External Advisory Board Report

Arkansas NSF-EPSCoR RII Track-1 project, Year 3

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

6 July 2023

Contributing External Advisory Board Members:

Dr. Scotty Strachan, Principal Research Engineer, Nevada System of Higher Education
Dr. James Caverlee, Professor of Computer Science and Engineering, Texas A&M University
Dr. Cihan Varol, Professor of Computer Science, Sam Houston State University
Mr. Kash Mehdi, Vice President of Growth, DataGalaxy
Dr. Hongmei Jiang, Professor of Statistics, Northwestern University
Dr. Srinivasan Parthasarathy, Professor of Computer Science, Ohio State University
Dr. Weisong Shi, Professor and Chair of Computer and Information Sciences, University of Delaware
Dr. Huan Liu, Professor of Computer Science and Engineering, Arizona State University

Introduction

The External Advisory Board (EAB) for the DART project has prepared the following review for Year 3 of the DART NSF-EPSCoR RII Track-1 project. On April 19-21 2023, EAB members Strachan, Caverlee, Varol, Mehdi, Jiang, and Barnhouse attended the in-person DART Year 3 all-hands meeting in Fayetteville, Arkansas, which resulted in the first draft of this document. In addition, the EAB was provided with a number of project documents that provided status and context for our Year 3 review:

- DART EAB Charge
- AR 1946391 Strategic Plan 2022 Update
- Year 3 EAB Overview Presentation
- Year 3 Narrative Report Draft
- DART Evaluation Report Year 2
- Responses to 2022 RSV Recommendations
- 1946391 vRSV Presentations
- DART Year 2 EAB Report
- DART Retreat Report 2022
- DART BP Plan
- UAMS Data Governance

This document collection was also used by EAB members who were not able to participate in the in-person meeting to still contribute to this review. Our review below consists of high-level summaries of accomplishments and concerns, with discussion and recommendations broken out by project components.

Overview & General Feedback

Overall, the DART project has put itself in a good place for the end of Year 3, with notable research outputs in terms of publications and spin-off proposal awards. Project leadership has done an admirable job in adapting to unforeseen circumstances as a result of the COVID pandemic, including challenges in institutional administrative support and personnel turnover. The EAB notes that not only has project leadership experienced turnover at the highest level (PI/PD), but that the state EPSCoR office is not provisioned with dedicated administrative support to the same level as other EPSCoR jurisdictions. These issues, combined with the abnormally large project scale (number of supported faculty and students) and associated task granularity have undoubtedly combined to create a “perfect storm” of challenges for all involved.

The EAB agrees with the current strategic plan revisions and intent that the final two years of the project be focused on raising awareness of data science in education, laying the foundations for coordinated technology infrastructure, and refining capabilities, practices, and policies around data governance.

Summary of Accomplishments

The potential of transformative research-driven educational infrastructure and demonstrated student diversity and talent across DART are to be commended and are exciting for the state. The amount of additional grant dollars awarded in DART-driven research topics is impressive after only three years, and the publication output is excellent. It is clear that research infrastructure and research-driven educational infrastructures being improved by DART are a combination of materials, human capital, and processes that have high probability to support and advance Arkansas’ objectives in the areas of data science and related workforce development.

Summary of Concerns

Key challenges remain for DART leadership to evaluate and prioritize as they consider the final two years of project activities.

These challenges are articulated in detail in the component-specific sections below, but can be generally categorized as:

(1) cross-component activity towards true multi-disciplinary data science;
(2) friction-free cyberinfrastructure for effective statewide data science;
(3) project-wide data/code sharing and security processes;
(4) industry-aligned measures of workforce/education success and skills development;
(5) clear assessment protocols for education and outreach activities.

Coordinated Cyberinfrastructure Component

General Feedback: As articulated in the DART Vision and Mission statements, the Cyberinfrastructure (CI) Component is situated as the core research infrastructure being developed and improved in the DART project, with significant dependencies propagating across other project Components and their objectives.

