2021 ASEE ANNUAL CONFERENCE
ASEE
Virtual Meeting | July 26–29, 2021 | Pacific Daylight Time
Paper ID #34337

2020 ETI Annual Summer School: Data Science and Engineering

Prof. Steven R. Biegalski, Georgia Institute of Technology

Steven Biegalski is the Chair of Nuclear and Radiological Engineering and Medical Physics Program at Georgia Institute of Technology. He has three degrees in nuclear engineering from University of Maryland, University of Florida, and University of Illinois, respectively. Early in his career Dr. Biegalski was the Director of Radionuclide Operations at the Center for Monitoring Research. In this position Dr. Biegalski led international efforts to develop and implement radionuclide effluent monitoring technologies. This work supported both US national capabilities and international treaties. Dr. Biegalski was a faculty member at The University of Texas at Austin for 15 years and held the position of Reactor Director for The University of Texas at Austin TRIGA reactor for over a decade. He has advised 25 Ph.D. students to graduation and holds Professional Engineering licenses in the states of Texas and Virginia. Dr. Biegalski's current research focus is on nuclear analytical methods, research isotope production, nuclear forensics, and nuclear non-proliferation.

Dr. Pavel V. Tsvetkov, Texas A&M University

Pavel V. Tsvetkov, Ph.D., is an Associate Professor in the Department of Nuclear Engineering, Texas A&M University. Dr. Tsvetkov's research program is focused on novel energy systems meeting global growing needs in sustainable resources. The project portfolio includes direct energy conversion, waste minimization efforts, novel reactor designs, instrumentation efforts, and data science and engineering for a broad range of applications targeting optimized designs and performance. He published over 300 papers in peer journals, conference proceedings and reports as well as served as an editor and major contributor for 14 books on energy, environment and nuclear energy.

Dr. Yuguo Tao, Georgia Institute of Technology

Yuguo Tao received B.S. and M.S. from Tianjin University in China, and PhD from University of New South Wales, Sydney, Australia. Prior to joining the Woodruff School of Mechanical Engineering in 2019, Dr. Tao was a Research Scientist at the School of Electrical and Computer Engineering, Georgia Institute of Technology since 2011. Dr. Tao's current works focus on supporting and assisting the overall management, execution and operations of the Consortium for Enabling Technologies and Innovation (ETI), and developing novel instrumentation at Laboratory for Nuclear Nonproliferation and Safety (LANNS).

Prof. Vladimir Sobes, University of Tennessee at Knoxville
Dr. Karl Pazdernik, Pacific Northwest National Laboratory
Simon Labov, Lawrence Livermore National Laboratory

Simon Labov is the Program Leader for Nuclear Detection Systems and the Nuclear Security Physics Group Leader at Lawrence Livermore National Laboratory in Livermore, California. Dr. Labov is an expert in nuclear detection systems, advanced spectral and multi-source analysis algorithms, distributed detector systems, and data analytics applied to nuclear threat detection.

Dr. David F. Williams, Oak Ridge National Laboratory

David F. Williams is a PhD chemical engineer with 37 years of professional experience at Oak Ridge National Laboratory and has publications in the nuclear fuel cycle domain spanning reprocessing, fuel fabrication, isotope production, advanced reactor concepts, and nonproliferation/nuclear security. Dave has been a principal investigator or managed projects/programs in each of these applications areas for the DOE Office of Science, Office of Nuclear Energy, Office of Environmental Management and multiple offices within DOE/NNSA.

Dr. James M. Ghawaly Jr., Oak Ridge National Laboratory

Dr. James Ghawaly Jr. is an applied data scientist in the Advanced Radiation Detection, Imaging, Data Science, and Applications group at Oak Ridge National Laboratory. James has a PhD in nuclear engineering, a MSc in computer engineering, and a BSc in nuclear engineering from the University of Tennessee Knoxville. His research is focused on the application of machine learning, accelerated computing, and other modern data science methods to the field of radiation detection, especially when it comes to anomaly detection in low signal to noise ratio environments. He has also performed fundamental research in the development of training algorithms for deep spiking neural networks.

