

T3-CIDERS

A Train-the-Trainer Approach to Fostering CI- and Data-Enabled Research in Cybersecurity

(revised 2024-06-01)



Masha Sosonkina
***Wirawan Purwanto**
Peng Jiang
Mohan Yang
****Hongyi Wu**



CyberTraining #2320998, 2320999



*Corresponding author. Email: wpurwant@odu.edu

** Affiliation: The University of Arizona

Outline

- Background
 - Challenges in adequate CI skilling to meet the demands of cybersecurity research & education
 - Previous training development: DeapSECURE
 - Impacts & Lessons learned
- New train-the-trainer program — T3-CIDERS
- Current status and timeline
- Building partnerships — engage with us!

Cybersecurity's Demand for Advanced Computing

Increasing demand for CI skills for cybersecurity research & education

- The world's becoming more “cyber” and “AI” == greater security challenges
 - AI-generated attacks, deepfakes, ...
 - Bots of IoTs
 - Complex & vulnerable cloud / cyber-physical infrastructure
 - Widespread hardware and software vulnerabilities
- Simultaneous rapid development of AI and CI:
 - What used to be niche computing ~~will~~ has become mainstream!
 - New algorithms, new software, new hardware...
 - Are we training enough people with AI/CI skills??
 - Traditional academia too slow to adapt to fill the skill gap

DeapSECURE — A Review

Data-Enabled Advanced *Computational* Training Platform for Cybersecurity Research and Education

- **Focus:** CI (computational) techniques *applied* to cybersecurity research
- **Goals:**
 - Infuse CI techniques into cybersecurity research and education
 - Prepare students to embark modern cybersecurity research
- **Topics:** HPC, big data, machine learning, parallel programming, crypto systems
- **Target audience:** undergraduate and graduate students
- **Prerequisites:**
 - Interest in cybersecurity
 - Basic programming skills

CyberTraining #1829771
(2018-2022)



Training Approach

- Tailor CI materials toward novice
 - Goal: Teaching important salient points to be knowledgeable enough to begin practicing the techniques and to learn further on his/her own
 - *Experts are not produced instantly but through continuous learning and practice*
- Motivate CI techniques from state-of-the-art cybersecurity research:
 - Real research use cases as “storylines” for hands-on activities
 - Introduce practical tools (libraries, frameworks) used in real-world research
 - Adopt “concept-by-example” approach
 - Similar approach to the Carpentries
- Assessment as an integral part
 - PRE- and POST-workshop surveys; knowledge assessment; focus group interviews
 - Iteratively improve our training program based on the feedback

Materials Available for In-Person and Online Training

DeapSECURE's open learning resources:

- Six Carpentries-style online lessons (“e-textbooks”)
- Hands-on datasets and files (one set per lesson)
- Jupyter notebooks for self-paced learning (2–3 / lesson)

The number of threads shows a multimodal distribution (two major peaks and two smaller peaks). It is actually interesting to plot the histogram separately for the two applications:

```
nthrds_FB = df2[df2['ApplicationName'] == 'Facebook']['num_threads']
nthrds_WA = df2[df2['ApplicationName'] == 'WhatsApp']['num_threads']

seaborn.distplot(nthrds_FB, kde=False, label="Facebook")
seaborn.distplot(nthrds_WA, kde=False, label="WhatsApp")
plt.legend()
```

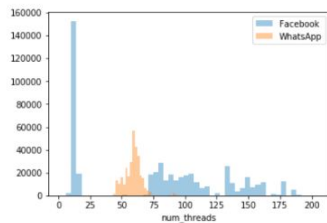


Figure: Histograms of `num_threads` grouped by application type


Characterizing Behavior of Different Applications

The last plot shows histograms of `num_threads` drawn for individual applications. A glance of this plot visual cues about the differences between the two applications being considered. Discuss the differences you can uncover using the histogram plot.


 Discussion

EXAMPLE 1: Dropping a useless feature. Create `df_mystery2` that does not have the `Unnamed: 0` field, which is not a feature at all.

```
[ ]: """Create new DataFrame which does not contain 'Unnamed: 0'.
Make sure to verify the result."""

#df_mystery2 = df_mystery.drop(#TODO, axis=1)

[ ]: df_mystery2 = df_mystery.drop(['Unnamed: 0'], axis=1)
df_mystery2.head()
```

EXAMPLE 2: Dropping all non-features. `ApplicationName` and `Unnamed: 0` are not features. Create `df_features_only` that has only features.

```
[ ]: """Create new DataFrame which does not contain 'Unnamed: 0' and 'ApplicationName'.
Make sure to verify the result."""

#df_features_only = df_mystery.#TODO
```

EXAMPLE 3: Dropping a useless feature forever. don't need to see it anymore.

