EAB Report for Project Year 2

(July 1 2021 – June 30 2022)

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

The DART External Advisory Board (EAB) consists of the following members:

Dr. James Caverlee	Texas A&M University
Dr. Weisong Shi	University of Delaware
Dr. Hoda Eldardiry	Virginia Tech
Dr. Donald Adjeroh	West Virginia University
Dr. Srinivasan Parthasarathy	The Ohio State University
Dr. Michael Khonsari	Louisiana Board of Regents and Louisiana State University
Dr. Huan Liu	Arizona State University
Dr. Dirk Reiners	University of Central Florida
Mr. James Deaton	Great Plains Network
Dr. Carolina Cruz Neira	University of Central Florida
Dr. Jason Leigh	University of Hawai‘i at Mānoa
Dr. Hongmei Jiang	Northwestern University
Dr. Noushin Ghaffari	Prairie View A&M University

The External Review Board (EAB) was provided the following information prior to a virtual meeting on May 19, 2022, to review the accomplishment of the project.

1. Year 2 annual report
2. Year 2 reporting tables
3. Year 2 highlights
4. A link to the Year 2 video presentations

A number of board members met with the DART team in-person in Arkansas during the all-hands annual meeting in May 2022. Due to scheduling conflicts, other Board members held virtual meetings in May, September and October to discuss the project accomplishments with the DART leadership. In what follows, a summary of the discussions in terms of achievements and recommendations for the various thrusts is provided.

1. Coordinated Cyberinfrastructure

The EAB noted an incredible effort at the recent DART Annual Meeting, onboarding researchers to the various computing clusters with one EAB member stepping through the process for access to both UAF's Pinnacle and UAMS's Grace. A wonderful demonstration of the progress in leveraging ARP within the DART project.

The Coordinated Cyberinfrastructure Team has a set of challenging aspects to address. First and foremost is the federated identity and access management of the respective campus research computing resources within the Arkansas Research Platform contributors. Federated access to computing resources is also an issue at the national level with shifts in progress during the migration from XSEDE to ACCESS. The Coordinated CI Team appears to have a great set of options and new campus participants are exploring and experimenting with a variety of solutions, commercial and open-source. The EAB suggests the team provide updates on their progress and selected path, as it would be of interest to the broader community.

Parallel to the federated access POCs, the EAB suggests the team explore methods to standardize the environment for collecting metrics from the ARP clusters. Recent NSF solicitations funding research computing have encouraged this; in fact, ACCESS has renewed investments in XDMoD. As ARP is further formalized and leveraging the respective computing clusters, collecting and reporting metrics across the participants becomes a priority to demonstrate the value. Open XDMoD would be a natural fit for this effort.

Cybersecurity within research cyberinfrastructure is a dominant concern during modifications and improvements. The EAB suggests the campus CISOs or designees of the participating campuses have a series of focused conversations and documentation of the practices and policies in place. In addition, they should include discussions with other regional campuses to compare existing multi-institutional research computing sharing practices. SHARP-CI has helped initiate these conversations, and the EAB suggests the team leverage the SHARP-CI team to continue this effort.

2. Data Life Cycle and Curation

The EAB is pleased to see that the Data Life Cycle and Curation component has made very good progress in the last year and is meeting many of its objectives. For example, faculty have implemented DWM in Python POC in 10th release, version 2.21; automated data cleansing to include data corrections based on record-to-record comparisons with blocks and clusters; developed a novel framework for scalable Entity Resolution using NLM for Locality Sensitive Hashing and Machine Learning achieving accuracy over 95% with a nearly linear runtime; developed a computational framework integrating multi-layer genomics data to identify transcriptome and pathway dysregulations in autism spectrum disorder; investigated the expression alterations of survival-related genes in various immune cell types when combining breast cancer bulk and single-cell RNA sequencing data; established a computational workflow of several combined machine learning approaches to identify biomarkers for both prostate cancer using metabolomics data and chemotherapy-induced cardiotoxicity among breast; downloaded more than 300k bacterial and archaeal genomes from the NCBI and complete set of genomes from Integrated Microbial Genomes and Microbiomes project; and developed and published a program, ProdMX, to speed up genome comparison more than a million-fold compared to traditional alignment methods.