Prof. Alfred Olivier Hero, University of Michigan

Alfred Hero is the John H. Holland Distinguished University Professor of Electrical Engineering and Computer Science and the R. Jamison and Betty Williams Professor of Engineering at the University of Michigan, Ann Arbor. He is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE) and the Society for Industrial and Applied Mathematics (SIAM). He is a recipient of the Fourier Award in Signal Processing from the IEEE. He is a Section Editor of the SIAM Journal on Mathematics of Data Science and a Senior Editor of the IEEE Journal on Selected Topics in Signal Processing.

2020 ETI Annual Summer School: Data Science and Engineering

Abstract

The Consortium for Enabling Technologies & Innovation (ETI) was established in 2019 to address emerging technologies within the context of nuclear nonproliferation. ETI creates a research and education environment to support cross-cutting technologies across three core disciplines: 1) computer and engineering science research specifically in a form of machine learning and high performance computing (HPC), 2) advanced manufacturing, and 3) nuclear detection technologies. For outreach and development, ETI hosted the first of three summer schools from August 24-28, 2020 with the theme of "Data Science and Engineering". The school was hosted in an on-line format and had over 200 participants. The recorded content is available on-line as a resource for students. This describes the hurtles and methods utilized to overcome obstacles limiting in-person workshops in 2020.

The summer school had four modules: 1) Fundamentals of data Applications, 2) Computational Machine Learning, 3) Bayesian Modeling and Inference, and 4) Data Science for Safeguards. Modules contained both lectures as well as student exercises. Poll Everywhere was utilized in some modules as an on-line method to engage large groups of students. Data based exercises were also conducted with students to ensure learning objectives were met. Upcoming ETI Summer Schools include Novel Instrumentation in 2021 and Advanced Manufacturing in 2022.

Background

The Consortium for Enabling Technologies & Innovation (ETI) is funded through the Department of Energy and is led by Georgia Institute of Technology.[1] ETI consists of twelve universities and twelve national laboratories as shown in Figure 1. ETI Figure 2 illustrates how this consortium addresses nuclear non-proliferation concerns through cross-disciplinary research in data science, advanced manufacturing, and novel instrumentation and sensors. Educational outreach is a key component of Department of Energy University Consortia. Other consortia including the Consortium for Monitoring Technologies and Innovation (MTV)[2], the Consortium for Verification Technology (CVT)[3], the Consortium for Nonproliferation Enabling Capabilities (CNEC)[4], and the Nuclear Science and Security Consortium (NSSC).[5]

ETI is organized so that advances in data sciences may be leveraged to promote the assessment of advanced manufacturing and novel instrumentation and sensors. To increase the advanced data science methods across all of ETI a summer school was offered August 24-28, 2020.

The goal of the Data Science and Engineering Summer School was to provide students a connection between nuclear non-proliferation applications and data science. Lectures provided a review of key topics and introduced data science methods via hands-on tutorials. Students were immersed in a collaborative environment.

The Data Science and Engineering Summer School was originally planned to be in-person. However, pandemic concerns in 2020 forced the school to transition to an on-line format. This transition had both benefits and challenges. One of the primary benefits was that the on-line format allowed for more participants resulting in 214 registrants for the course. The large number of registrations exceeded expectations and provided a challenge on how to maintain active engagement with the students.

