Lessons from an Online Multidisciplinary Undergraduate Summer Research Program

Matthias K. Gobbert[1] and Jianwu Wang[2]

[1] Dept. of Mathematics and Statistics, University of Maryland, Baltimore County,
Baltimore, MD 21250, USA, gobbert@umbc.edu
[2] Dept. of Information Systems, University of Maryland, Baltimore County,
Baltimore, MD 21250, USA, jianwu@umbc.edu

Abstract. During 2018, 2019, and 2020, the UMBC CyberTraining initiative "Big Data + HPC + Atmospheric Sciences" created an online team-based training program for advanced graduate students and junior researchers that trained a total of 58 participants. The year 2020 included 6 undergraduate students. Based on this experience, the authors created the summer undergraduate research program Online Interdisciplinary Big Data Analytics in Science and Engineering that will conduct 8-week online team-based undergraduate research programs (bigdatareu.umbc.edu) in the summers 2021, 2022, and 2023. Given the context of many institutions potentially expanding their online instruction, we share our experiences how the successful lessons from CyberTraining transfer to a high-intensity full-time online summer undergraduate research program.

Keywords: Online Education, Big Data, High-Performance Computing, Multidisciplinary Education, Team-Based Learning

1 Introduction

Next to theory and experimentation, computation has become the third pillar [1] and data-driven science has become the fourth pillar of the scientific discovery process [2] for many disciplines and critical to their research advances, such as bioinformatics, physics, computational chemistry, and mechanical engineering. It demands requirements on a training explaining how data and computation related techniques can help scientific discovery. Yet such a "Data + Computing + X" training is often missing in current curriculum design.

In 2017, the U.S. National Science Foundation (NSF) published the solicitation "Training-based Workforce Development for Advanced Cyberinfrastructure (CyberTraining)" designed to address this national need. This program continues currently with solicitation number NSF 19-524. Four faculty from three departments across two academic colleges at UMBC joined in response and proposed the UMBC CyberTraining initiative to create the nationwide online team-based training program "Big Data + HPC + Atmospheric Sciences" (cybertraining.umbc.edu) for students in three disciplines (Computing, Mathematics, and

Physics) to foster multidisciplinary research and education using advanced cyberinfrastructure (CI) resources and techniques. Figure 1 illustrates graphically the connection between the disciplines.

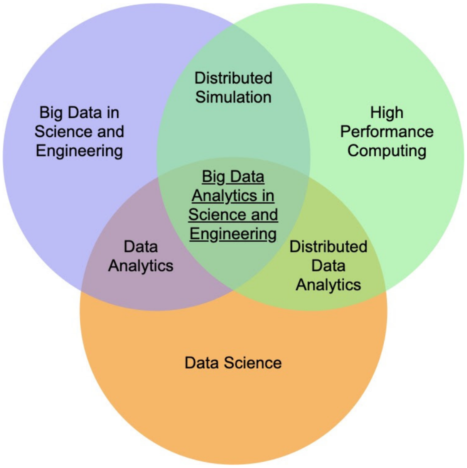
Big Data in Science and Engineering
Distributed Simulation
High Performance Computing
Big Data Analytics in Science and Engineering
Data Analytics
Distributed Data Analytics
Data Science

Fig. 1. Illustration of the connections between Big Data, HPC, and their applications in science and engineering.

2 The CyberTraining Program on Big Data + HPC + Atmospheric Sciences

The CyberTraining program teaches participants how to apply knowledge and skills of high-performance computing (HPC) and big data to solve challenges in Atmospheric Sciences. We focused on the application area of atmospheric physics and within it radiative transfer in clouds and global climate modeling, since these topics are important, pose computational challenges, and offer opportunities for big data techniques to demonstrate their impacts. The NSF funded our proposal in the inaugural year 2017 (OAC–1730250) for training programs conducted in 2018, 2019, and 2020 [3].

This program trained 58 participants and we reported on our experiences in conducting such training online and team-based with participants ranging from undergraduates (NSF-funded through an REU Supplement in Year 3), graduate students, post-docs/non-TT faculty, and TT (tenure-track) junior faculty [4]. We specifically describe how to practically create the necessary training material, chiefly the tapings of lectures for later asynchronous online delivery of contents and homework, during Year 1, and how to accomplish this in an institutionally supportive environment, but without the type of resources an institution with an institutional focus on online teaching would have. Thus, we wish to share our experiences to regular faculty, who might want to add aspects of online teaching to their repertoire. Table 1 summarizes the profile of the participants for our program over the three years.