DART has identified major challenges as being (generalized from project objectives and activities):

(1) interoperability between existing HPC and data storage systems at main AR campuses;
(2) universal, secure access to CI for DART personnel;
(3) coordinated CI architecture and development for the state;
(4) scaling up computational research and data science on both local and cloud CI systems;
(5) developing functional restricted data environments.

These challenges are common to the national research computing and data community, and so DART identifying and tackling these issues is consistent with trends and leading concerns nationwide. It is no surprise that some of these challenges are proving difficult for DART to address, as there are complex technical and political factors that mitigate design and implementation even if resources are brought to the table. At the same time, participation in project meetings by high-level technology leaders is unusual for EPSCoR and brings significant potential for success. *Accordingly, DART should narrow focus to key CI priorities in the remaining two years of the project to ensure that key foundations are solidly in place so that these research infrastructures can be utilized and built upon beyond the DART funding timeline.*

Accomplishments: DART has made good progress in some key areas of CI development that are setting the stage for Arkansas higher-ed to leap forward in terms of technology capabilities in research and education. Specifically,

- In-state identity federation for main campuses and systems is underway, and once complete will enable students and researchers from main Arkansas campuses to log in and access primary HPC compute systems and data storage using their managed, secure university credentials.
- High-speed networks for an Arkansas Research Platform: 100Gbps network connections for UAF and UAMS were implemented, laying the groundwork for friction-free exchange of data between systems, federated HPC queue partitions, or multi-cloud environments in the future.
- Regional CI coordination activities with ARE-ON and GPN have led to engagement with regional computing facility grants (OSU supercomputer and NRP nodes, for example). As regional Team Science continues to evolve, especially between GPN members, active participation/contribution by Arkansas CI personnel on regional infrastructure will be critical for AR faculty to take leadership roles in larger research projects.

- State-wide CI gap analysis via engagement with the CaRCC RCD Capabilities Model is crucial for DART to facilitate CI awareness across campuses and identify critical opportunities for CI improvement now and in the future.
- Training for DART researchers and students on Open OnDemand and CLI HPC access is the first step to scaling Data Science research beyond the desktop and into highly-parallel computational environments.
- Plans for managing CUI are absolutely critical for Data Science activities during and after DART project timelines.
- HIPAA-compliant storage at UAMS is a great first step in formalizing controlled-data environments for researchers working on sensitive/proprietary/DoD projects in the near future.

Concerns: In general, the plans for CI development during the DART project were ambitious, even if no serious systemic disruptions occurred during the project period. Given the launching of the project during the most disruptive timeframe in the last generation (C19 pandemic), the assessment, planning, coordination, and co-investment processes necessary across all major Arkansas higher-ed technology organizations are going too slowly to fully realize the vision of a truly friction-free and sustainable ARP before the end of DART funding. Key concerns contributing to this outlook are:

- A process for planning CI at the state level or across the institutions is not yet formalized.
- Roles of various campus CI elements/personnel in CI ecosystem development/implementation are not clear.
- There is not enough effort being put into survey, assessment, and benchmarking of CI capabilities from both the researcher (faculty and students) perspective as well as the campus administration (Research, IT, Education offices) perspectives.
- There is not a systematic approach to enabling access to national CI resources for underserved students and researchers from 2-4 year schools, which would require national identity federation and focused facilitation/familiarization activities on national platforms.
- Administration, operation, and development of key CI elements (campus HPC and data storage centers) are becoming more (not less) hampered by enterprise IT network management processes.
- Actual resources/needs to realize the ARP vision (and regional competitiveness/leadership) are not well articulated or delineated.
- DART data cybersecurity and access has been left in the hands of individual researchers and their respective campus IT support, which does not ensure cross-lab or cross-component exchange of data in a standardized manner and hampers team data science approaches.
- There is a lack of formalized standard CI architecture demonstrating how identity/access (and associated security) enables availability of various systems, applications, and communication protocols.
- There is a lack of centralized catalog of project algorithms, APIs, and protocols, scripts, etc. for enhanced activity tracking and measurement.