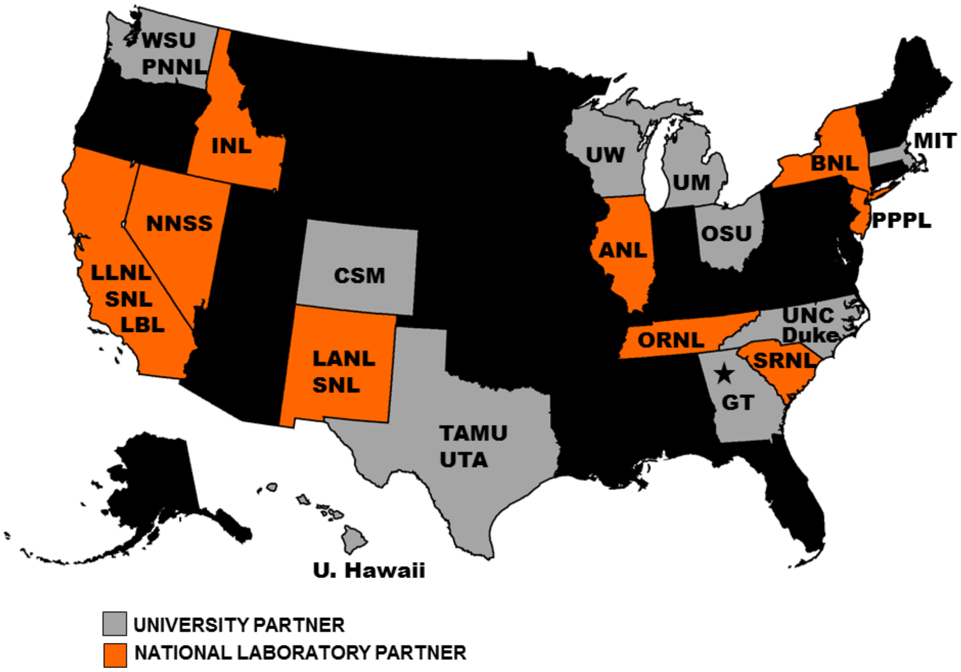

Figure 1. Universities in ETI

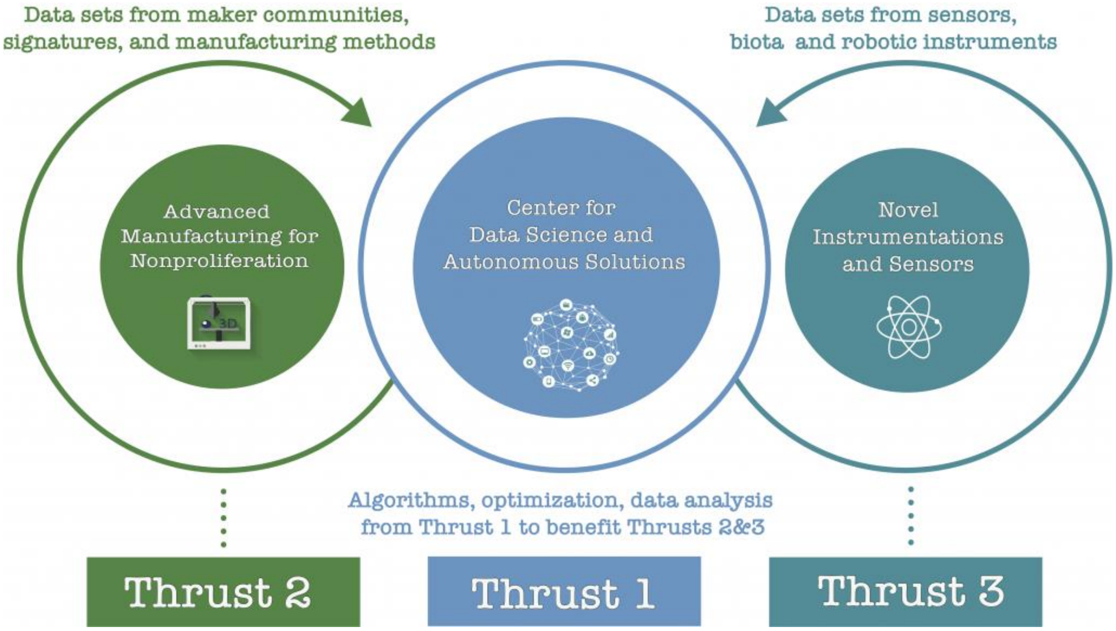
Data sets from maker communities, signatures, and manufacturing methods
Data sets from sensors, biota and robotic instruments
Advanced Manufacturing for Nonproliferation
3D
Center for Data Science and Autonomous Solutions
Novel Instrumentations and Sensors
Algorithms, optimization, data analysis from Thrust 1 to benefit Thrusts 2&3
Thrust 2
Thrust 1
Thrust 3

Figure 2. Three thrust areas

Course Description

Figure 3 includes the background of students that registered to attend the Data Science and Engineering Summer School. Approximately half of the students were graduate students. The next largest group was National Laboratory Researcher. One high school student participated in the school.