Table 1. Profile of participants for our training program.

	under-graduate	graduate	postdocs	faculty	total participants	female participants	total teams
Year 1	0	9	4	3	16	7	5
Year 2	0	14	2	1	17	6	5
Year 3	6	11	4	4	25	14	8
Total	6	34	10	8	58	27	18

3 REU Site: Online Interdisciplinary Big Data Analytics in Science and Engineering

Based on the experience with online team-based CyberTraining on Big Data, we applied and were awarded a summer undergraduate research program, called an REU (Research Experiences for Undergraduates) Site by the U.S. National Science Foundation (OAC–2050943) for programs in summers 2021, 2022, and 2023. The proposal uses the experience to create a novel online team-based REU Site on the same topic for undergraduate students. We share here lessons learned of how to conduct a summer research program online.

Recruiting, participants, projects. We received more than 120 applications of highly qualified and motivated students. The 9 participants include 3 females, 1 Native American, 2 African Americans, 3 Asians, 3 Caucasians, 1 Hispanic, and 1 student with a disability. 4 of the 8 NSF-funded participants are from institutions with limited research opportunities according to the Carnegie Classification of Institutions of Higher Education.

The 2021 program consists of 2 teams working on the projects (i) Big Data and Machine Learning Techniques for Sea Ice Prediction and (ii) Big Data and Machine Learning Techniques for Medical Image Classification. The projects involve collaborators in the application areas, Yiyi Huang from NASA Langley Research Center and Jerimy Polf from the University of Maryland School of Medicine, respectively. The research uses the CPU, GPU, and Big Data clusters in the UMBC High Performance Computing Facility (HPCF, hpcf.umbc.edu), giving participants a real-life experience on a shared distributed-memory cluster running Linux with batch scheduling of jobs, etc.

Schedule and ancillary activities. Figure 2 provides a schematic of the 8 weeks by distinguishing the three phases of instruction, research, and dissemination, which have distinct purposes, but necessarily need to overlap in time. The first week of the program consists of training on this cluster (Linux OS, slurm scheduler, power of parallel computing) followed by a concise introduction to data science, machine learning, and their software on our cluster (Python, NumPy, Pandas, matplotlib, Tensorflow, Keras, Horovod). The teams are guided by the authors of this note as faculty mentors and by one dedicated graduate assistant for each team; these are our own experienced PhD students/candidates, who also serve as TAs during the training. Simultaneous to this training, each mentor has the students do some background reading on their team's project, and the week ends with presentations by the outside collaborators to *all* teams, followed by social meeting of each team with its collaborator. This sets the stage for weekly progress updates by each team to all participants to widen the communication experience beyond the team.

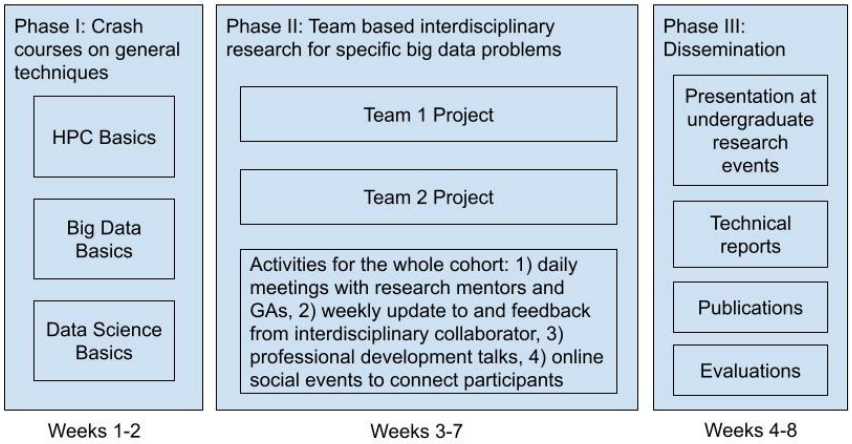
Phase I: Crash courses on general techniques
HPC Basics
Big Data Basics
Data Science Basics
Phase II: Team based interdisciplinary research for specific big data problems
Team 1 Project
Team 2 Project
Activities for the whole cohort: 1) daily meetings with research mentors and GAs, 2) weekly update to and feedback from interdisciplinary collaborator, 3) professional development talks, 4) online social events to connect participants
Phase III: Dissemination
Presentation at undergraduate research events
Technical reports
Publications
Evaluations
Weeks 1-2
Weeks 3-7
Weeks 4-8

Fig. 2. Structural outline of the 8-week schedule in three overlapping, but distinct phases.

The bulk of the weeks of the program are dedicated to conducting the research, with a transition from background reading to experimentation with the data to formulating a plan of attack and conducting the simulations. Throughout this, we introduce the need to document and disseminate by adding to the slide deck every week and writing the technical report one section at a time, beginning with background and literature review. The last week of the program culminates with each team giving a full oral presentation and publishing a technical report in the HPCF series. Along with the research training, the participants get exposure to a full set of professional development activities such as presentations by the Dean of the Graduate School about graduate school applications and by staff from the Career Center on services of such an office that exists also at the participants' home institution. We also invite a wide range of VIPs to the program, from the department chair to the university president to explain the structure of a university more in depth than many know at this point.

