements
	Goal 2.2 (DC2) **Staff:** Talburt, Wang, Tudoreanu, Pierce, Liu, Rainwater **Partnerships:** UAMS, UAF. UALR **Facilities:** Campus offices, labs, IT Services and Networking **Equipment:** Local	Objective 2.2: Build a POC for Positive Data Control (PDC)	**Workshops, demonstrations, and trainings:** -- Three (3) demonstrations of data exchange between PDC Systems	-- Citations of papers and reports -- DART Git repository accesses, downloads, and branches show acceptance and use of algorithms within DART -- Public Git repository accesses, downloads, and branches show adoption of algorithms in the data	

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
	Resources **Infrastructure:** Globus File Sharing, GitHub		**Publications, presentations, and reports:** -- Three (3) conference presentations -- Three (3) research papers **Applications and platforms:** -- One (1) working prototype of a positive data control system -- One (1) novel design of a system to control data access and movement in synchronization with a data catalog and data governance standards	science community -- Read/write firewalls around enterprise data mirror "external connection firewalls" around the enterprise intranet.	
	Goal 2.3 (DC3) **Staff:** Ussery, Byrum, Jun, Yang, Liu, Rainwater **Partnerships:** UAMS, UAF, UALR **Facilities:** UAMS/UA HPC infrastructure **Equipment:** EMC object store on the UAMS HPC	Objective 2.3.a: Standardize pipelines for genome and proteome storage, retrieval, and visualization	**Workshops, demonstrations, and trainings:** -- Host one (1) workshop on how to use these standardized pipelines **Publications, presentations, and reports:**	-- Citations of papers and reports -- DART Git repository accesses, downloads, and branches show acceptance and use of algorithms within DART -- Public Git repository accesses, downloads, and branches show adoption	

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
	Infrastructure: Globus File Sharing, GitHub	Objective 2.3.b: Automate quality scores for biological sequence data Objective 2.3.c: Apply machine learning methods to systems biology	-- Five (5) papers published on standardized pipelines for genomics and proteomics. -- Five (5) conference presentations -- Five (5) Journal publications -- Two (2) PhD dissertations **Datasets and algorithms:** -- At least one (1) standardized database with genomics and proteomics quality scores shared with DART researchers -- One (1) algorithm (code with associated training and testing data)	of algorithms in the data science community--Pipelines for storage and retrieval of millions of genomes and proteomes are possible. -- Visualizing massive amounts of biological sequence data is possible -- New data quality dimensions and metrics are in use.	
Research Theme 3: Social Awareness	**Goal 3.1 (SA1)** **Staff**: Xintao Wu, Qinghua Li, Anna Zajiciek	Objective 3.1.a: Identify potential vulnerabilities of deep learning algorithms Objective 3.1.b: Develop a universal threat- and privacy-aware deep learning framework	**Workshops, demonstrations, and trainings:** -- One (1) tutorial given at major AI conference **Publications, presentations, and reports:** -- Two (2) conference papers -- One (1) journal paper	-- Citations of papers and reports -- DART Git repository accesses, downloads, and branches show acceptance and use of algorithms within DART -- Public Git repository accesses, downloads, and branches show adoption of algorithms in the data science community	-- Public debate and policy, as evidenced in press reporting, government regulations, and company policy discusses the methods, algorithms, and findings related to social awareness -- Channels for collecting and publishing personal

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		Objective 3.1.c: Conduct comprehensive evaluations of the proposed framework and models	-- One (1) thesis -- One (1) proposal		information and sources of breaches are reported and adopted by the general public -- Privacy preserving data analytics in genomics and health data analytics are evident. -- Influence policies for developing and adopting cryptography based privacy protection models and change the practice of insufficient trust evaluation/enforcement for machine learning
	Goal 3.2 (SA2) **Staff**: Chenyi Hu, Ningning Wu, Xintao Wu	Objective 3.2.a: Improve crowdsourcing data quality with considerations of uncertainty	**Publications, presentations, and reports:** -- Two (2) conference papers -- One (1) journal paper -- One (1) thesis -- One (1) proposal		
		Objective 3.2.b: Enhance available inference and learning models with novel algorithms for improved effectiveness and efficiency			
		Objective 3.2.c: Verify and validate the robustness and trustworthiness of information from crowdsourcing data			
	Goal 3.3 (SA3) **Staff:** Ningning Wu, Qinhua Li, Chenyi Hu, Xintao Wu	Objective 3.3.a: Investigate on personal identifying information and their privacy issues	**Publications, presentations, and reports:** -- Two (2) conference papers -- One (1) journal paper -- One (1) thesis -- One (1) proposal		
		Objective 3.3.b: Investigate appropriate multimodal deep learning techniques to identify discriminative and stigmatizing information			