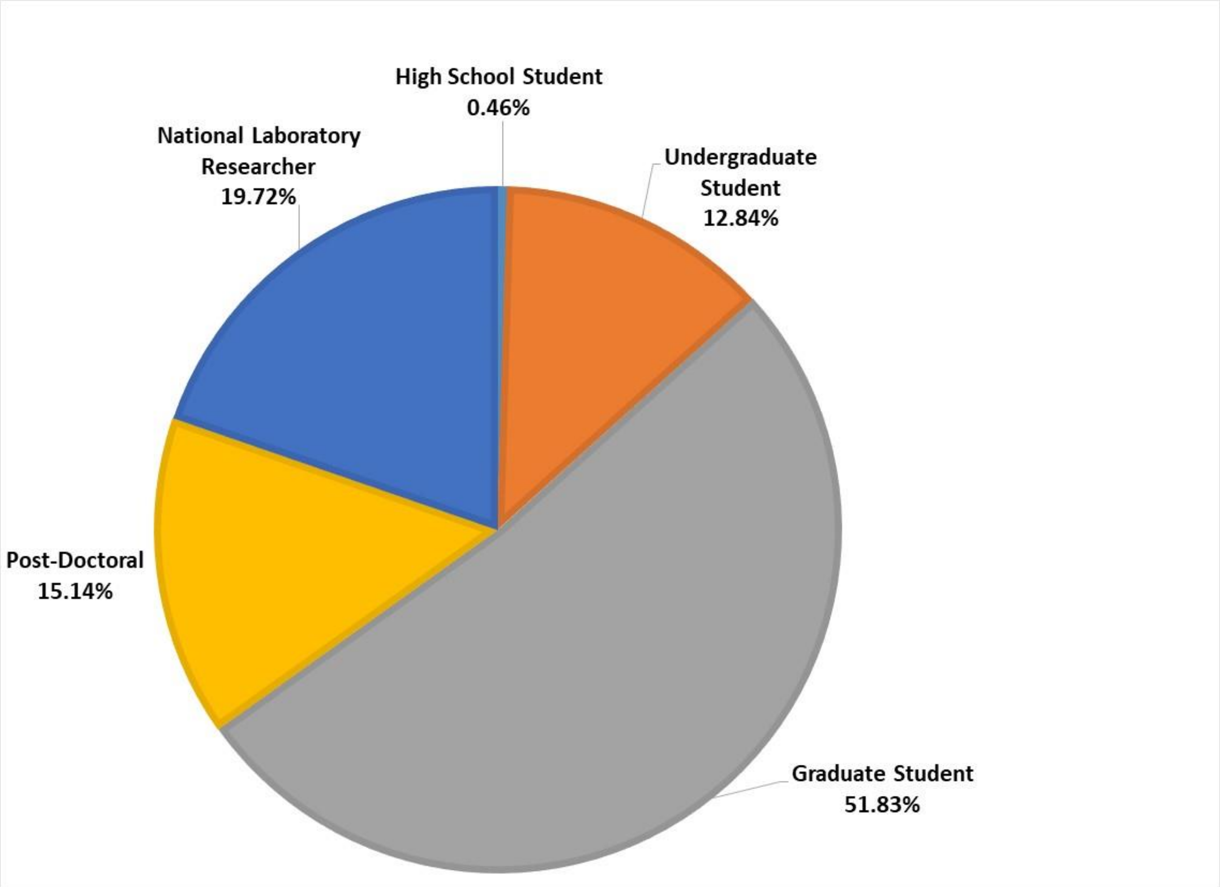

Figure 3. Distribution of students registered to attend the Data Science and Engineering Summer School