Hint: This is an in-place operation which alters `df`.

```
[ ]: """Write a code to remove 'Unnamed: 0' co
```

Module 1: Introduction to HPC

Introduction to high-performance computing on a Linux cluster: UNIX shell interaction, SLURM job scheduler, parallel job launch. ([Lesson site](#))

Module 2: Dealing with Big Data

Introduction to Pandas, a powerful data processing framework capable of handling large amounts of data in an efficient manner. ([Lesson site](#))

Module 3: Machine Learning

Machine learning is an approach to program a computer to perform certain intelligent tasks without being explicitly programmed to do so. ([Lesson site](#))

Module 4: Deep Learning Using Neural Networks

Neural network is a powerful approach to machine learning that can yield extremely high accuracy on complex cognitive tasks. ([Lesson site](#))

Module 5: Cryptography for Privacy-Preserving Computation

Homomorphic encryption enables untrusted parties to perform computation while preserving the privacy of sensitive data ([Lesson site](#))

Module 6: Parallel and High-Performance Computing

<https://deapsecure.gitlab.io/lessons/>
<https://gitlab.com/deapsecure/>

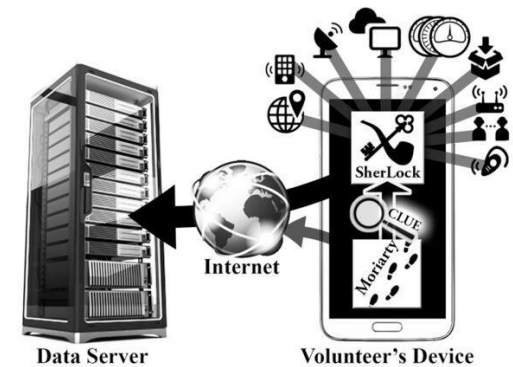
Training Contents and Format

DeapSECURE lessons are highly hands-on:

- Based on well-defined research problems
- Hands-on learning style is *similar* to the Carpentries
 - Teach the CI techniques on HPC while working toward the solution of the problem
 - Suitable for “participatory live coding” instructions
 - Some lectures still needed to teach the basic concepts of CI methods
 - Web-based lessons written in the style of Carpentries lesson
- Computer programming is required
 - Primary language: Python
 - UNIX shell required for working with HPC
- Using (mostly) standard libraries and tools:
 - Pandas, Matplotlib, scikit-learn, TensorFlow, Keras, mpi4py, Python-Paillier

Hands-on Based on Real-World Research Problems

- **Spam email analysis (HPC)**
 - Dataset: Untroubled spam collection (Bruce Guenter)
 - Task: Discover countries of origin based on email headers
 - Goal: Parallel-processing of many independent tasks
- **Privacy-preserving techniques with homomorphic encryption (CRYPT, PAR)**
 - Task: Encrypt data (numbers, images) and perform computations in encrypted form
 - Goal: Understand performance bottleneck(s) and cut down execution time by parallelization with MPI
- **Mobile device security (BD, ML, NN)**
 - Dataset: Sherlock smartphone dataset (Ben-Gurion Univ)
 - Task: Explore Sherlock “applications” data, identify patterns, create classifier for smartphone apps
 - Goal: Expose full cycle of data analytics and machine learning



Lessons and Hands-On Activities

Table 1: The DeapSECURE lesson modules (since Fall 2019)

Lesson Description	Hands-on Activities	Toolkits
Introduction to HPC and how to access, use and program HPC systems	Analyzing countries of origin from a large collection of spam emails; using parallel processing on HPC to speed up data processing	UNIX shell commands, SLURM
Advanced cryptography for privacy-preserving computation	AES ciphertext cracking; “King Oofy” privacy-preserving census; Paillier encryption of bitmap image data	AES-Python [19], Python-paillier [5]
Parallel programming with MPI	Parallelization of image Paillier encryption	mpi4py [6], Python-paillier
Big data (BD) analytics	Processing, cleaning, analyzing, and visualizing large Sherlock dataset	Pandas, Matplotlib, Seaborn
Machine learning (ML) modeling	Classification of smartphone apps based on system utilization data using classic ML methods	scikit-learn [14]
Neural networks (NN) for deep learning modeling	Building neural networks to classify smartphone apps	TensorFlow [2] and Keras [4]

Making Sense of Data with Visualization

Now that we have learn some basic skills on `Series` and `DataFrame`, let us try one more exercises to familiarize ourselves with the mystery of Sherlock data. We will also make use of visualization here; a latter episode will be dedicated to more in-depth visualization techniques.

