EAB Report May 2021

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	Wayne State University
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 Hawaiʻi at Mānoa
Dr. Hongmei Jiang	Northwestern University
Dr. Noushin Ghaffari	Prairie View A&M University

The Board was provided the following information prior to a virtual meeting on May 18, 2021, to review the accomplishment of the project.

1. Sections 4 and 5 of the RII proposal
2. Strategic Planning document
3. Year 1 Annual Report

The Board met with the Project Director and DART's leads and co-leads of different thrusts from 8:00-10:00 a.m. on May 18. The first hour of the discussion was a brief presentation of how the project was selected at the State level, its rationale, and the management organization, followed by specific goals of each thrust. The second hour was devoted to Q&A and discussion among the EAB and the project personnel. The purpose of the meeting was to provide initial feedback to the leadership in anticipation of the all-hands, face-to-face meeting scheduled to take place on September 13-14 in Arkansas, where there will be direct interaction with the rest of the project personnel and students.

The Board met internally from 8:00-10:00 a.m. on May 20th and discussed and summarized the team's accomplishments. In what follows, a summary of the discussions in terms of achievements and recommendations for the various thrusts are provided.

1. Coordinated CyberInfrastructure

The Board appreciates that the team has worked with the TrustedCI and EPOC groups to seek guidance on implementing security measures and data transport capabilities (such as establishing a ScienceDMZ and using Globus).

The Board is impressed by the detailed gap analysis, needs assessment, and updated cyberinfrastructure plan that the team has put together in the first year. It is commendable that an advisory board has been established for the cyberinfrastructure thrust, as this is a very complicated task in this project that needs to be compatible with all existing investments from different institutions.

Efforts to leverage The Carpentries and its structured lessons and workshops are commended in providing a path to utilize the CI. The panel appreciates that it will also have an impact in broadening participation across several facets of DART's initiatives.

The Board suggests doing a thought experiment on likely application use cases to ensure the bandwidth requirements for data movement across HPC/storage/visualization centers are being met by the proposed networking infrastructure updates. The Board also recommends that, in addition to implementing a robust CI for the science teams, the team should continue to innovate on or look to innovations in CI. Perhaps this can be supported through Seed funded projects, which can turn into proposals to NSF's Office of Advanced Cyberinfrastructure solicitations.

2. Data Life Cycle and Curation

The Board appreciates the team's objective to automate data cleaning and integration. The Board was particularly impressed with the detailed description of the team's effort on multi-threaded, distributed, resilient, and scalable data collection, curation, analysis, and visualization.

The Board recognizes the critical importance of this effort as it is foundational to the success of other parts of the project. The Board is looking forward to observing achieved progress on the interface and impact of the project methodologies for handling data volume, variety, velocity, and veracity.

3. Social Awareness

Social awareness research is a timely topic with the rapid development of AI technology. There are 7 research goals (SA1-SA7). Researchers include some world-leading experts on the topic who have produced influential papers. They collaborate with each other pursuing 7 different research goals.

The Board appreciates that the DART team has surveyed the SA research landscape. We encourage the team to produce a survey paper compiling the current state-of-the-art in the field. It would be great to see how students involved in the project collaborate and expand their social networks in these multidisciplinary research projects. Given the recent emphasis by some of the engaged industrial partners on AI and data ethics and fairness, some activity in these areas could be further encouraged, perhaps via targeted seed grant calls for proposals. The EAB recommends that measures of success and assessment on the social awareness thrust are clearly laid out to ensure progress with respect to stated objectives.

4. Social Media

Social media and networking analytics research have many scientific challenges: detection of mis/dis-information, the ways in which it is disseminated, and the scope of impact; analytically assessment of the collective impact of social media and networking on societal polarization and other social phenomena; and visualization of large social network data. In the 9-month reporting period, researchers have made solid progress on the four research goals (SM1 - SM4). SM1, SM3, and part of SM4 focus on the development and design of their proposed systems. SM2 has made impressive progress in terms of scientific research.

Recommendations: For SM1, SM3, and SM4, it would be timely to plan some scientific research hypotheses that can be verified using the platforms and systems developed by their research teams, and so scientific research findings can be obtained and shared via various dissemination channels. The coordination among the four teams will lead to concerted efforts.

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. All projects are currently on track. Notably, some projects have already resulted in conference papers and poster presentations.