The Data Science and Engineering Summer School introduced students to nuclear nonproliferation challenges and introduced data science tools that are relevant to addressing these problems. Methods were showcased to the attendees through lectures delivered by experts from universities and national laboratories. Teaching modules offered introductions on data science methods and included hands-on illustrations and tutorials covering applications of data science and engineering for nonproliferation. Lectures included discussions on existing challenges in the field coupled with hands-on experiences. Modules were organized to boost students' knowledge of data science and hone their skills under constraints of a single week program.

course was delivered via the WebEx platform. Within WebEx, students were limited to utilize the chat tool for communication. Instructors also established a Slack channel to enhance student-to-student and student-to-instructor communication. Presentation files and data sets were uploaded to Google Drive for student access. Poll Everywhere was utilized my instructors so they could receive prompt feedback on discussions and example problems during a lecture. Extensive question and answer sessions were held at the end of each day and at the end of the week.

The Data Science and Engineering Summer School was divided into four modules. One of these modules were addressed on each of the first four days of the school. The last day of the school provided a summary, review exercises, and extensive discussion.

Module 1 - Fundamentals of Data Applications

This module will provided an introduction to basic statistical concepts and models. Fundamental distributions start with binomial processes and introduce a range of distribution types. Methods were introduced for testing distribution types. Data comparison methods are discussed leading to comparison methods that utilize associated uncertainty. An outline of this module is as follows:

1) Definitions
2) Statistical models
3) Testing of distribution types
4) Data comparison
5) Data comparison of data points and their associated uncertainty

Module 2 - Computational Machine Learning

This module covered the basic elements of machine learning with emphasis on the computational underpinnings of supervised learning. Prototype machine learning (ML) methods, including the k nearest neighbor (kNN) classifier and the perceptron, and optimization based methods, such as ridge regression, logistic regression, and support vector machines were discussed. Empirical risk minimization formed the backdrop for iterative optimization methods. Basic elements of deep learning for supervised classification, with emphasis on the convolutional neural network (CNN) were covered. While the course focused on computational aspects of ML, some probabilistic approaches to ML were also discussed. Python demos were used to illustrate the concepts on data drawn from diverse application areas. While not required, it was helpful if students had some familiarity with concepts of introductory probability & statistics, matrix linear algebra, multivariate calculus and numerical optimization. An outline of this module is as follows:

1) Brief introduction to ML concepts, prototypes, and cross-validation
2) Empirical risk minimization, optimization, iterative optimization
3) Linearly weighted predictors: ridge regression, logistic regression
4) Max margin classifiers: perceptron and support vector machine
5) Deep learning classifiers: multilayer perceptron and CNN.

Module 3 - Bayesian Modeling and Inference

This module presented multi-modal data fusion in the context of Bayesian modeling and inference. Bayesian modeling combined with structured probabilistic graphical models comprise a powerful framework for reasoning over complex heterogeneous data sources. Graphical models encode probabilistic relations via modular composable structures for data integration. They allow for both physics-based and learned sensor models, uncertainty and risk quantification, are robust to missing data, and support multiple reasoning tasks. This module provided the mathematical foundations while demonstrating the concepts using practical real-data examples and python notebooks. The use of Markov-chain Monte Carlo methods for inference was emphasized. The module assumed some student familiarity with probability, matrix linear algebra, random processes, Python, and Jupyter. An outline of this module is as follows:

1) Representations for inference uncertainty quantification, and risk
2) Probabilistic graphical models
3) Multi-modal data fusion
4) Introduction to Bayesian nonparametric models
5) Sensor data examples and exercises

