Cybersecurity and Data Science Curriculum for Secondary Student Computing Programs

Richard Lawrence
rarensu@tamu.edu
HPRC[1]

Zhenhua He
happidence1@tamu.edu
HPRC[1]

Dhruva K. Chakravorty
chakravorty@tamu.edu
HPRC[1]

Wesley Brashear
wbrashear@tamu.edu
HPRC[1]

Honggao Liu
honggao@tamu.edu
HPRC[1]

Sandra B. Nite
s-nite@tamu.eduu
HPRC[1]

Lisa M. Perez
perez@tamu.edu
HPRC[1]

Chris P. Francis
chrispeterfrancis@tamu.edu
HPRC[1]

Nikhil Dronamraju
nikhildronam@tamu.edu
HPRC[1]

Xin Yang
karen89@tamu.edu
Medical College of Wisconsin
Milwaukee, WI, USA

Taresh Guleria
taresh@tamu.edu
HPRC[1]

Jeeeun Kim
jeeeunkim@tamu.edu
Texas A&M University
Computer Science & Engineering
College Station, TX, USA

ABSTRACT

Computing programs for secondary school students are rapidly becoming a staple at High Performance Computing (HPC) centers and Computer Science departments around the country. Developing curriculum that targets specific computing subfields with unmet needs remains a challenge. Here, we report on developments in the two week Summer Computing Academy (SCA) to focus on two such subfields. During the first week, 'Computing for a Better Tomorrow: Data Sciences', introduced students to real-life applications of big data processing. A variety of topics were covered, including genomics and bioinformatics, cloud computing, and machine learning. During the second week, 'Camp Secure: Cybersecurity', focused on issues related to principles of cybersecurity. Students were taught online safety, cryptography, and internet structure. The two weeks are unified by a common thread of Python programming. Modules from the SCA program may be implemented at other institutions with relative ease and promote cybertraining efforts nationwide.

KEYWORDS

Outreach, Cybersecurity, Data Science

1 INTRODUCTION

The explosion of computing in everyday life has led to increased interest in teaching computing skills to students of all ages. The projection of information technology and computer science job growth is 13 percent from 2016 to 2026 [11]. Computing education in K12 schools has lagged behind the demand due to lack of resources, including both physical resources and qualified teachers. Previously, the Summer Computing Academy (SCA) at Texas A&M has had success in introducing a variety of computing concepts to secondary school students, especially at the high school level and early college level [4][5]. This program utilizes well reviewed and effective pedagogical techniques to create an effective learning approach. The incorporation of hands-on projects related to the educational topic creates engaging experiences for the students, with the dual goal of internalizing foundational knowledge and driving further interest in the subject.

In addition to the general need for computing education, two computing subfields have been identified as having especially serious unmet workforce needs: Data Science and Cybersecurity. The growth of computing infrastructure has led to an explosion of data volume in multiple scientific and industrial disciplines. In particular, subjects like Genomics in the field of Biology and Machine Learning suffer from a wealth of data. At the same time, the growth of computing infrastructure has led to an explosion of threats, including cyber bullying among child peer groups, cyber crime such as ransomware, and even cyber warfare. It is essential that future efforts to educate students include a focus on these high demand computing subfields.

This paper presents improvements to the Summer Computing Academy curriculum for the purpose of better introducing the topics of Data Science and Cybersecurity to secondary school students. This is beneficial to the students who participate in our summer camps, who develop increased interest in science and computing after participating. In particular, students are made aware of career opportunities in Data Science and Cybersecurity. In addition, the improvements are also of benefit to high school computer science educators who could adopt our methods into their own practice.

[1]Texas A&M University High Performance Research Computing, College Station, TX

https://doi.org/10.22369/issn.2153-4136/14/2/2

In addition, several non-Python-based activities reinforced the cybersecurity concepts: hacking a C program with memory overflow, social engineering to "steal" the instructor's password, and investigating the local structure of the internet. This activity reinforced the importance of protecting personal information with strong passwords and careful consideration of what is posted online.