An interesting practical aspect of an online program is the need to be mindful of participants from a range of time zones. We ask applicants already to confirm that they will be able to work on an Eastern time zone, where the host institution is located, but we typically schedule formal meetings not earlier than 10 or 11 a.m. Eastern time to accommodate participants on the West Coast. The teams themselves, both individually and in subgroups, definitely work earlier already, depending on where they are located.

Tools and techniques specific to online learning and research. The online format of our program requires the careful and deliberate choice of tools and techniques to facilitate online collaboration and particularly the creation of a community among each team and among all participants.

Webex: We choose Webex[3] as the meeting and messaging platform. In this software, running on any operating system including smart phones in a web browser or an app, we create, using Webex jargon, a "Team" for the REU Site 2021, under which we create "Spaces" for Instruction, Team 1, Team 2, and Staff. Each space is set up for a specific set of members (e.g., Team 1 or 2 only, or staff only), comprises a synchronous online meeting space, a chat for messages that persist (not just during one online session), a contents area for files and links, and more. The online meeting area lets all attendees see each others in video feed, allows for screen sharing by any member, provides for recording capability in the cloud, and a connection to the messaging area of that space.

Google Drive: Our cluster in the UMBC High Performance Computing Facility provides shared storage for data and code, organized by Unix groups, which facilitates shared access to files. But we report that undergraduates are very accustomed to Google Docs (documents, slides, etc.) and thus we leverage this by using Google Drive for each team to store their research files. This readily facilitates the joint creation of slides for the weekly talks on Friday afternoon to all participants on the week's progress, which the teams give by smoothly transitioning from one student to another from section to section. We observe that the number of participants familiar with LaTeX is small, but these all used it through Overleaf, an online implementation for collaborative editing.

LinkedIn: We created a group in LinkedIn for our program and invited all participants to it, as well as established links between all of us. The first author of this note guided an REU Site program from 2010 to 2017 using the same idea. The long-term benefits of this approach are by now paying off: We get regular exciting updates on the educational and career achievements of the alumni of that program, e.g., graduating from top-notch universities, accepting jobs at equally exciting institutions and companies, and generally staying in touch. This setup also allows students from several years to be in touch with others beyond one year's cohort.

Social program: Research in an online format, where participants never meet each other in person, is difficult enough, and team-building a most substantial challenge, but the historically most significant outcome of a summer REU Site

[3] In early 2021, the product formally known as Webex Teams became simply Webex, the new mainline product, while the original Webex was renamed Webex Meetings.

of making live-long friends is even more difficult to replicate! First on the team-building for research, we accomplish this by posing challenging, tightly timed, urgent homework, which helps the team members to learn each others' strengths and weaknesses under pressure. Second for the social aspect, we find that what used to take place casually by joint lunches and evenings at dormitories needs to be created more deliberately. We thus use the daily check-up meeting of one hour to include open-ended discussions, such as on their impressions of attending a PhD defense, chats about career goals, review and feedback of each resume/CV, and finally strictly getting-to-know-each-other icebreakers (e.g., all team members locating one thing in common, then pairwise introductions first within each team, then beyond each team).

4 Conclusions

We hope that the lessons from our experience at conducting fully online research training programs are helpful, also beyond the specific contexts, such as the remarks about software tools used. These lessons might be important to many of you, since after the experience with an extended pandemic in 2020–21, we might all see opportunities as well as challenges in conducting not just teaching, but also particularly collaborative research among several institutions fully online, and the tools and ideas, e.g., about team-building, are applicable in that situation.

Acknowledgment

This work is supported in part by the U.S. National Science Foundation under the REU Site (OAC–2050943), CyberTraining (OAC–1730250), and MRI (OAC–1726023) programs. The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF).

References

1. Computational Science: Ensuring America's Competitiveness. https://www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf (2005)
2. National Academies of Sciences, Engineering, and Medicine and others, future directions for NSF advanced computing infrastructure to support US science and engineering in 2017–2020 (2016)
3. Wang, J., Gobbert, M.K., Zhang, Z., Gangopadhyay, A., Page, G.G.: Multidisciplinary education on Big Data + HPC + Atmospheric Sciences. In: Proceedings of the Workshop on Education for High-Performance Computing (EduHPC-17), p. 8 pages (2017). DOI 10.13016/M2KS6J78R
4. Wang, J., Gobbert, M.K., Zhang, Z., Gangopadhyay, A.: Team-based online multidisciplinary education on big data + high-performance computing + atmospheric sciences. In: The 16th International Conference on Frontiers in Education: Computer Science & Computer Engineering (FECS'20) (accepted (2020)). URL CyberTraining_FECS2020.pdf