Module 4 - Data Science for Safeguards

This module covered a variety of data science techniques applied to nuclear safeguards problems, with a specific focus on unsupervised learning methodology and natural language processing. Introductory concepts of nuclear safeguards and data science were introduced first. Unsupervised approaches covered include clustering, such as k-means and Gaussian mixture models, and dimension reduction techniques, such as principal component analysis and factor analysis. The section on natural language processing covered the entire workflow required for such tasks, including preprocessing of the text, numeric vectorization, and both supervised and unsupervised learning techniques. The module was geared towards researchers of all abilities but was more easily understood by those with some familiarity with optimization, matrix linear algebra, Python, and Jupyter notebooks. An outline of this module is as follows:

1) Introduction
    a. Safeguards 101
    b. Data Science 101
    c. Machine learning terminology
2) Unsupervised learning
    a. Clustering
    b. Dimensionality reduction
    c. Unsupervised deep learning
3) Natural language processing
    a. Preprocessing
    b. Numeric vectorization
    c. Analysis

Poll Everywhere

Poll Everywhere is a real-time engagement tool that was integrated into the Data Science and Engineering Summer School Power Point lectures. This allowed instructors to directly ask students questions via multiple formats including multiple choice, surveys, question & answers, and clickable images. Students provided their input through internet browsers or text messaging on their mobile phone. Results are shown live on the Power Point slide integrated within the presentation. Students could change their answers based on instructor feedback or may provide additional answers following class discussion.

Figure 4 shows the result of a simple multiple-choice poll where students were asked if the Zn data passed a statistical test. For this result, the correct answer was “No” and it was clear that most of the students responded correctly. However, the feedback showed that some students did not fully understand the statistical test results. The instructor was able to address the misunderstanding. Later in the lecture, a similar poll was provided for a different data set. In that case all the students responded correctly indicating that learning objectives were met for this part of the lecture.

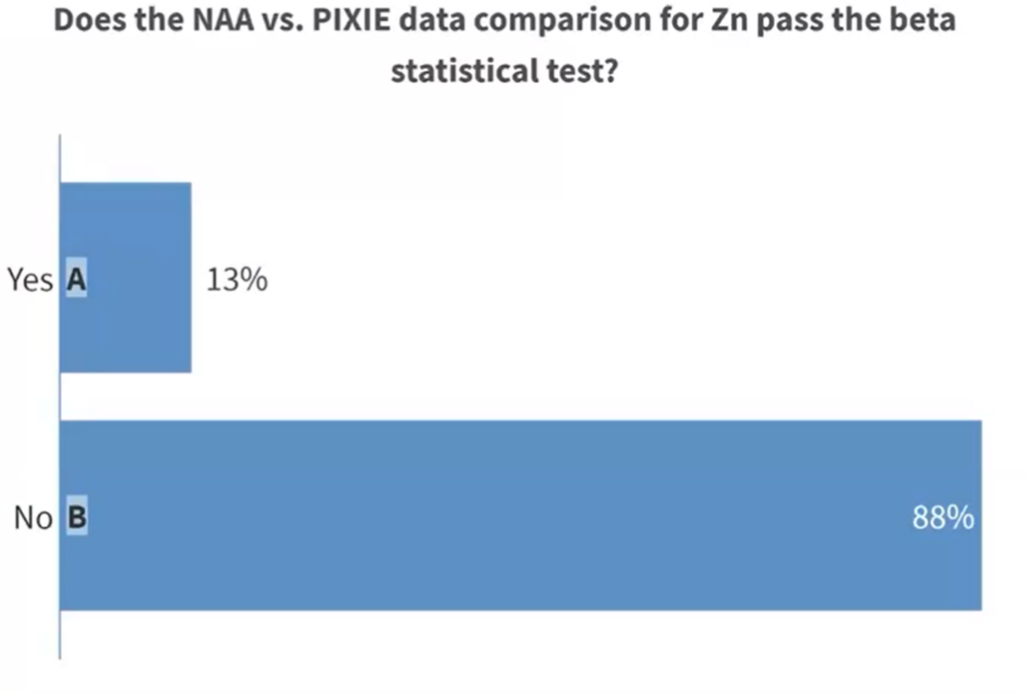

Figure 4. Poll Everywhere simple multiple-choice poll.