The EAB suggest that the team in this component share their findings with the DART team and the broad scientific community. The idea of holding a team retreat this fall to revise the DC strategic plan is very good. It has the potential to bring the whole team on the same page in terms of language and terminology, to make a cohesive goal for this component, and the adoption of the proposed technologies.

Since facilitating the collaboration among multiple campuses in Arkansas is one of the goals of this project, the EAB suggests exchanging Ph.D. students among different campuses during the summer might be a

good idea. Also, during the retreat or annual conference, bringing diverse groups together to discuss the proposal submission plan for the next year might be a good direction too.

The EAB is delighted to see that the industry collaborations in this team are strong and continue to grow.

3. Social Awareness

The social awareness (SA) team has done a great job writing literature review papers and survey papers on privacy, statistical fairness, and causal fairness. They have developed various algorithms to provide privacy preservation, fairness, safety, and robustness of data analytics, data collection, data sharing, and decision making, with multiple publications in leading conferences and multiple awarded grants and training of numerous graduate students. All seven projects have been conducted as scheduled. It seems that SA1 and SA2 have been completed. The EAB would like to see the SA team plans to expand the activities, have better communication of SA work and results, and invite a social scientist member to the team.

4. Social Media & Networks

The Social Media (SM) and Networks component has four research thrusts: (1) Mining cyber argumentation data for collective opinions and their evolution; (2) Socio-computational models for safer social media; (3) Auto-annotation of multimedia data; and (4) Informing disaster response with social media. Based on the project and external evaluation reports, the EAB believes that good progress has been made and that this component is meeting most of its objectives. The SM team has determined key features and software design document for the cyber social network platform; developed, tested, and deployed a data collection framework with real-time dashboard to monitor progress with alerting capabilities; revised the taxonomy to characterize OIE based on social media platforms; studied cyber campaigns and characteristics of platforms and involved information actors and selected Hurricane Harvey to represent a large-scale and geographically widespread disaster scenario. Research into the auto-annotation of multimedia data goal appears to be trailing the progress being made in the other goals/objectives, as noted by the external reviewers. Please address the concerns raised by the evaluator, including “obtain and index content types for at least two disaster scenarios” and “developing GIS system to display real-time road status inputs” have not been done yet.

The EAB suggests the project team consider the following questions raised in the RSV report as potential areas of interest for the next step, including: (i) how to incentivize users to provide trusted information; (ii) how to develop trust models and handle biases; and (iii) how to reliably assess information-driven context or situation awareness leading to safety and security. The number of papers generated in this component is beyond expectation; however, the EAB suggests that the team keep improving the quality of the work by publishing in top conferences in the field. Targeting the conferences listed on ankings.org might be a good start.

Also, as one of the co-leaders of the team left the university, the EAB suggests that the project leadership take action as soon as possible to revise the SM strategic plan in consideration of the new team member's expertise.

5. Learning and Prediction

The objective of Learning and Prediction focuses on applying statistical methods and advanced deep learning techniques to analyze high-dimensional, dynamic, and unstructured data. Despite the turnover experienced by the team, research activities and progress according to the strategic plan have been successfully maintained and met many of its objectives. The EAB noted that the team implemented the EAB suggestions from last year with regard to greater elaboration on its achievements. The EAB suggests connecting Learning and Prediction more closely with other project areas, such as Social Awareness.

It is indeed exciting to note that this component has grown to cover a wide range of important activities that are core to the DART mission. Therefore, the EAB suggests that the team consider structured/facilitated ideation workshops involving the different LP project members with project members from other components (e.g., Data Curation, Social Awareness, Social Media and Networks, Education) to facilitate cross-fertilization of ideas and possibly address the RSV concern of impact and outreach both within DART and its industry partners and across the broader community.

6. Education

The educational goal of DART is to integrate the research from other components, coupled with industry needs, to develop a statewide data science educational ecosystem, including technical certificates, associate's degrees, and bachelor's degrees in data science and data analytics across the state of Arkansas. As noted in last years' report, the EAB was particularly impressed with how quickly the team has achieved this objective by working through various layers of bureaucracy across the state.