Statistics of a Column

The `describe` operation also works on a `Series` object. Print the descriptive statistics of just the `CPU_USAGE` column.

Solution

Let us look more closely into the statistics of one column:

```
print(df_mystery['vsize'].describe())
```

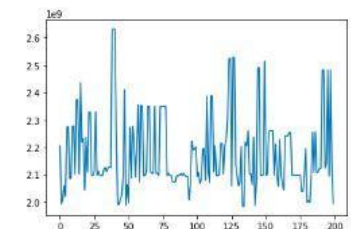
```
count    2.000000e+02
mean     2.175793e+09
std      1.417478e+08
min      1.983623e+09
25%      2.095809e+09
50%      2.107628e+09
75%      2.252854e+09
max      2.633568e+09
Name: vsize, dtype: float64
```

As we mentioned earlier, the quantities returned by the `describe` method collectively make up the *descriptive statistics* about this data. These numbers speak the range of the values, the mean and median, as well as the spread of the data around the mean and median values.

For many people, it is much easier to “see” these values in a graph. Therefore let us turn to visualization to help put the statistics in perspective. First let us plot the raw values of `vsize`. There is a built-in plotting utility in `pandas` to do so.

```
df_mystery['vsize'].plot(kind='line')
```

This statement asks `pandas` to print an (x-y) plot; by default two nearby data points are connected by lines. On a Jupyter notebook, the plot will be displayed inline when the code above is executed:



In-Person Workshop Format (2018–2020)

Six workshops throughout an academic year, or a week-long summer institute

- Each workshop:
 - 30-minute research presentation by a cybersecurity researcher
 - 2.5 hours of intro to a CI topic (some lectures, mostly hands-on)
- Hands-on intro to CI methods:
 - Platform: ODU's Turing HPC cluster with Intel CPUs & NVIDIA GPUs
 - Goal: Applying CI methods on HPC to work out research problems
 - “Participatory live coding” teaching style
 - Some exercises done in small groups
 - Workshop teaching assistants (WTAs) available assist learners
- Attendance: between 11–32 per workshop
- Comprehensive assessment: perception, opinion (feedback), knowledge questions

In-Person Workshop Experience Recap

- Participant's feedback / follow-up:
 - Hands-on activities as a valuable aspect of the workshop
 - Learned new technologies, methodologies, software tools, and computational resources
 - A few students pursued to take courses in HPC
- Challenges:
 - Participant attrition toward the end of semester
 - Very limited time to teach contents adequately
 - Students have difficulty following instructors on command-line interface
 - Large differences in participants' entrance skills and knowledge

Conversion to Online Training

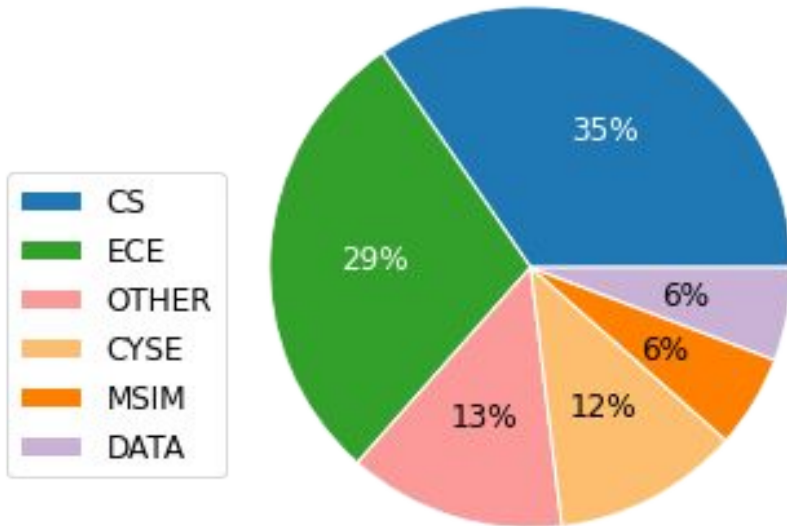
- Challenges with online workshops:
 - No indirect feedback from learners; very little direct feedback
 - Need to accommodate different learning styles & speeds
- Key adaptation strategies:
 - Using Jupyter notebooks for most modules (except HPC)
 - Abbreviated version of the web-based DeapSECURE lessons
 - Mixed text / codes to support self-paced learning
 - Learners work through series of code cells to complete and execute them
 - 2–3 notebooks per lesson
 - Identifying critical concepts and hands-on experiences
 - Focus on delivering these during the workshops
 - Additional details left for learners to pursue on their own