2 METHODS

2.1 Theoretical Framework

The original basis of the theoretical framework used in the camps came from experiential learning as it was expressed by John Dewey. It applies to formal and informal learning environments and is appropriate for summer camp instruction. It posits that knowledge envelope through experience, practice, and engagement [7]. Many teaching and learning philosophies and frameworks have been built upon this original work. One of those frameworks is the 5-E model of instruction that was originally developed in the field of science education but has been adapted to other STEM education fields. In this model, there is an initial activity to *engage* students in the topic of study. Secondly, students have the opportunity to *explore* ideas and concepts related to the topic at hand. Thirdly, the *explicit* portion involves formalizing the learning with formal definitions of terms, use of formulas, and solution methods. The fourth phase is the *elaborate* phase during which students solve problems and learn more deeply and the topic and applications. Finally, the *evaluation* occurs in order to assess students' learning [3]. In the SCA, students developed knowledge in this way. They engaged in hands-on learning through coding in Python and exploring websites with databases of information related to data science.

2.2 Learning Cycle

To address the need for greater understanding of and interest in careers in data science and cybersecurity, a curriculum consisting of supporting topics and primary topics was interweaved. In addition, it was taken into account that secondary students' ability to focus for longer periods of time is still developing. The first session was a discussion of what data science is all about, with a variety of examples given. Students participated in the discussion as the presenter asked questions to help them discover the use of data science applicable to their lives. This discussion was the initial *engagement* activity for the camp. Students then learned some Python programming that would be the foundational learning as they moved into the next phases. Websites with databases for genomes and DNA strands were provided with the opportunity for students to *explore* various diseases and corresponding DNA strands. Most *explicit* knowledge was gained with additional Python programming and a visit to a DNA processing lab to see the work in action and learn about the research currently underway. The *elaborate* phase flowed directly from the previous phase as students used coding skills and new skills and knowledge of 3D modeling, 3D printing, and visual programming to assemble the electronics from a connection diagram, run the device and observe the functionality. Finally, the *evaluation* of the learning was demonstrated in the team presentation on a topic related to the learning in the camp.

2.3 Curriculum Details

The summer computing academy was a two week program for 7-12th grade students. Each week of five days consisted of four full days and one half day. The two weeks had independent registration, which allowed students to attend one or both weeks. A high fraction of the students attended both weeks. For the purpose of the curriculum, we consider the students who attended both weeks to have attended a single two-week summer camp, because the materials in the two camps were largely unique. Despite the limitation of designing two curricula such that each could stand on its own, the overall two-week summer camp was still highly synergistic. The students who attended both were able to leverage what they learned in the first week and apply it during the second week. This was accomplished by giving the two weeks their own independent themes. The first week had the theme of data science. The second week had a theme of cybersecurity. However, the two weeks were unified by their common use of the Python programming language. Similar but not identical elements of Python syntax were taught in the two weeks so that the students could derive benefit from both, but still be able to participate in the activities even if they only attended one week. Both themes included projects to allow students a hands-on experience. The foundational learning for the summer camp was Python programming language. In order to complete their activities, the students needed the following elements of Python: Variables, Conditionals, Arrays, Dictionaries, and Loops. Other foundational topics were omitted to conserve time.

The first week of the summer camp had a theme of Data Science. For the Python-based activities, the students used the Python libraries NumPy, Matplotlib, Pandas, and Scikit-learn. The exercises were selected to give students a feel for research in data science, including data analysis, visualization, and machine learning. The contents includes how to load different formats of data files, manipulate the data for preprocessing, select best plotting methods for visualization, etc. Students were able to utilize the learned data skills and create linear regression and clustering models for different problems from the Scikit-learn machine learning library. Clustering analysis can be very helpful in cyber attacks detection. For example, Distributed Denial of Service (DDoS) attacks can be proactively detected by clustering analysis of its architecture consisting of the selection of handlers and agents, the communication and compromise [8]. These activities were designed to allow students to internalize the foundational Python knowledge and enhance their learning experience. In addition, several non-Python-based activities reinforced the data science concepts: searching databases for genetic information, 3D modeling and robotics, and a student team presentation on a topic of choice.