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		Objective 3.3.c: Develop a user-centric privacy monitoring and protection framework			
	Goal 3.4 (SA4) **Staff:** Lu Zhang, Xintao Wu, Zhenghui Sha, Anna Zajicek	Objective 3.4.a: Explore deep learning-based techniques to detect cross-media discrimination.	**Workshops, demonstrations, and trainings:** -- One (1) tutorial given at major AI conference		
		Objective 3.4.b: Design generative adversarial models to remove cross-media discrimination.	**Publications, presentations, and reports:** -- Two (2) conference papers -- One (1) journal paper -- One (1) thesis -- One (1) proposal		
		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.	**Datasets and algorithms:** -- One (1) algorithm (code with associated training and testing data)		

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
	Goal 3.5 (SA5) **Staff:** Zhenghui Sha, Lu Zhang, Xintao 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	**Publications, presentations, and reports:** -- Two (2) conference papers -- One (1) journal paper -- One (1) thesis -- One (1) proposal **Datasets and algorithms:** -- One (1) algorithm (code with associated training and testing data)		
		Objective 3.5.b: Network-based modeling of customer preference incorporating marketing parameters			
		Objective 3.5.c: Design of marketing strategies with fairness consideration and validate the approach			
	Goal 3.6 (SA6) **Staff:** Huang, Li, Mary, Ussery, CI ARP team	Objective 3.6.a: Design and develop machine learning algorithms and software, and advanced security and privacy technologies, for privacy-preserving data analytics.	**Publications, presentations, and reports:** -- Two (2) conference papers -- One (1) journal paper -- One (1) thesis -- One (1) proposal		

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		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.	**Datasets and algorithms:** -- One (1) algorithm (code with associated training and testing data)		
		Objective 3.6.c: Test the algorithms and technologies to work with a wide range of data types and high-dimensional heterogeneous data sources; Develop and deploy bioinformatics workflows into the private cloud environment, the Arkansas Research Platform ARP.			
	Goal 3.7 (SA7)Staff: Qinghua Li; Xiuzhen Huang; Ningning Wu	Objective 3.7.a: Develop privacy-preserving federated learning methods through combining cryptography techniques and privacy models	**Publications, presentations, and reports:**-- Two (2) conference papers-- One (1) journal paper-- One (1) thesis-- One (1) proposal		

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		Objective 3.7.b: Explore how to protect the privacy of classification input data from the server hosting machine learning models	**Datasets and algorithms:**-- One (1) algorithm (code with associated training and testing data)		
		Objective 3.7.c: Assess/Protect the trustworthiness of training data and machine learning models			
Research Theme 4: Social Media and Networks	**Goal 4.1 (SM1)** **Staff:** Zhan, Adams, S. Yang	Objective 4.1.a: Develop a cyber discourse social network platform	**Publications, presentations, and reports:** -- Five (5) peer-reviewed journal and/or conference papers (articles) **Assessments, questionnaires, and surveys:** -- One (1) IRB-approved questionnaire for collecting discourse data **Applications and platforms:** -- One (1) cyber discourse social network platform	-- Cyber-discourse platform sees use within and outside of DART -- DART Git repository accesses, downloads, and branches show acceptance and use of algorithms within DART -- Public Git repository accesses, downloads, and branches show adoption of algorithms in the data science community	-- Cyber discourse platform is used for social media and network discourse data collection and analysis --Policies and public debate consider deviant cyber campaigns --New studies in this research discipline cite datasets provided by this research -- Perceptual information is considered in the development of new smart technologies designed to influence social and ethical
		Objective 4.1.b: Collect data using the developed cyber discourse social network platform			

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		Objective 4.1.c: Develop natural language processing algorithms to analyze discourse data collected by the platform as well as existing data	**Datasets and algorithms:** -- Two (2) advanced natural language algorithms (code with associated training and testing data)		behavior --Realtime, disaster-relevant data posted to social platforms and collected from various imagery sources see incorporation into disaster response planning.
	Goal 4.2 (SM2) **Staff:** Agarwal, Trudeau, Zhan, Milburn, Dagtas **Equipment:** COSMOS Data Servers **Infrastructure:** Social media data acquisition license	Objective 4.2.a: Characterize online information environment (OIE) Objective 4.2.b: Develop socio-computational models to identify key actors and key groups of actors Objective 4.2.c: Study tactics, techniques, and procedures (TTPs) of deviant cyber campaigns Objective 4.2.d: Develop socio-computational models to measure power of a cyber campaign	**Publications, presentations, and reports:** -- One (1) taxonomy -- One (1) journal paper -- Two (2) conference presentations **Applications and Platforms**: -- Socio-computational models for OIE and TTP implemented in a web-based application **Datasets and algorithms**: -- Two (2) socio-economic models and associated datasets	-- Citations of papers, presentations, and reports -- Socio-computational platform sees use within and outside of DART -- DART Git repository accesses, downloads, and branches show acceptance and use of algorithms within DART -- Public Git repository accesses, downloads, and branches show adoption of algorithms in the data science community -- Findings are incorporated in courses taught by the PI imparting new skills to analyze social media and social networks. -- Datasets are accessed by academics and industry under NSF guidelines and terms of	