Online Workshop Format (2020–2021)

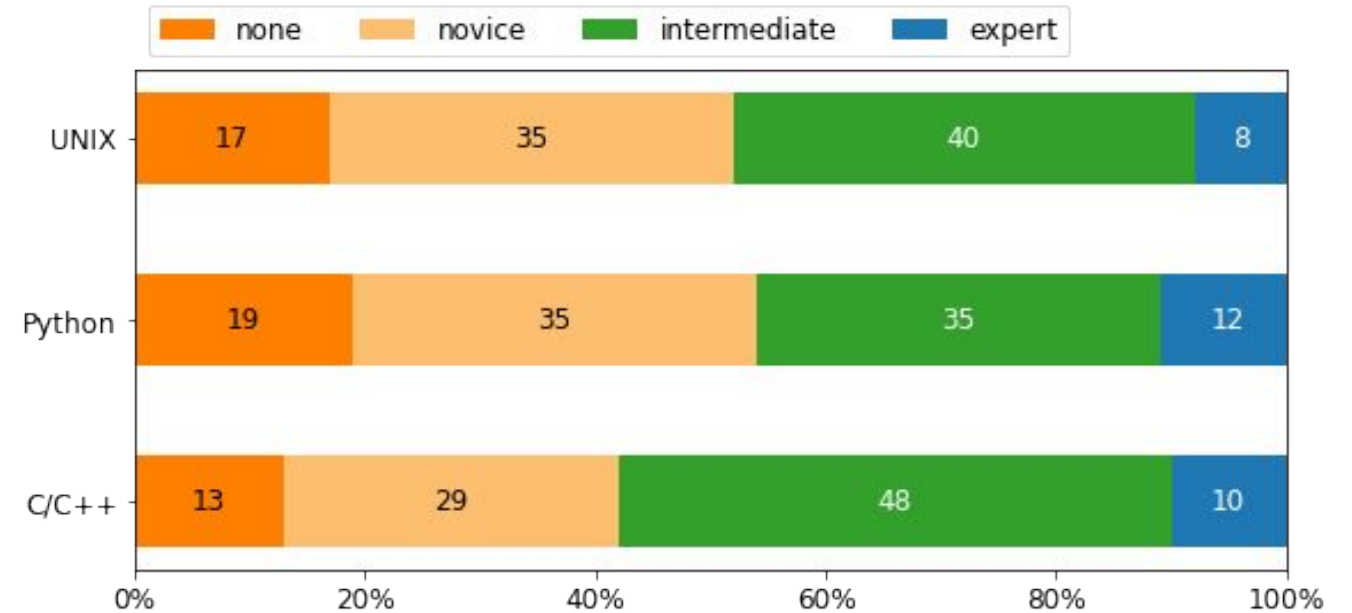
- Each workshop:
 - Three 1-hour sessions
 - Each session contains a brief lecture, followed by a hands-on breakout session
 - Avoid prolonged time on one activity; minimize Zoom fatigue
 - Kahoot quizzes were used to engage learners and break silence
- Hands-on intro to CI methods:
 - Platform: ODU's Wahab HPC cluster + Open OnDemand interface
 - Utilizing Zoom *breakout room* for hands-on
 - Target: 5–6 participants per *room*, led by one WTA
 - “Participatory live coding” teaching aided by Jupyter notebooks
- Attendance: 7–23 per workshop

