2024 IEEE 10th Conference on Big Data Security on Cloud (BigDataSecurity) | 979-8-3503-8952-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/BigDataSecurity62737.2024.00009

An Education Program for Big Data Security and Privacy

Bhavani Thuraisingham*, Kim Nimon#, Latifur Khan*
bhavani.thuraisingham@utdallas.edu, knimon@uttyler.edu, lkhan@utdalas.edu
*The University of Texas at Dallas, #The University of Texas at Tyler

***Abstract*— This paper describes the course that we developed at the University of Texas at Dallas on Big Data Security and Privacy with funding from the National Science Foundation. The project started in 2017 and was completed in 2023. We have taught the course starting Fall 2020, Fall 2021, Fall 2022 and we are teaching it for the fourth time in Fall 2023. This is an extremely popular course and we limit the enrollment to 80 students. Students have some background in cyber security and big data analytics. This paper provides an overview of the organization of the course, the course content, and learning outcomes and the evaluation carried out. The members of the project include those who have expertise in (i) Cyber Security, (ii) Big Data Analytics, and (iii) Education Research.**

Keywords: Big Data, Security and Privacy, Trustworthy Machine Learning, Artificial Intelligence and Security, Education Research, Learning Outcomes

I. INTRODUCTION

Large amounts of data are being collected, stored, managed analyzed and shared for a variety of applications. The amount of such data collected daily could be in the range of zettabytes and even exabytes. This data could include heterogeneous data types such as text, images, audio, and video. In addition, we need techniques to retrieve this data rapidly. Such large amounts of data is known as Big Data. The management and analytics of such data is called Big Data Management and Analytics.

While large amounts of data are critical for a variety of applications, there are also major concerns. These include data privacy where it is now possible to collect massive amounts of data and analyze the data that could violate the privacy of individuals. Furthermore, the analytics techniques could also be biased and show preferences to certain segments of the population. In addition,, these techniques could be attacked to produce incorrect results. While it is important to conduct research in Big Data Security and Privacy which is also sometimes used interchangeably with Trustworthy Machine Learning (ML), we also need an educated workforce to tackle the many challenges involved. We have developed a strong and comprehensive education program at the University of Texas at Dallas under the umbrella of Trustworthy Artificial Intelligence (AI) that includes courses in Trustworthy Machine Learning, Big Data Security and Privacy, Secure Cloud Computing and Analyzing and Securing Social Media.

This paper will describe the education program that we have developed in Big Data Security and Privacy at the University of Texas at Dallas. We have also taught variations of this course that also includes Trustworthy Machine Learning. The organization of this paper is as follows. Section 2 describes our framework for Big Data Security and Privacy that has guided us in our curriculum development. The contents of the course are discussed in Section 3. The learning outcomes and the education research and evaluation of the course we carried out are discussed in Section 4. We have educated a global workforce based on the material developed or the course. Therefore our outreach activities including in Africa are discussed in Section 5. The paper is concluded in Section 6.

II. A FRAMEWORK FOR BIG DATA SECURITY AND PRIVACY

Figure 1 (taken from Thuraisingham, 2022), illustrates the framework we have developed for our education program in Big Data Security and Privacy. It is a layered architecture. Aspects such as Security, Privacy, Dependability, Integrity, Fairness, and Provenance cut across all the layers.

Layer 1 is the Infrastructure layer where we focus on Secure Cloud Computing. We offer modules that include a comprehensive overview of the cloud including access control models and visualization techniques. We also discuss attacks to the cloud as well as cloud governance. We expect that in the future this layer could likely be replaced by the Quantum Computing layer. The next layer (Layer 2) is the Big Data Layer which is essentially the data layer. Here, we teach the foundations including securing the Hadoop and MapReduce systems. We also discuss access control models at this layer and illustrate attacks to the big data. We address topics such as Big Data governance. In addition techniques for secure query processing and information sharing in the secure cloud are discussed. The next layer (Layer 3) is the Trustworthy Machine Learning layer. At this layer we discuss machine learning techniques, privacy aware machine learning, adversarial attacks to machine learning, and fairness in machine learning. We also discuss governance in AI/ML systems. The next layer (Layer 4) is the application layer. The applications could include domain applications such as healthcare and finance or technology applications such as social media and IoT. We discuss how social media systems could be analyzed

979-8-3503-8952-4/24/$31.00 ©2024 IEEE
DOI 10.1109/BigDataSecurity62737.2024.00009

and secured. We also discuss how IoT systems can be secured. In addition, we discuss governance in systems such as social media and IoT. In the future we are planning to include applications such as secure transportation systems.