The second week of summer camp had a theme of Cybersecurity. For the Python-based activities, the students relied on fundamentals of Python and libraries taught in the first week. The two exercises were selected from the subject of Cryptography: a regular cipher, and the RSA algorithm. The students created their own ciphers in Python from scratch, and used a toy model of RSA also written in Python. Neither used any libraries, but instead reinforced concepts of programming. The cipher is the simpler of the two Cryptography tools. The progression of cipher to RSA provides both a technical and conceptual bridge, since the RSA algorithm is complex and

Table 1: Camper demographics

Gender	M	F				Total
	44	19				63
Ethnicity	A	W	AA	O	NR	Total
	28	23	1	1	10	63
School Type	Pub	Pri	Cha	HS	NR	Total
	42	4	2	5	10	63
Grade Level	8	9	10	11		Total
	2	14	16	12	19	63
Age	13	14	15	16	17	Total
	3	17	13	10	20	63

M = Male; F = Female; A = Asian; W = White; AA = African American/Black; O = Other; NR = Not Reported; Pub - Public; Pri = Private; Cha = Charter; HS = Homeschool

mysterious to younger students. The RSA algorithm is important to introduce to early learners because it is a fundamental technology of the internet, including established protocols such as HTTPS and newer ones such as Blockchain.

Students were also taught how to import Scikit-learn machine learning models to solve image classification problems. Image classification is proven to be very important in cybersecurity. Cybersecurity researchers built machine learning models to detect malware [1] and phishing websites [2] from the images created by applications. In this activity, we used the MNIST (Modified National Institute of Standards and Technology) handwritten digit image dataset [6]. The students learned about image representation, training a machine learning model, evaluating the machine learning model, making predictions, and analyzing the cases when the model makes correct and incorrect predictions on the images. This activity will help the students to be prepared to understand how image classification is applied to detect cyber threats.

3 RESULTS

3.1 Camper Demographics

The students in the camp were diverse in their interests, backgrounds, prior knowledge, and ethnicity. Table 1 provides some demographics of the campers. Some campers attended both camps, but they are represented only once the table. As can be seen in the table, there were only about half as many girls as boys. Most of the campers attend public school, and the vast majority of students were Asian or White.

3.2 Camper Reflections on Learning

Each day students were asked to complete a reflection that asked them to choose just one activity from the day on which to reflect. They were asked why they chose the activity and what they learned from the activity. The three most common activities mentioned were Python programming, visiting the DNA processing lab, and 3D design and printing. It is not surprising that Python programming was mentioned because students knew the camp would be a computing camp, and many high school computing classes use programming languages required by the particular courses. Python is not a prescribed language for advanced placement courses. The DNA processing lab visit tied in very closely with the learning they experienced in the classroom as they learned about data science, and this visit brought it all to life. Some mentioned that they did not expect to do work in this area in the future, but they still chose that activity because of the real-life aspect and connection to the learning activities in the classroom. The 3D design choice was a little surprising in that it was expected that many students would have had this experience previously. However, they would likely not have connected their 3D design object to a microcontroller to control its movement.

4 DISCUSSION

As has been shown in similar camp experiences, the opportunity to learn more about career opportunities in computer science and furthering their knowledge about several topical areas in the field may increase interest in pursuing careers in these fields [9]. In addition to learning more about different career possibilities, students gained introductory knowledge about actually doing some of this work. Gaining knowledge increases confidence, and confidence in an academic area increases the likelihood that a student will decide to pursue a career that involves that knowledge [10]. Students reflected on that knowledge and commented on their enjoyment of learning more about programming in Python and 3D design. Entering college with the confidence that they can be successful in computing or any of the many fields that require some knowledge of computing can increase the interest in a variety of STEM fields of study.