Assessment: Online Workshop 2020–2021

Academic Majors



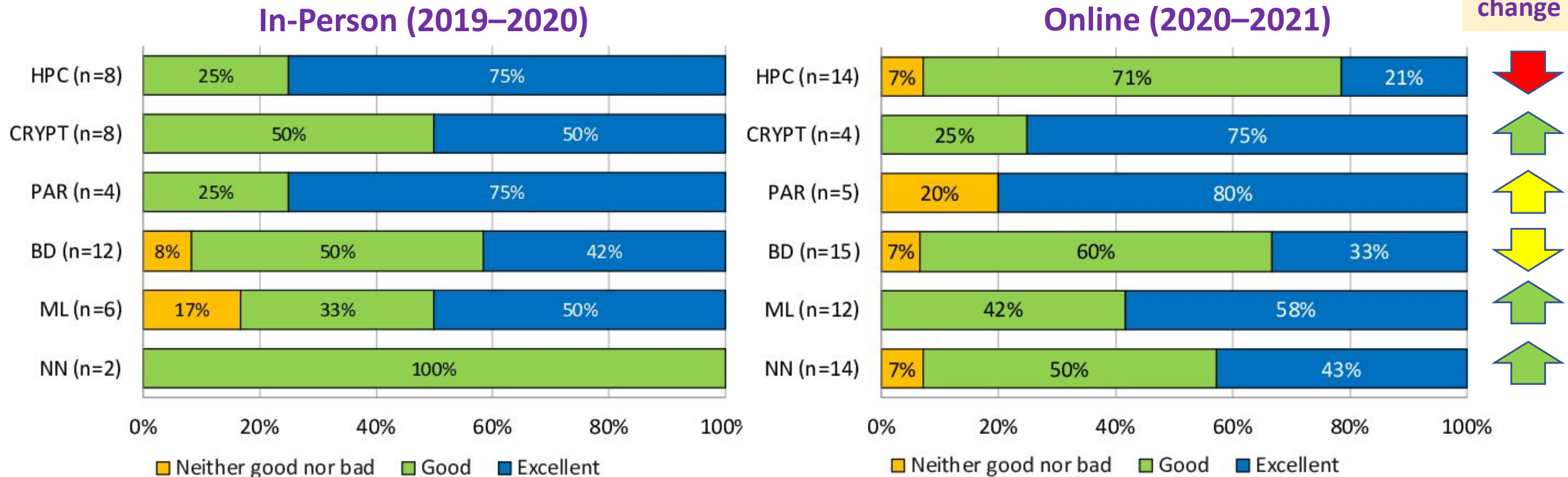
Programming Skills (self-eval)



- Over 75% came from computer-related fields (CS, ECE, Cybersecurity [CYSE])
- Programming skills split 50/50 between none/novice and intermediate/expert

Assessment: Online Workshop 2020–2021

PERCEPTION: OVERALL RATINGS



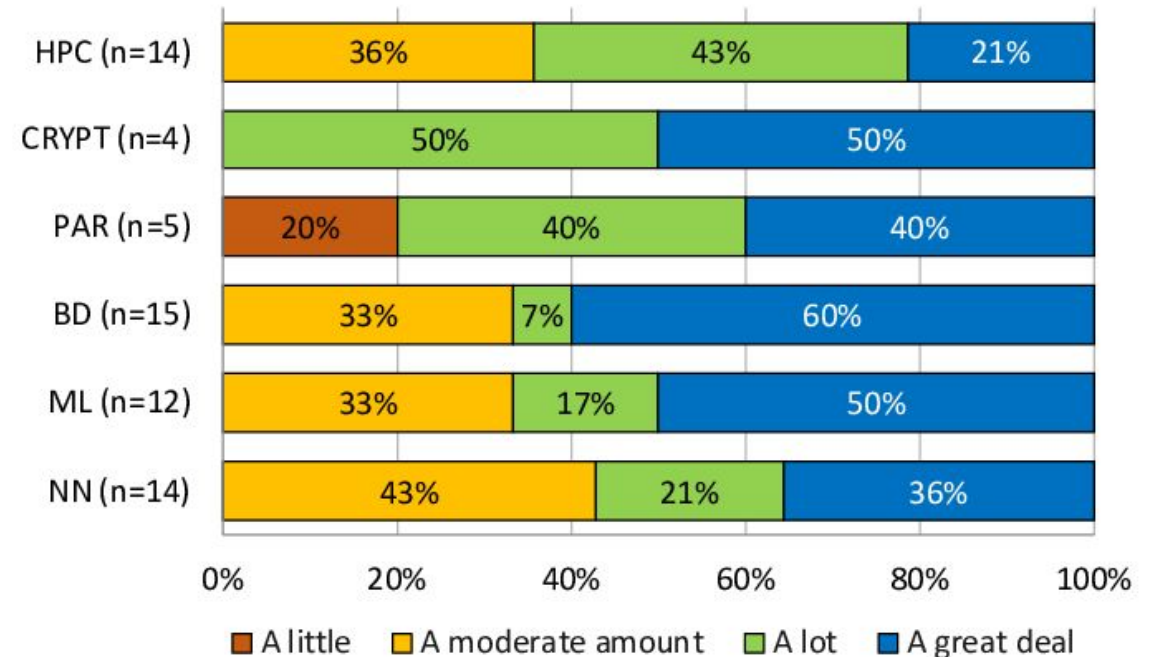
- Ratings approximately similar in both modes --- Improved lesson vs. tougher remote learning?
- HPC rating dropped: tough to teach CLI-based topic (UNIX-shell intensive)?
- Somewhat consistent patterns: PAR > CRYPT ; ML > BD

Assessment: Online Workshop 2020–2021

PERCEPTION: HOW MUCH HAVE YOU LEARNED?