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
				service agreement with social media companies	
	Goal 4.3 (SM3)Staff: Dagtas, Trudeau, Milburn**Partnerships:** UALR, UA Law School	Objective 4.3.a: Develop multimedia indexing methods for social media data Objective 4.3.b: Design and implement deep learning methods for multimedia data Objective 4.3.c: Build Integrated smart applications based on unstructured multimedia data	**Publications, Presentations, and Reports**-- Two (2) journal articles-- One (1) conference paper**Datasets and Algorithms:**-- Three (3) algorithms ---- 1 x Indexing---- 1 x deep learning for multimedia data---- 1 x integrated smart applications	-- Citations of papers, presentations, and reports-- Socio-computational platform sees use within and outside of DART-- DART Git repository accesses, downloads, and branches show acceptance and use of algorithms within DART-- Public Git repository accesses, downloads, and branches show adoption of algorithms in the data science community-- Findings are incorporated in courses taught by the PI imparting new skills to analyze multimedia data and other unstructured data	
	Goal 4.4 (SM4) **Staff:** Milburn, Dagtas, Liao, Zhan, Cothren, Ussery, Talburt,	Objective 4.4.a: Extract and index content describing transportation infrastructure status from social platforms	**Publications, presentations, reports:** -- Three (3) journal articles -- Two (2) conference papers	-- Citations of papers, presentations, and reports -- Socio-computational platform sees use within and outside of DART	

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
	Nachtmann, Rainwater, Celebi, Karim	Objective 4.4.b: Fuse data from social platforms describing transportation infrastructure status with other data sources Objective 4.4.c: Assess credibility of data inputs from Objectives 4.4.a and 4.4.b Objective 4.4.d: Develop routing algorithms that use inputs from Objectives 4.4.a-4.4.c to support routing for disaster response	-- Two (2) case studies of natural disasters scenerios **Applications and Platforms:** -- GIS routing platform infromed with social media feeds **Datasets and Alogirthms:** -- One (1) schema for mapping each datum to a probability describing its credibility -- Four (4) algorithms	-- DART Git repository accesses, downloads, and branches show acceptance and use of algorithms within DART -- Public Git repository accesses, downloads, and branches show adoption of algorithms in the data science community -- GIS platform is used by other researchers to better understand how social media input can be used dynamically for better routing	
Research Theme 5: Learning and Prediction	**Goal 5.1 (LP1)** **Staff:** Liu, 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	**Publications, presentations, reports:** -- Two (2) manuscripts submitted for publication -- Two (2) student theses or dissertations proposed -- Two (2) student theses or dissertations defended	New statistical and machine learning paradigms; Formal comparison of ensemble learning with traditional methods.	-- Documented understanding of complex/dynamic relationship between event processes and covariates information; Proposed interpretable intervention, control and optimization actions

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		incorporated into a tree-based method. Objective 5.1.b: Create the Gradient Boosting method for Recurrent Event Analytics, which integrates the boost trees with classical statistical methods allows dynamic feature information. Objective 5.1.c: Perform comparison study between the methodologies above and identify future research directions	-- One (1) submitted research proposals		related to recurrent event analysis--Promote acceptance of AI algorithms in on-the-edge devices and low-cost computers --Teach new lecture content in Deep Learning; -- Demonstrate reduced computational complexity and improve performance accuracy in AI -- Create research environment in higher education that trains students to experience and execute real-world AI applications -- Intervention, control and optimization policies via improved feature utilization; New statistical learning-based research practices; Research findings shared with industry to support practice of regular and rigorous data collection -- Motivate use of the integrated framework to improve decision making in business problems
	Goal 5.2 (LP2) **Staff:** Rainwater, Liu	Objective 5.2.a: Develop methodology integrating the marked temporal point process (MTPP) with long short-term memory networks (LSTM) Objective 5.2.b: Create scalable implementation of MTTP/LSTM approach applicable to real-world data analysis scenario	**Publications, presentations, reports:** -- One (1) conference paper -- Two (2) conference presentations -- Two (2) journal publications -- One (1) case study **Workshops, demonstrations, and trainings:**	Documented improved integration of large-scale computing knowledge with marked temporal point process (MTPP) and deep learning; Curation of discrete marked temporal point process (MTPP) data set; Benchmark results comparing marked temporal point process	