Recommendations: DART should use their remaining project time and resources to establish the process/procedural infrastructure to advance research technology capabilities in the near future. This will require evaluation and focus on specific areas and activities within CI, and also establishing key coordination and planning elements for the state/universities. Primary recommendations in this area include:

- Create a formal statewide CI “advisory” or “development” group with participation from all major research institutional technology organizations and the regional networks (ARE-ON & GPN), in order to accelerate knowledge exchange, language and priority alignment, and professional development of enterprise IT personnel in the concepts and requirements of research technology support.
- Charge this statewide group with developing an AR CI Plan (as a section of the state S&T Plan) that addresses 1-10 year CI objectives, required resources, and timelines.
- Ensure that multiple personnel in this group are active contributors to the national research computing and data community in order to facilitate alignment and adoption of national best practices in CI (for example: CaRCC, NSF Campus Champions, Research Software Engineers, etc.).
- Focus active development on long-term foundational CI objectives, including identity federation (internal to AR and external to national/international systems), research network management and facilitation for CI developers and sysadmin, identifying security models for CI, planning CI ecosystem elements and scope, implementing regular assessment and benchmarking of CI, and planning sustainability of CI.
- As iterated by the year #2 EAB, create a formal standardized Data Sharing/Privacy/Contingency plan for DART, in collaboration with institutional CISOs and research policy personnel, that sets baseline expectations of cross-component data transparency and sharing mechanisms.
- The EAB suggests building a common system architecture diagram to promote visibility into how different systems and applications interact (especially with respect to identity, access, and security) along with data and code source traceability.

Data Lifecycle & Curation Component

General Feedback: Overall, the researcher's use of the parameter data discovery process with 14 input parameters to tokenize data, generate groups with similar features, establish pairwise linking, and find relevant components is impressive. The DC component clearly demonstrates a strong mechanism for parameter discovery, and the traceable steps make it easy to follow the process. However, it might also be worthwhile for the researcher to explore user curation patterns based on different use cases and stress test the processes to ensure that they are robust and can handle various scenarios. This would provide an additional layer of validation and enhance the credibility of the results.

Accomplishments: DART created an autoencoder technique that enhances the capability of unsupervised and self-supervised deep learning approaches to handle and categorize significantly larger datasets. The team also solved the “out-of-memory” problem caused by the creation of shared memory tables/dictionaries.

Concerns: There is significant dependency on Hadoop MapReduce, which is becoming more of a legacy big data approach in terms of performance and industry utilization. Alignment of skills learned by students with industry needs on more modern technologies (*e.g.*, Hadoop vs. Apache Spark).

Recommendations

- Beyond protecting and securing data access, it would be advantageous for the team to consider more visible metadata mechanisms that promote automated data discovery, along with meaningful enrichment attributes to classify data's lifecycle. This includes identifying what the data are, their source, and who to contact if there are any questions. Especially in the industry context, attention to data management and discovery is needed in order to fully realize the value of a developed solution.
- The DC team's focus on Hadoop environments could benefit from exploring more modern technologies, such as Apache Spark. Spark offers a streamlined programming model and efficient data processing, making it a valuable skill to acquire.

Learning & Prediction Component

General Feedback: The learning and prediction (LP) team employs statistical modeling and machine learning to understand the complex data in many fields and to help people make decisions. The data involved in the projects include recurrent event data, medical segmentation on less labeled data, and image data. By engaging these diverse data types, the DART team remains flexible to pivot to partnership opportunities with industry.

Accomplishments: The project team has been very productive with a total of 37 publications including 18 journal articles, 18 conference proceedings and 1 book, as of April 20, 2023. Furthermore, Emre Celebi and colleagues guest-edited two special issues in the area of medical image analysis with high impact factors. Ahmad Al-Shami received the Faculty Excellence award for research at Southern Arkansas University. These are excellent academic achievements and benchmarks for DART performance.