The application of developed/to be developed methods on transaction data fits well with the learning and prediction objective. In particular, predicting opioid usage is a commendable application. However, it is not clear if there are any preliminary results on these methods and their verification in year one or if it will be accomplished in the upcoming years. Overall, the prediction and learning objective needs more elaboration on its achievements and its future goal since it is one of the most important components of the DART.

Also, given the issue of trust and privacy that is at the core of the research in this project, there should be more effort on how the project has considered IRB issues related to accessing certain types of data required for the proposed work. This issue is related to several themes in the project, beyond Social Media and Learning & Prediction.

6. Education

The objectives to develop a consistent and collaborative interdisciplinary multi-college and multi-institutional B.S. and Associate degree and certificate program in Data Science and Analytics, and coordinate a team to create a Statewide path for Data Science for the Arkansas Department of Education are, indeed, laudable. Members of the EAB were particularly impressed by the fact that the team was able to coordinate a statewide effort in a short time, working through various bureaucratic elements. The summer camp engagement (planned) was notable as well. DART provides training and educational opportunities for undergraduate and graduate students and K12 teachers and postsecondary faculty members.

The EAB recommends that measures of success and assessment on the educational thrust are clearly laid out to ensure progress with respect to stated objectives. Continuous evaluation of programmatic elements is essential. Another suggestion is to inculcate the discussion of ethics in data analysis in the curriculum. The Board understands that such a course is currently under development.

EAB recommends providing more details on diversifying and inclusion aspects of the educational efforts. It was mentioned in the report that the MSI, K12 and K20 would be included in the educational efforts. It would be helpful to see more details on the team's plans on including MSI. There are sufficient details on K12 and K20 inclusion plans.

7. Workforce Development and Broadening Participation

In the Seed mini-grants program, the first round was offered in October 2020, and two awards of $5,000 were issued. One was awarded to the Arkansas Regional Innovation Hub for a virtual field trip project entitled "STEM Saturdays" and another one to the Henderson State University STEM Center for a project entitled "ConneCTED: Developing Communities of Practice with an Emphasis on Computational Thinking and Engineering Design".

It is excellent to see that all 55 student research assistantship positions were filled during Year 1.

The EAB noted that the DART team has not yet established a regular meeting focused solely on opportunities for students in ACDS. Like DART, the operations of ACDS have been
COVID-impacted. However, Under COVID-impacted operations, ACDS has focused on the development of apprenticeship opportunities through partners like the Arkansas Coding Academy and the Forge.

8. Communication and Dissemination

Communication within a large project and to wider stakeholders and the general public is a non-trivial problem.

The EAB appreciates their creation of a networking diagram illustrating the connections between researchers and the various thrusts they are involved in.
https://app.powerbi.com/view?r=eyJrIjoiN2JiNmQ5NTctODlhMy00M2ZjLThlNmQtZjRjYWJlNTljYzQ3IiwidCI6Ijc5Yzc0MmM0LWU2MWMtNGZhNS1iZTg5LWEzY2I1NjZhODBkMSIsImMiOjN9

The Board encourages the team to continue to improve on this visualization so that it can be used by the DART team members as well as external stakeholders to identify synergistic opportunities for collaboration. For example: when one clicks on an individual team member, it would be useful to show a brief pop-up "baseball card" of the team member and home page URL, current EPSCoR project. Similarly, it would be useful to show a pop-up showing some brief info about the project when clicking on the project nodes. Given the strength of the IAB, one can imagine adding industrial partners as a 3rd category of nodes to show which industry groups are working with which projects (to the extent the industrial partners will allow).

The Board very much appreciates the numerous introductory videos and encourages the team to continue to produce them. In particular, the initial high-level overview of the problem each thrust was attacking was understandable to a non-technical audience and highly compelling. We also encourage future videos to be produced by individual projects or project members, giving individual investigators opportunities to expand their communication repertoire beyond just publishing papers.

In summary, the EAB is very impressed with DART's progress in a short length of time, particularly during the current pandemic and looks forward to the possibility of a face-to-face meeting in September 2021. For the all-hands meeting, it would be helpful to have detailed information about the plans to develop novel methods, especially in the areas of "Data Life Cycle and Curation" and "Learning and Prediction". We also look forward to a presentation by the external evaluator. The EAB is particularly interested in learning more about DART's accomplishments toward tech transfer and sustainability planning.