Each layer takes advantage of the technologies provided by the layers below it. For example, the Big Data layer takes advantage of the secure cloud. The Machine learning layer takes advantage of the cloud and big data. The application layer takes advantage of machine learning, big data, and the cloud. As stated earlier, security, privacy, integrity, and fairness cut across all the layers.

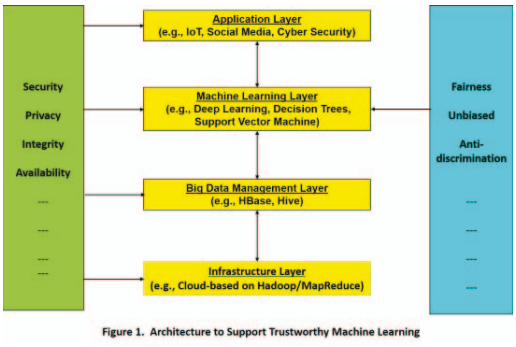

Figure 1. Architecture to Support Trustworthy Machine Learning

III Curriculum for Big Data Security and Privacy

Our Approach: We received funding from NSF to enhance our education program in Big Data Security and Privacy. While our main focus was on security and privacy for big data and machine learning, we also included aspects such as fairness and governance.

We used two reference books for the course and they are:

Big Data Analytics with Applications in Insider Threat Detection, CRC Press January 2018, B. Thuraisingham, P. Pallabi, M. Masud, and L. Khan, (Thuraisingham et al, 2017/2018).

Secure Data Science: Integrating Cyber Security and Data Science, CRC Press, April 2022, B. Thuraisingham, M. Kantarcioglu, and L. Khan (Thuraisingham et al, 2022)

The first book was influenced by the proposal we put together for the project in December 2016. The second book was influenced by some of the material we used for the course. In addition, we are preparing a third book on Trustworthy Machine Learning that is based on the framework of Figure 1.

The Course Work consisted of the following:

Two Term Papers: 10 points Each (Survey and/or Critiquing papers)

Team Programming Project: 26 points

Two Exams: 26 points each

Team Paper Presentations: 2 points

Total: 100 points

Students give team presentations on the programming project related to Big Data Security and Privacy on the last day of class as well as give demonstration of the system developed. Students also present a paper on Big Data Security and Privacy as a team.

Course Content: Lectures: Note that a lecture may span multiple dates or multiple lectures are taught during a lecture period. Lectures are grouped into modules. In particular, the course consists of the following six modules.

Module 1: **Background for Cyber Security and Big Data** that provides the background information both in cyber security and big data. Much of the material for this module is provided in the form of YouTube lectures by the course instructor (Thuraisingham).

Module 2: **Introduction to Big Data Security and Privacy** focusing on an overview of the topics as well as a general overview of special big data systems.

Module 3: **Secure Big Data Infrastructures that discusses Secure Cloud Computing** focusing on the Security for the Hadoop/MapReduce Platform.

Module 4: **Big Data Security and Privacy Functions**: This is the core module that discusses various concepts in Big Data security and privacy including access control model, privacy for big data, secure big data storage, secure big data query processing, attacks to big data, Big Data Governance, Integrating Cloud and Machine Leaning to detect attacks on Big Data, and Secure Big Data products and prototypes.

Module 5: **Trustworthy Machine Learning**: This module consists of lectures that cover Securing machine learning techniques, privacy aware machine learning and fairness in machine learning. One can consider this area to be Big Data Analytics.