Recommendations: The EAB recommends more explicit collaboration with other domain experts and researchers to test and apply the developed methods and models into practical use, and more importantly share the programming code and generated data with the DART community.

Social Media Component

General Feedback: The social media component in DART is a very strong research area in the program. Social movement mobilization, hate speech, discrimination, and sentiment analysis are the key topics investigated by the PIs within the last year. The EAB thinks that good progress has been made and that this component is accomplishing the majority of its objectives based on the project reports.

Accomplishments: The work in the Social Media component of DART resulted in several journal articles along with conference proceedings. Based on Year 2 EAB feedback and personnel changes in the Social Media group, new links between Social Media (SM) and Social Awareness (SA) teams were constructed since new participants are working in both of these areas.

Concerns: High turnover of personnel in the Social Media team remains a problem, in terms of continuity for project objectives, inter-component collaboration, and probably student progression as well. This results in limited research direction/coordination for all of the teams.

Recommendations

- The EAB suggests that the project team consider the following questions for future directions: 1) How to embed privacy of data/information, secure communication, a forensically sound process for information-driven context data collection, 2) How to build trust models and deal with biases, and more importantly 3) How to utilize/expand Social Media teams' actions/project scopes to support the overall research infrastructure and other teams in the project.
- Although the number of publications in the area is satisfactory, the EAB recommends the teams target top-tier journals and/or conferences such as IEEE Transactions, ACM, or Elsevier journals, and highly ranked peer-reviewed conferences.
- The EAB applauds the action of the PI team on adding personnel who can contribute to both SM and SA but also suggests that the PI team have contingency plans implemented and ready for personnel changes.

Social Awareness Component

General Feedback: The social awareness (SA) team is organized around a broad effort to develop new approaches centered around privacy, fairness, safety, and robustness in the context of data analytics and learning algorithms. In the past year, the team has continued a strong research effort, even in the face of significant turnover in personnel.

Accomplishments: The team has published high-quality work on fairness from multiple dimensions (e.g., intersectional fairness, fair regression, fairness in bandits, etc.), safety (e.g., fraud detection, hate speech detection), and robustness (e.g., anomaly detection). Methodologically, the team has drawn upon and advanced research in causal inference, counterfactual reasoning, deep learning, and information theory among many others. The summary slides provided to the EAB reference 5 journal, 5 book, and 25 conference publications, which is a reasonable amount for the number of people involved. The EAB is pleased to see collaboration among team members, including some cross-disciplinary work (e.g., computer scientists X. Wu and L. Zhang have been collaborating with sociologist A. Zajicek). The SA team has secured additional funding and has engaged with researchers from other teams (LP, DC, and WD) to secure follow-on collaborative research funding.

Concerns: The project strategic plan highlights a number of specific goals (*i.e.*, SA1 through SA7, though SA7 has been removed due to personnel change). It appears most of the progress so far is aligned along a few of the goals with less activity in some other ones. The turnover in personnel is understandable due to the high demands for experts in this area, but it is worth noting as a concern.

Recommendations

- Please update the project website with up-to-date publications and consider organizing (or at least annotating) by research area.
- With the personnel changes, it would be good to focus remaining efforts on areas of strength (*e.g.*, fairness) that can bring national and international recognition to the research quality of the team.
- There may be opportunities to connect the research advances to infrastructure advances made as part of the project, and to foster additional cross-campus collaborations.

Education Component

General Feedback: Overall, the project in this area shows a strong foundation in education development and shared resources, creating a supportive environment for talent to grow and succeed. The availability of channels such as summer research, certification, and degree plans also demonstrates a commitment to continuous learning and improvement. However, there seems to be a gap in promoting team science and cross-lab work within the project, with limited evidence suggesting that faculty members and students from different universities are actively engaging and collaborating with each other and not just some tools. It may be worth exploring ways to further promote cross-university collaboration, such as implementing shared technology and code repositories to facilitate more meaningful interactions between team members. A clear assessment plan of programmatic objectives still needs to be developed to ensure programmatic relevance and sustenance after grant award is completed.