ACKNOWLEDGMENTS

The authors would like to thank staff, student workers and researchers at Texas A&M HPRC, Yang Liu, Michael Dickens, the Laboratory for Molecular Simulation, Dr. Steve Johnson, TAMU Engineering Innovation Center, TAMU IT, TEES IT, TexGen, Division of Research, the Texas Engineering Experiment Station IT, TAMU CPM and TAMU Provost IT for supporting the HPRC SCA program at Texas A&M. We gratefully acknowledge support from the National Science Foundation (NSF) Award OAC #1730695 "CyberTraining: CIP: CiSE-ProS: Cyberinfrastructure Security Education for Professionals and Students", and NSF Award OAC # 1925764 "CC: Cybterteam: South West Expertise in Training Education and Research", and Texas Workforce Commission award #1622SMP001.

REFERENCES

[1] Irina Baptista, Stavros Shiaeles, and Nicholas Kolokotronis. 2019. A novel malware detection system based on machine learning and binary visualization. In *2019 IEEE International Conference on Communications Workshops, ICC Workshops 2019 - Proceedings (2019 IEEE International Conference on Communications Workshops, ICC Workshops 2019 - Proceedings)*. Institute of Electrical and Electronics Engineers Inc., United States. https://doi.org/10.1109/ICCW.2019.8757060 2019 IEEE International Conference on Communications Workshops ; Conference date: 20-05-2019 Through 24-05-2019.

[2] Luke Barlow, Gueltoum Bendiab, Stavros Shiaeles, and Nick Savage. 2020. A Novel Approach to Detect Phishing Attacks using Binary Visualisation and Machine Learning. In *2020 IEEE World Congress on Services (SERVICES)*. 177–182. https://doi.org/10.1109/SERVICES48979.2020.00046

[3] Rodger W. Bybee. 2014. The BSCS 5E instructional model: Personal reflections and contemporary implications. *The Journal of Computational Science Education* 51 (2014), 10–13. Issue 8. https://doi.org/10.2505/4/SC14_051_08_10

[4] Dhruva K. Chakravorty, Marinus "Maikel" Pennings, Honggao Liu, Xien Thomas, Dylan Rodriguez, and Lisa M. Perez. 2020. Incorporating Complexity in Computing Camps for High School Students - A Report on the Summer Computing Camp at Texas A&M University. *The Journal of Computational Science Education* 11 (Jan. 2020), 12–20. Issue 1. https://doi.org/10.22369/issn.2153-4136/11/1/3

[5] Dhruva K. Chakravorty, Marinus "Maikel" Pennings, Honggao Liu, Zengyu "Sheldon" Wei, Dylan M. Rodriguez, Levi T. Jordan, Donald "Rick" McMullen, Noushin Ghaffari, and Shaina D. Le. 2019. Effectively Extending Computational Training Using Informal Means at Larger Institutions. *The Journal of Computational Science Education* 10 (Jan. 2019), 40–47. Issue 1. https://doi.org/10.22369/issn.2153-4136/10/1/7

[6] Li Deng. 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine* 29, 6 (2012), 141–142. https://doi.org/10.1109/MSP.2012.2211477

[7] John Dewey. 1938. *Experience and Education.* Macmillan, New York, NY.

[8] Keunsoo Lee, Juhyun Kim, Ki Hoon Kwon, Younggoo Han, and Sehun Kim. 2008. DDoS attack detection method using cluster analysis. *Expert Syst. Appl.* 34 (2008), 1659–1665. https://api.semanticscholar.org/CorpusID:6987397

[9] S. Nite, Ali Bicer, Kimberly Currens, and Rayan Tejani. 2020. Increasing STEM Interest through Coding with Microcontrollers. In *Proceedings of 2020 IEEE Frontiers in Education Conference: Education for a Sustainable Future*. 1–7. https://doi.org/10.1109/FIE44824.2020.9274273

[10] Sandra B. Nite, Devyn Chae Rice, and Rayan Tejani. 2020. Influences for Engineering Majors: Results of a Survey from a Major Research University. In *2020 ASEE Virtual Annual Conference Content Access*. ASEE Conferences, Virtual On line. https://peer.asee.org/34825.

[11] U.S. Bureau of Labor Statistics. 2018. Computer and information technology occupations. Retrieved August 28, 2023 from https://www.bls.gov/ooh/computer-and-information-technology/home.htm