Figure 5. Shows another example of Poll Everywhere with multiple answers allowed for this question. Answers B, C, and D were all correct. The results show that no students selected A or E which were not correct. Since many students did not recognize B and D as valid answers, the feedback provided an opportunity for the instructor to explain the validity of all the correct answers.

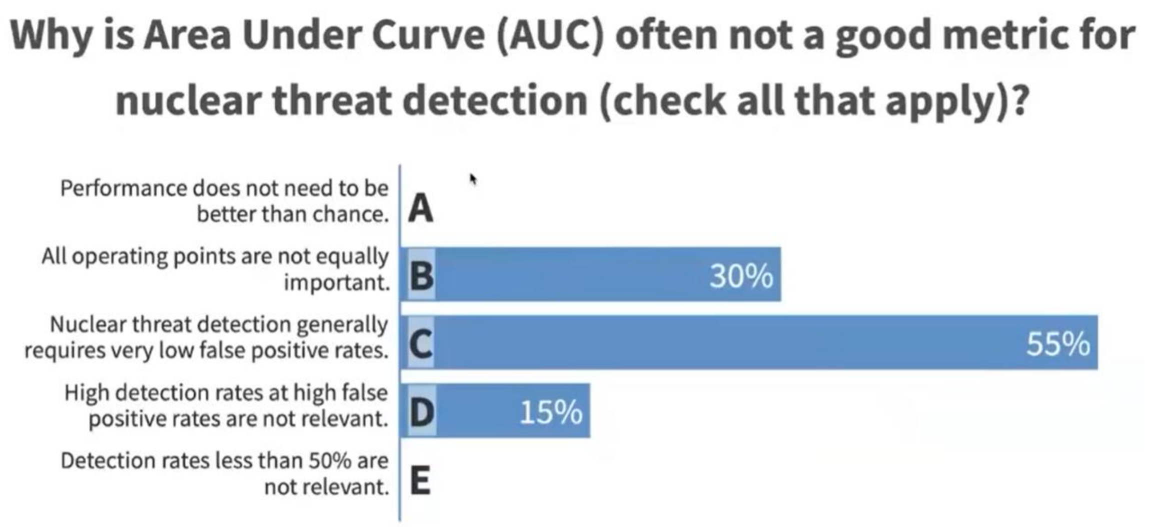

Figure 5 Poll allowing the selection of multiple answers.

Student Survey

Students were sent a survey at the end of the Data Science and Engineering Summer School. The results shown in Figure 6 indicate a high overall satisfaction with the course material and delivery methods. Students felt that the course met their expectations with informative lectures and problem sessions. On the negative side, students felt that it was difficult to maintain full engagement for an entire week of on-line lectures and discussion. There also appear to have been some difficulties properly targeting the highly varied range of student backgrounds.

Students overwhelmingly praised the utilization of Poll Everywhere within the lectures. The survey results showed that the students thought that Poll Everywhere was “fun”, “useful”, and “engaging.” The students suggested broadening the use of Poll Everywhere across a larger range of course modules. However, some students mentioned that they did not like how Poll Everywhere made them create their own account and login for use. These students would prefer to maintain their anonymity and privacy.