Accomplishments: There are now three universities – University of Arkansas, the University of Central Arkansas, and Arkansas State University – with four-year data science programs, with a plan in place to have all three on a common 8 semester plan. There is also good progress on a 2+2 plan (which could potentially transform educational opportunity and skills development for underserved groups), and in finalizing a study abroad program. Good engagement of project findings both locally (at the ACC Annual Fall 2023 Conference) and nationally (at the ASEE 2023 Annual Conference) on topics related to data science careers in community college and improving a B.S. data science program. Good broadening participation effort through the Arkansas Summer Research Institute (with over 300 applicants).

Concerns: While data for each cohort shows growing participation, there is a lack of measurement for individual students that made the leap from k12 → college → undergrad → graduate → doctorate. For example: how could one track if there is continuous interest retained by newly nurtured students? If the 2+2 plan does not result in at least one functional implementation between institutions, the 2+2 program risks never being fully developed. Considering that this is one potential standout research-driven infrastructure improvement deliverable, the team should consider how to ensure that this is given the best chance to succeed through suitable assessment protocols.

Recommendations

- EAB suggests developing clear assessment protocols to understand how well learning objectives are being met programmatically (not just in individual courses). We suggest developing an assessment plan that includes elements of both direct and indirect measures of assessment in both the near term (time of graduation) and longer term (3 to 5 years after graduation) to evaluate if program objectives are indeed being met and up to date with latest developments in the data science arena.
- EAB suggests utilizing industry and state government collaborations to create an online education component to train the workforce in Arkansas with introductory-level data science knowledge.

- EAB suggests more field trips for researchers promoting cross-lab and cross-institution for team science development.
- EAB suggests investment in data storytelling development of researchers. For the initiatives taken by the DART in the education field, the work, experiences, and findings could be shared more in regional or national venues.
- EAB recommends implementing a mechanism to track the education “continuous” lifecycle of student talent profiles across various stages such as K12, undergraduate, graduate, and post-graduate. This will enable forecasting of the growing data science workforce and ensure broadening participation in the field.

Other Elements (Communication, Workforce Development, Outreach, Partnerships)

General Feedback: The EAB addressed portions of these topics as part of the feedback for each of the project components, and did not perform a separate review on DART project structure or ancillary organization. Overall, we feel like the project is doing as well as it can to pivot off of the COVID activity restrictions and personnel turnover, and is particularly well-connected with industry in Arkansas due to the placement of the EPSCoR PD within the Department of Commerce.

Formal Recommendations (Written response required)

Recognizing that there are only two years left in DART, and personnel turnover remains a challenge, the EAB recommends that the team pick feasible, simple, and effective end-of-project goals that will ensure the success of DART in its research infrastructure improvement mission. The main areas of concern and recommendation based on the above EAB report can be distilled to the following:

(1) Establish or articulate a mechanism that enables and facilitates cross-component activity towards true multi-disciplinary data science; perhaps focused on regular student interchanges across all components, a common code repository, common machine-readable metadata standards, a project data commons, or some combination of these.

(2) Complete initial stages of a friction-free cyberinfrastructure for effective statewide data science; roadmaps and/or pilot implementations for some combination of the following: regional CI assessment, formal regional CI planning, permanent Arkansas CI working group, global identity federation for AR higher-ed, common research security approach, campus enterprise IT/networking alignment/support, ARP infrastructure status and monitoring dashboard, one-stop ARP onboarding/offboarding workflow, and/or common communications channels for ARP users, facilitators, and engineers.

(3) Focus on industry-aligned measures of workforce/education success and skills development for components: are participating scientists partnered with industry effectively where lab workflow and therefore student skills development are concerned? How much alignment is there between the research components and the Education portions of DART where specific skills and tools for Data Science are concerned?