Online (2020–2021)

- HPC & PAR are challenging topics
 - HPC: CLI-based nature?
 - PAR: advanced topic for most learners
- BD > ML > NN
- Reasons for “not enough learning”?
 - Contents too difficult or too easy?
 - Issues in delivery?
 - Analysis of knowledge questions may shed light on this matter (upcoming).



Assessment: Online Workshop 2020–2021

PERCEPTION: MOST VALUABLE AND LEAST VALUABLE ?

Most Valuable	% respondents
CI topics	40
Hands-on	27
Jupyter & tools	23
Instruction & TAs	13

Least Valuable	% respondents
Challenging pace / insufficient time	7
Difficulty of materials	3

- Value #1: CI topics
- Value #2: Hands-on learning on Jupyter
- Pace of the materials and length of time are still an issue

Overlapping responses: Sum of % respondents not to add up to 100%

DeapSECURE Impacts & Lessons Learned

- Intro to CI methods with hands-on components was well received
- Over 4 years: 3 workshop series, 2 summer institutes, 2 pilot online workshops
- 200 students (35 from outside ODU)
- Week-long “Summer Institute” format works best
- Online hands-on workshops can work with proper adaptation
 - In-person format still more effective
 - How to make online workshops more engaging?
 - What are ways to improve overall learning experience?
- Programming prerequisites are important
 - Suggest learners to take UNIX shell and/or Python programming lessons (from SwC or others) prior to joining DeapSECURE workshops

Open Source Release & Community Adoption

- All six lessons are now open-source! <https://gitlab.com/deapsecure/>
 - License compatible with the Carpentries (CC-BY-4.0).
 - Additional developer's repos will be released in due time.
- Toward Community Adoption
 - Virginia-wide pilot online workshop (Fall 2021; 50–60 attended)
 - Virginia-wide “Community Interest Survey”
 - Responses: 9 total from faculty, staff, administrators
 - Recommending workshops to students: 8
 - Teaching DeapSECURE lessons: 3
 - Modify & customize lessons: 5
 - Contribute & further develop: 4

Can we scale up the training effort & impact?

Scaling Up CI Training

- **Scaling up \neq massive online training**
 - Pure online training lacks social / communal aspect of learning!
 - Broadening participation to smaller / minority-serving institutions
 - Variety of learners need adaptations of training materials, approach, emphasis, etc.
 - **Producing trainers** is the way to go to scale up the training to meet the diverse needs!
- Needs a community to further develop and sustain the training
 - Need to produce *both* learners *and* trainers
 - Broad input and perspective will make better training and community!
- Virginia has a cyber research/innovation community!

Scaling Up CI Training

- Virginia has a cyber research/innovation community!
 - **Commonwealth Cyber Initiative** (cyberinitiative.org)
 - **Vision:** “To establish Virginia as a global center of excellence in cybersecurity research and serve as a catalyst for the commonwealth's economic diversification and long-term leadership in this sector.”
 - **Mission:** “To serve as an engine for research, workforce development, and innovation at the intersection between cybersecurity, autonomous systems, and intelligence.”
 - Funded by the Commonwealth of Virginia
- Basic CI trainings such as DeapSECURE will enable more students and researchers to take on cyber-related research projects relevant to CCI



Seeding T3-CIDERS: Lesson Developer's Training

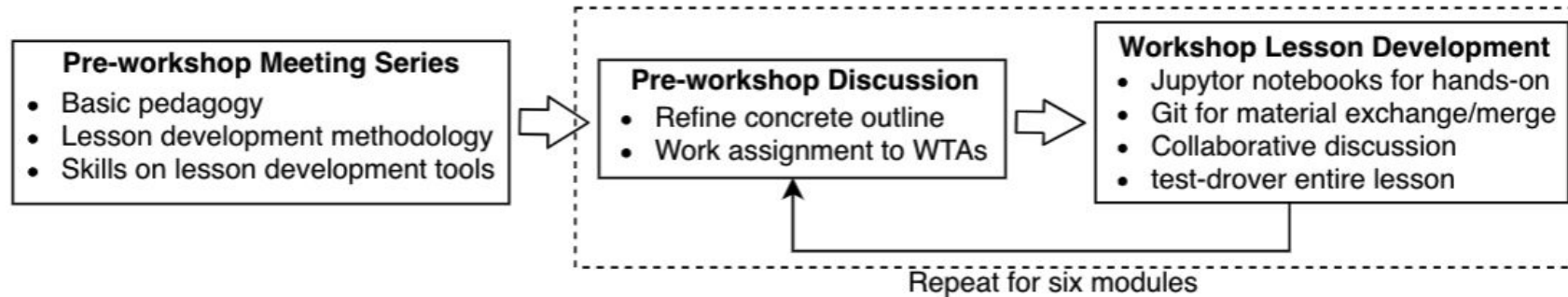


Figure 2: Overall process of the lesson developer's training.