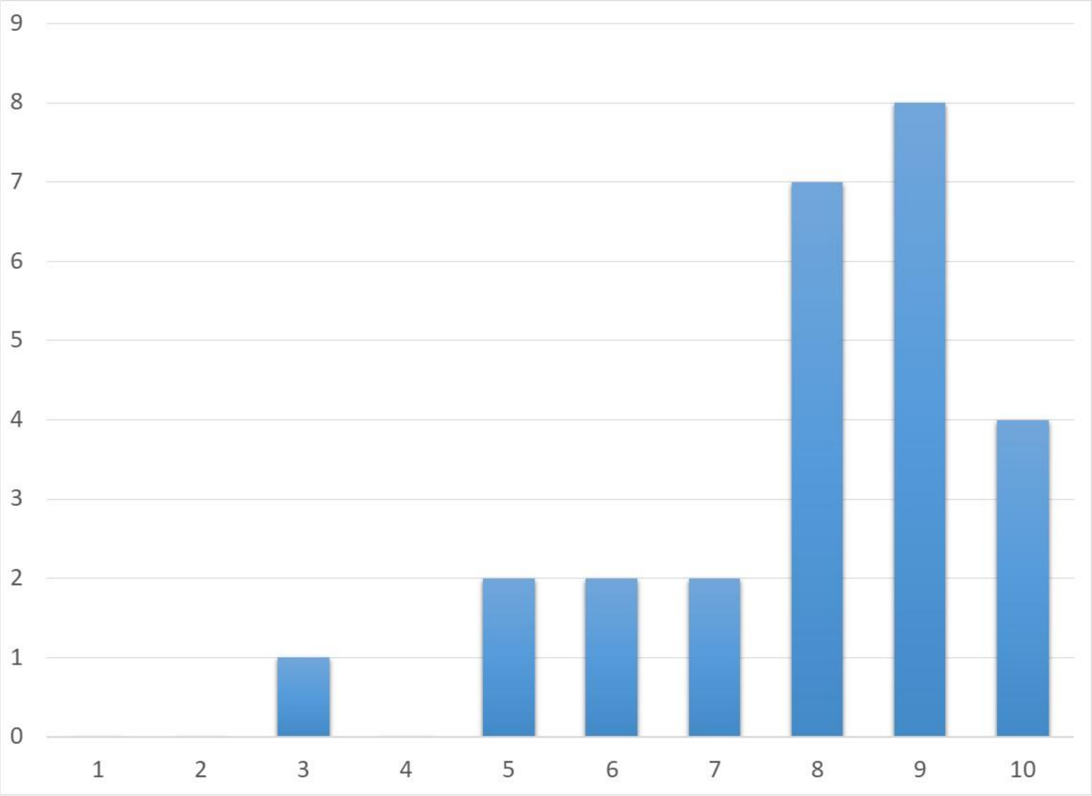

Figure 6. Student satisfaction survey result (1= low satisfaction, 10= very high satisfaction)

Conclusion

The ETI Data Science and Engineering Summer School was successfully delivered in an on-line format to an audience of over 200 students. All lectures from this class were recorded and are available via the ETI website (https://eti.gatech.edu/eti-annual-summer-school-2020/). It was a successful delivery with positive reviews from the students.

Many lessons were learned under the challenges of on-line course delivery to a large student audience required due to limitations in in-person education in 2020. One important lesson is that clear directions for access to the virtual platform and course material must be provided well in advance to the students. Another lesson is that it is very difficult to maintain engagement of students over a full week of on-line instruction. Paths for discussion and feedback during the course are essential. Methods to improve these interactions is necessary to further improve delivery of on-line workshops.

Poll Everywhere was successfully utilized. Students recommended expansion of this tool for future course delivery. Tools similar to Poll Everywhere are available, so an evaluation of options will have to be conducted prior to future utilization.

ETI will host two additional summer schools. The Novel Instrumentation Summer School is planned for 2021. The Advanced Manufacturing Summer School is anticipated to be held in 2022. While these future summer schools may be in-person, the lessons learned from this experience will be utilized for these future course deliveries.

References

1) Consortium for Enabling Technologies and Innovation (ETI), https://eti.gatech.edu/
2) Consortium for Monitoring Technologies and Verification (MTV), https://mtv.engin.umich.edu/
3) Consortium for Verification Technology (CVT), https://cvt.engin.umich.edu/
4) Consortium for Nonproliferation Enabling Technologies (CNEC), https://cnec.ncsu.edu/
5) Nuclear Science and Security Consortium (NSSC), https://nssc.berkeley.edu/

Acknowledgements

This material is based upon work supported by the Department of Energy / National Nuclear Security Administration under Award Number(s) DE-NA0003921.

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

LLNL operates under Contract DE-AC52-07NA27344.

Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.