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		Objective 5.2.c: Evaluate and assess MTTP/LSTM approach on real-world discrete data sets	-- One (1) graduate seminar -- One (1) industry workshop	(MTPP) against other deep learning approaches	
	Goal 5.3 (LP3) **Staff:** Celebi, Kursun, Luu, Kim, Karim	Objective 5.3.a: Extract explanatory features from Deep Network Objective 5.3.b: Address high dimensionality issues in Deep Reinforcement Learning (DRL) using algebraic and topological methods Objective 5.3.c: Designing a novel rewarding model, and addressing interpretability issues in DRL	**Publications, presentations, reports:** -- Six (6) conference papers -- Two (2) Journal publication -- Two (2) Master's theses; **Workshops, demonstrations, and trainings:** -- One (1) workshop -- One (1) teaching module -- One (1) Special topics undergraduate class offering **Datasets and Alogrithms:** -- One (1) dataset	Identification of explanatory features in Deep Networks; Demonstrated ability to design improved reward function and explain causal relationships in DRL; Training of graduate and undergraduate students	
	Goal 5.4 (LP4) **Staff:** Khoa Luu, Ngan Le	Objective 5.4.a: Create Novel Deep Learning Networks Executable with Reduced Computational Resources and Assess Performance	**Publications, presentations, reports:** -- One (1) journal article -- Two (2) conference papers -- One (1) invited	Train graduate students and honors undergraduate students; Disseminate new low-computational cost Deep Learning algorithms,	

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		Objective 5.4.b: Address Low-cost Deep Learning Algorithmic Analysis and Challenges	presentation at an Arkansas Institution -- One (1) invited presentation elsewhere **Workshops, demonstrations, and trainings:** -- One (1) conference workshop -- One (1) tutorial -- One (1) teaching module	frameworks, and supportive libraries; Contribute new dimensions to analyze the computational time, resource consumption, and performance in AI algorithms	
		Objective 5.4.c: Explore Low-cost Deep Learning Applications in Natural Images and Medical Images			
	Goal 5.5 (LP5) **Staff:** Zhang, Nachtmann	Objective 5.5.a: Design advanced feature engineering techniques for high-dimensional temporal data	**Publications, presentations, reports:** -- Three (3) journal publications -- Three (3) conference presentations -- One (1) doctoral dissertation	Improve knowledge in feature engineering with transaction data; Establish abilities to incorporate feature engineering in prediction; Employ the skills of the integrated framework to solve business problems	
		Objective 5.5.b: Create an improved prediction and decision-making framework incorporating feature engineering with health transaction data			
		Objective 5.5.c: Employ and validate the new framework for prediction and decision making with business transaction data			

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
Research Theme 6: Education	**Goal 6 (ED)Staff:** Fowler, Schubert, Addison	Objective 6.1.a: Middle school coding block	9 week curriculum developed in consideration of standards, piloted in at least 6 schools		Students successfully complete coding block and increase in enrollment in computer science courses at high school level.
		Objective 6.1.b: Create model postsecondary programs in 3 phases at collaborating institutions	Curriculum implemented at three cohorts of 3+ campuses each,		Curriculum has been implemented on two- and four-year campuses. Programs are beginning to produce graduates who have completed the new curriculum.
Research Theme 7: Workforce Development	**Goal 7 (WD)** **Staff:** Fowler, Schubert, Addison	Objective 7.1.a: K12 Teacher PD	4 training workshops and 8 webinars, platform for resource sharing and troubleshooting		400+ K12 teacher participants confidently teach tech and leadership skills, 4+ students present at EAST conference
		Objective 7.1.b: Education & Broadening Participation Seed Grants	TBD	TBD	TBD
		Objective 7.1.c: Provide funding for faculty at collaborating institutions to learn new skills and tools in data science	50 faculty trained	Developing human infrastructure needed to achieve implementation	Each collaborating institution has faculty resources needed to teach DS curriculum
		Objective 7.1.d: Career Development Workshops	15 workshops hosted		Increased research competitiveness with progress in proposals submitted and collaborative projects

Research Theme	Input(s)	Objectives	Output(s)	Short-term Outcomes	Medium Term Outcomes
		Objective 7.2.a: Student Support at Participating Institutions	75+ undergraduates supported and trained, 75+ graduate students supported and trained	Network of student participants established	Students with rich educational experiences matriculating and getting jobs in relevant industries
		Objective 7.2.b: Summer Internships	20+ internships placed	Students are aware of the opportunities and there is completion for placements.	Interns graduating and entering workforce
		Objective 7.2.c: Research-based capstone projects	9+ Capstone project documentation disseminated to ARHE community	Research faculty are contributing a variety of potential capstone projects. Students have capstone opportunities relevant to their chosen specialization in data science.	Students are graduating with meaningful and relevant capstone experiences.
		Objective 7.2.d: Arkansas Summer Research Institute	5 ASRI workshops hosted, 200+ students trained	Positive evaluations and feedback from student participants	Students with enhanced DS/CS skills matriculating through colleges and universities