- Training workshop TAs to become lesson developers *and* trainers
 - Weekly meetings (2-hour long)
 - Discuss lesson goals, structure (outline), hands-on activities
 - Collaborative platform and tools [Git/Gitlab, Jupyter notebook, Jekyll]
 - Test-drive teaching & hands-on

T3-CIDERS Overview

A Train-the-Trainer Approach to Fostering CI- and Data-Enabled Research in Cybersecurity

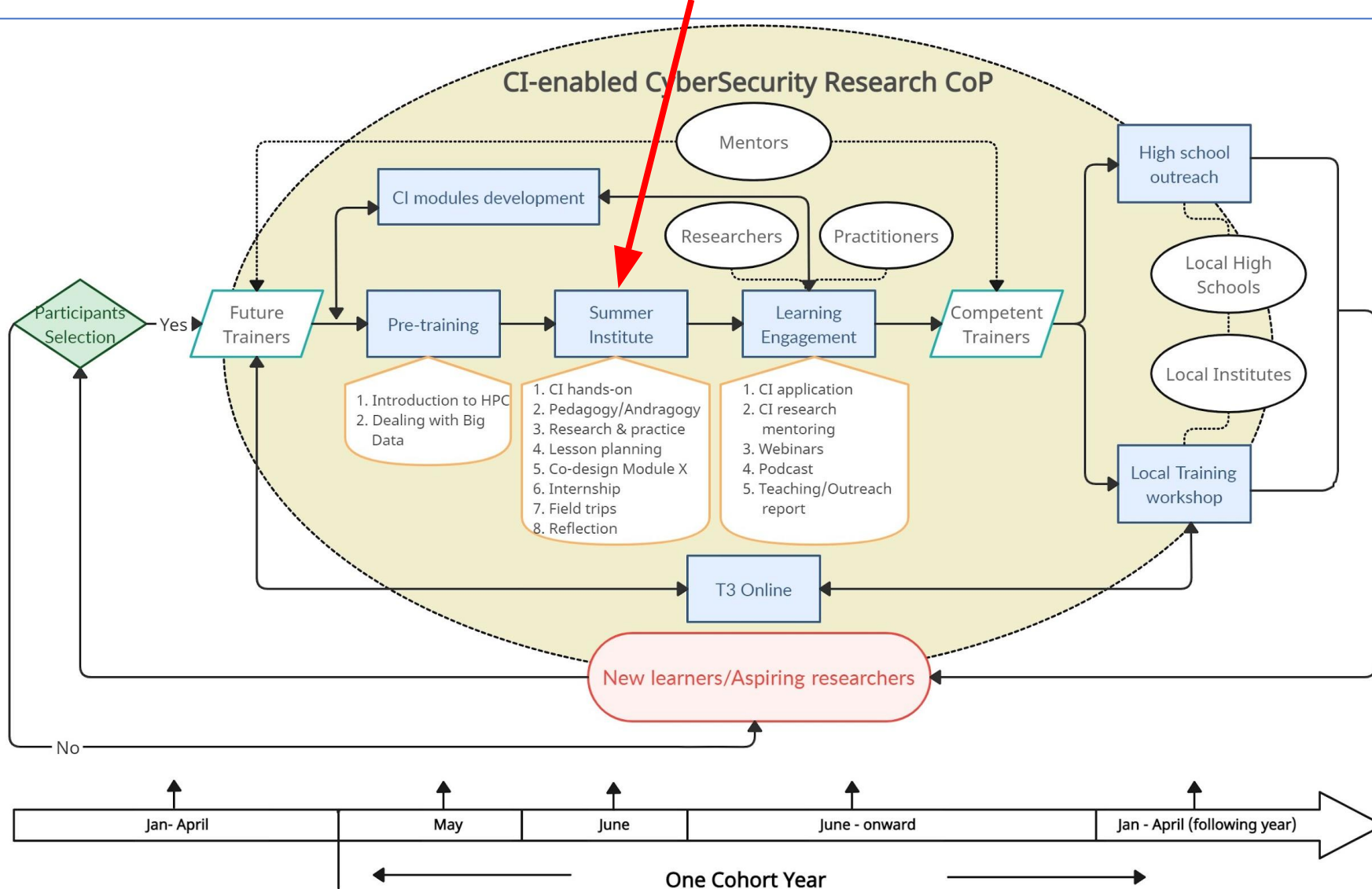
- **Overarching goal:** Fostering a CI-enabled cybersecurity research community of practice to accelerate state-of-the-art research & development in cybersecurity and related fields
- **Means:**
 - Preparing competent trainers to broaden utilization of advanced CI
 - Leveraging & further improving DeapSECURE lesson materials
 - Create additional training materials as needed
- **...all driven by the cybersecurity research community!**