Module 6: **Hosting Applications (Social Media/IoT) on Secure Big Data Infrastructures**. This module focuses on hosting secure social media applications as well as IoT applications on Big Data Infrastructures. Topics include security and privacy for social media, access control and privacy models for social media, attacks to social media, and fake news detection for social media. We also discus techniques for securing IoT systems.

Module 7: **Emerging Topics for Big Data Security and Privacy**. This module discusses novel topics such as Knowledge Graphs for Security, integrating blockchain with big data, Integrating Big Data Security and Privacy with applications such as transportation and space systems. Integrating quantum computing with big data is also discussed.

IV LEARNING OUTCOMES AND COURSE EVALUATION

Learning Outcomes:

- Students will be able to understand the three layer architecture for big data security and privacy that consists of the Infrastructure Layer, the Big Data Management and Analytics Layer and the Application Layer.
- Students will be able to develop access control models for big data, understand several types of

attacks and develop solutions such machine learning techniques to address the attacks.

- Students will be able to develop techniques for secure and private big data including risk-based secure and private data storage and secure big data query processing.
- Students will be able to understand interdisciplinary topics such as Big Data Governance that also discusses a framework.
- Students will understand the security issues for various commercial big data management systems.
- Students will learn to host Social Media Applications on Secure Big Data Infrastructures including understanding aspects of security and privacy for social media.
- Students will be able to understand social and psychological theories and the use of Machine Learning for Fake News Detection in Social Media.
- Students will also learn about the emerging directions integrated with Big Data security and Privacy including IoT, Blockchain, and Quantum Computing.

Course Evaluation: In Fall 2022 (the third time the course was offered), we sought to determine if, as a consequence of completing the course, students' self-beliefs improved. Consistent with prior literature, we assessed students' computing self-efficacy, grit, and STEM identity. Students were emailed a link to an electronic survey as we were concluding the course.

The survey incorporated a post-then design (Nimon et al, 2019). Consistent with the post-then design, we asked participants the same set of questions twice. First, they assessed themselves as the course was concluding (post) and second, they assessed themselves retrospectively before the course started (then). This design was especially helpful as IRB approval was not received until after the course started thereby not allowing a traditional pretest-posttest design. The design is also helpful as participants rate themselves with the same set of knowledge, skills, and attitudes and avoids response-shift bias that can occur with traditional pretest-posttest designs.

We assessed computing self-identity with the 2-item measure from Kreth et al. (2019) which has been used in evaluating students in computer science courses. A sample item for computing self-identity scale is :I am comfortable with learning computing concepts." The computing self-identity scale was measured with a five-point Likert scale (*Strongly Disagree* to *Strongly Agree*).

We assessed grit with Duckworth and Quinn's (2007) short grit scale. The scale is organized around two constructs, each containing four items. A sample item for the consistency of interest scale is "I often set a goal but later choose to pursue a different one." A sample item for the perseverance of effort scale is "I finish whatever I begin." The grit scales were measured with a five-point Likert scale with anchor points of "not at all like me" and "very much like me."

Measure	Thentest	Posttest	Change
Consistency of interest[a,b]	M = 3.75; SD = 1.03	M = 3.71; SD = 0.83	t = -0.31; p = .76; d = -0.08
Perseverance of effort[a]	M = 4.18; SD = 0.79	M = 4.04; SD = 0.73	t = -0.95; p = .36; d = -0.25
Self-efficacy	M = 4.07; SD = 0.94	M = 4.36; SD = 0.72	t = 1.53; p = .15; d = 0.41
Self-identity	M = 4.57; SD = 2.90	M = 5.64; SD = 2.02	t = 2.16; p = .05; d = 0.58

Note. n = 14. [a]measure of grit. [b]reverse-coded.

Table 1: Evaluation Results

We used a single item measuring for assessing STEM identity following the work of McDonald et al. (2019). As in McDonald et al., we created an identity overlap measure where students were instructed to select the picture that best described the current overlap of the image they have of themselves and their image of what a big data and cyber security professional was. Identity was measured with a 11-point scale (ranging from 0 to 10).