For its second-year updates, the EAB noted that the DART educational team reports significant postsecondary initiatives across six campuses spanning both community colleges and universities within the state. Additionally, some key efforts in the K-12 space were also noted. Finally, the DART ED team reports several strong government, industry, startup (including through the NSF I-Corps program) and non-profit collaborations which can help improve educational outcomes and opportunities for graduates.

The EAB reiterates its suggestion from last year to emphasize the importance of continuous evaluation of programmatic elements – this becomes increasingly important as the program's first graduates join the workforce next year. The EAB was pleased to see examples of ethics in data science curriculum development and would like to see more details on this element in next year's report.

The EAB does have several suggestions for the DART team to consider. First –given the impressive array of collaborations identified – board members suggest leveraging these industry and government collaborations to help amplify and highlight the importance of data science careers in local industries. Second, such connections can be leveraged in capstone projects within the curricular infrastructure being proposed. Third, research projects from other components of DART (e.g., seed grant recipients) can serve be leveraged in such courses and can serve as exemplars of research in pedagogy. Finally, as the first graduates come through the program, the DART ED team needs to start to think about holistic programmatic outcome assessments.

7. Workforce Development and Broadening Participation

One possible way to attract more attention to the SURE program could be to reach out to the smaller universities in Arkansas, for instance, the PUIs, and the HBCUs. These are more likely to have students with less access to research opportunities in areas covered by the DART project. Thus, working with faculty

at these smaller institutions/universities and then reaching out to their students could improve awareness of the SURE opportunity for such students. Also, providing the SURE program as a summer internship for the target students might help enhance interest in the program.

To encourage faculty in the smaller universities to be more involved with the SURE program, DART could also provide incentives for faculty members that take up SURE students, for instance, by providing summer support to the faculty. Since most of these faculty members are involved in teaching to cover their summer, this could also help them put more effort into their own research activities, especially during the summer.

8. Communication and Dissemination

Communication and dissemination are extremely important, and maintaining internal communication across campuses and research is challenging. In what follows, the EAB provides a variety of possible technology stacks that have been used in different campuses successfully, either of which are strong contenders for use in DART. The two main ecosystems are Microsoft Teams and Open Source.

MS Teams can serve as a highly integrated platform for chat, conference calls and file transfer. The main challenges people have faced are permissions and user management. Depending on the campus subscriptions, adding people from different organizations to a Teams group can be difficult. If all campuses in DART already have Microsoft Teams subscriptions, this may be a more or less complicated option. One pro for MS Teams may be centralized user management based on AD. This may simplify the setup here if this is set up at the central office level. Teams is not as powerful for real-time communications, which may be a desirable feature for close collaboration between DART members and groups.

An alternative tech stack is based on separate, freely available components. The main piece that we have seen being used is a discord for real-time communications, which is already used by many contributors already and makes it fairly simple to add a DART-specific group. For longer-term communications, we have seen good success with using web fora like phpBB or similar systems for recorded conversations. For maintaining shared definitions or onboarding documents, a shared document platform like a wiki like MediaWiki (or many others) has proven useful. Other shared document options are possible, but Wikis seem common enough for people to accept them fairly easily and successfully. This will need some seeding, possibly from the central coordination group for DART, but once the basics are provided, Wikis tend to grow well. For maintaining more project progress-related communications, workflow tools like Trello or Asana can easily maintain tasks and progress histories in an easy-to-understand and digest format. When used consistently, they can significantly simplify new user onboarding. Neither is Open Source, but they have free usage tiers for smaller projects. There are Open Source alternatives like taiga.io and others if needed.

The EAB believes that different technological solutions will enable effective post-covid communications and would encourage the project team to pick the easiest to integrate and closest to what is being used in the different groups already. We have had limited success with adding too many new components to existing environments unless no tech stack is overlapped between the admittedly pretty heterogeneous and fairly large number of groups in DART. We would encourage the central group to survey tools already in use and find the intersection between all the DART groups.

For dissemination, the project website is the primary tool, and it already does so very well. Including a section that clearly highlights project outcomes would be beneficial.