CyberTraining #2023998–9
(2023-...)



Schedule & Timeline

2024 Summer Institute: Jul 29 – Aug 2 (ODU)



Eligibility and Benefits

- **Who can join?**

- Faculty
- Researchers
- Practitioners with interests in cybersecurity, cyber-related fields
- Graduate students, if paired with a faculty member

- **Benefits:**

- Access to hands-on training modules in advanced CI relevant to cybersecurity and cyber-related research (DeapSECURE)
- Learn evidence-based training and design methodology
- Obtain support for your training activities in your local institutions
- Join a CI-enabled community of practice
- Receive monetary assistance for participation

CyberTraining #2023998–9
(2023-...)



Partner with Us!


#1 Most Important!

Join our Interest List:
tinyurl.com/t3ciders-signup

Promote our 2024 program
(click on flyer on the right)

Website: sites.wp.odu.edu/t3-ciders/

Email: t3ciders@gmail.com



T3-CIDERS

A Train-the-Trainer Approach to Fostering CI- and Data-Enabled Research in Cybersecurity

Cyberinfrastructure (CI) such as advanced computers, big data, and artificial intelligence are game changers in cutting-edge cybersecurity research. Our Train-the-Trainer program will equip you to teach computational and data competencies to embark on state-of-the-art research in cybersecurity!

Project Overview

Accelerate state-of-the-art research & development in cybersecurity and related fields by:

- Preparing competent trainers to broaden utilization of advanced CI
- Fostering a CI-enabled cybersecurity research community of practice

What will you learn?

T3-CIDERS will introduce effective training and instructional design methods, to prepare you to teach advanced CI domains such as:

- high-performance computing
- big data
- machine learning
- cryptography
- parallel programming

Who can join?

- Faculty
- Researchers
- Practitioners with interest in cybersecurity, cyber-related fields
- Graduate students, if paired with a faculty member

Summer Institute dates:
July 29-August 2, 2024
Old Dominion University
Norfolk, VA

Benefits:

- Access for hands-on training modules in advanced CI relevant to cybersecurity and cyber-related research
- Learn evidence-based training and design methodology
- Obtain support for your training activities in your local institutions
- Join a CI-enabled community of practice
- Receive monetary assistance for attendance

Train-the-Trainer Outcomes:

- Apply CI techniques in cybersecurity research
- Develop & conduct introductory CI trainings for local academic communities.

Join our Interest List:
tinyurl.com/t3ciders-signup

Visit our website:
sites.wp.odu.edu/t3-ciders/

The T3-CIDERS training program is a collaborative project of Old Dominion University and the University of Arizona, funded by the U.S. National Science Foundation CyberTraining grants #2320998 and #2320999.

Registration for Cohort 2024 Is Open!

Program Schedule

Pre-training (virtual, synchronous / asynchronous)	July 2024
Summer Institute (in person, ODU campus)	July 29 – Aug 2, 2024
Learning engagements (virtual, monthly)	Sept 2024 – Jul 2025
Local training activity (at least once, format TBD)	Sept 2024 – Jul 2025

For more details:

<https://sites.wp.odu.edu/t3-ciders/cohort-2024/>

Registration Form



https://odu.co1.qualtrics.com/jfe/form/SV_8FYKx10afzcpeR0



Summary & Acknowledgments

- T3-CIDERS is more than train-the-trainers!
 - Community building
 - Producing CI-competent trainers & practitioners
- We want your participation!
 - Promote T3-CIDERS
 - Join interest list
- Project site: sites.wp.odu.edu/t3-ciders/
- Contact us: t3ciders@gmail.com

Funding & support:



NSF/OAC #2320998-9

Graduate/under assistants (2023-2024):

Kristin Herman, Nolan Lovett,
Dorothy Parry, Jiawei Chen, Chunyu Hu, Kayla J. Curtis