Fourteen students completed the survey in its entirety. Nine respondents (64%) identified as being a computer science major and 1 student identified as a software engineering major. All respondents identified as full-time students. The majority of respondents identified as male (64%), Gen Z (64%), and international (79%). As identified in Table 1, students who completed the course in Fall 22 improved on self-efficacy and self-identity but not on either measure of grit.

V OUTREACH ACTIVTIES

In addition to educating our students at UT Dallas, we have also educated a global workforce. Specifically, we teach a 40 hours class at the University if Dschang in Cameroon Africa (virtually) in Trustworthy Machine learning which includes much of the material developed for this course. In addition, we have also given keynote and featured address at outreach events such as Women in Communication Engineering, Women in Data Science/Data Mining, and at distinguished seminars. In addition, we are now part of the USDOT $20M University Technology Center in Transportation Systems Security led by Clemson University and we will be delivering courses utilizing the material covered in this course with a new application area and that is Transportation Systems. In addition, we will also continue to develop curriculum that integrates cyber security and

privacy with Data Science, Machine Learning, Artificial Intelligence and also delve into new application areas such as Space Systems in addition to Transportation systems.

VI CONCLUSION AND FUTURE DIRECTIONS

This paper has provided the details of our curriculum development effort in Big Data Security and Privacy. We first discussed a framework that guided the course development. Next we provided the details of our curriculum. This was followed by a discussion of the learning outcomes and the course evaluation. Finally, we discuss out outreach efforts in educating a global community.

We believe that the development of this course is a success story that is a collaboration between Computer Scientists and Education Researchers. We will continue to follow this approach in our course development in Blockchain and AI+Security (Artificial Intelligence and Security). For example, in Spring 2024 we are teaching courses on Blockchain (at the undergraduate level) as well as a course on Large Language Models at the PhD level that includes units on cyber security. We will also continue to educate a global community including female students and those from the underrepresented minority communities. Our second offering of the course in Trustworthy Machine Learning is planned to be taught virtually at the University of Dschange in Cameroon Africa for PhD students in Fall 2024. In addition, a major focus area will be a course on Artificial Intelligence and Security (AI+Security, also called Trustworthy AI)) that includes the contents of our Big Data Security and Privacy course, our Trustworthy Machine Learning course, and our Large Language Models and Security course.

REFERENCES

Duckworth and Quinn, 2009: Duckworth, A. L., and Quinn, P. D. (2009). Development and validation of the Short Grit Scale (GRIT–S). *Journal of Personality Assessment*, *91*(2), 166-174.

Kreth et al, 2019: Kreth, Q., Spirou, M. E., Budenstein, S., and Melkers, J. (2019) How prior experience and self-efficacy shape graduate student perceptions of an online learning environment in computing, Computer Science Education, 29(4), 357-381..

McDonald et al. 2019: McDonald M. M., Zeigler-Hill V., Vrabel J. K., and Escobar M. (2019). A single-item measure for assessing STEM identity. *Frontiers in Education,* 4(*78*).

Nimon et al, 2011: Nimon, K., Zigarmi, D., & Allen, J. (2011). Measures of program effectiveness based on retrospective pretest data: Are all created equal? *American Journal of Evaluation, 32,* 8–28. doi:10.1177/10982140103878354

Thuraisingham, 2022: Thuraisingham, B., Trustworthy Machine Learning, IEEE Intelligent Systems, 2022.

Thuraisingham et al, 2018: Thuraisingham, B., Pallabi, P., Masud, M., Khan, L. Big Data Analytics with Applications in Insider Threat Detection*, CRC Press*, 2017 December/ 2018 January.

Thuraisingham et al, 2022: Thuraisingham, B., Kantarcioglu M., Khan, L. Secure Data Science: Integrating Cyber Security and Data Science, *CRC Press*, 2022.

ACKNOWLEDGEMENT: "This material is based upon work supported by the National Science Foundation under Award No. (FAIN): DGE-1723602. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation."