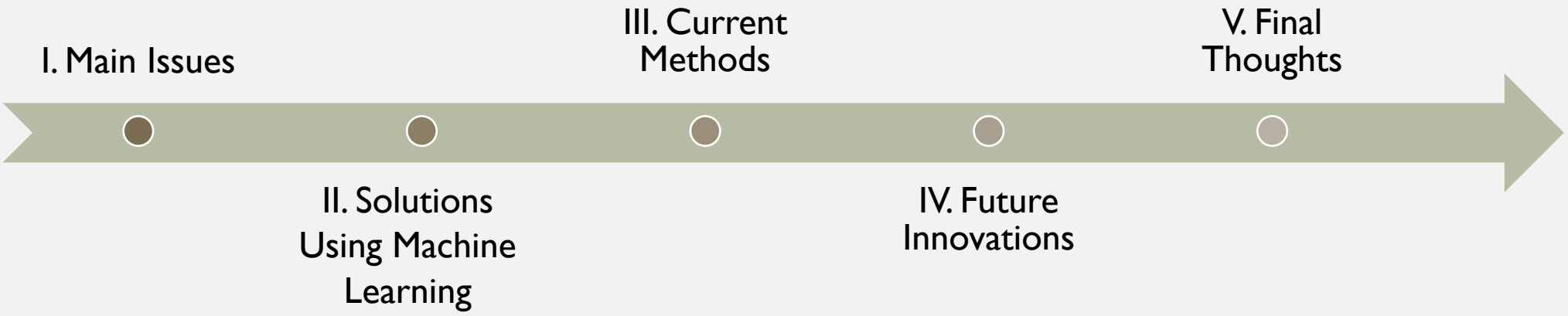
CYBERSECURITY USING MACHINE LEARNING

Current Approaches and Future Innovations

Sabrina Slattery | Saint Leo University

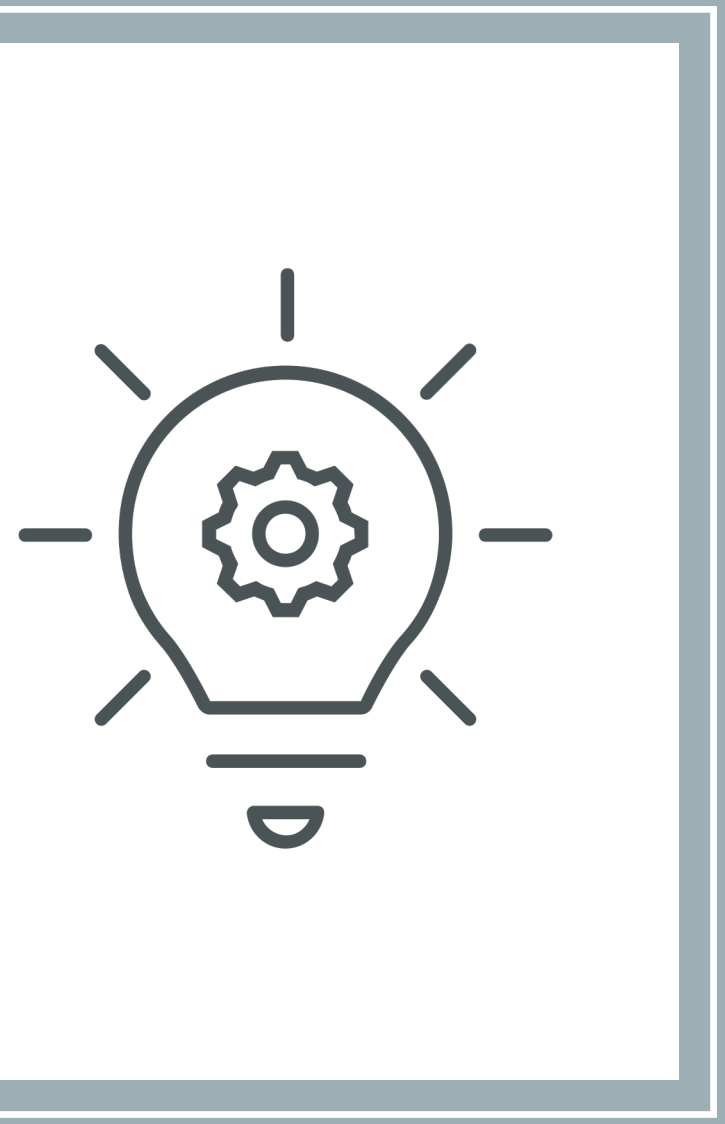
OVERVIEW

We will explore current methods being used in cybersecurity and future advancements being anticipated in the field of ML-based cybersecurity



I

MAIN ISSUES



DATA BOOM

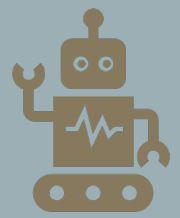
Internet access has increased from 413 million to **3.4 billion** over the course of 16 years

In 2017, the average internet user encountered around 312.5 megabytes of data each day –

the equivalent of a dictionary's worth of text

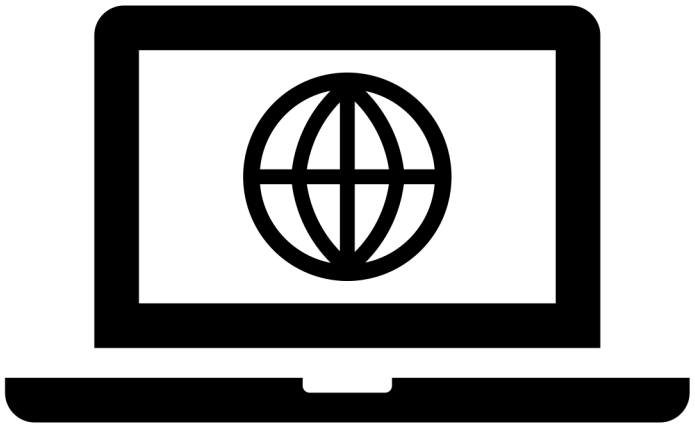
EVERYDAY TECHNOLOGY EITHER USE OR PUT OUT DATA

RELIANCE ON THESE DEVICES, TECHNOLOGIES, AND THE NETWORKS THAT CONNECT THEM, MAKE USERS **INCREASINGLY MORE VULNERABLE** TO CYBERATTACKS





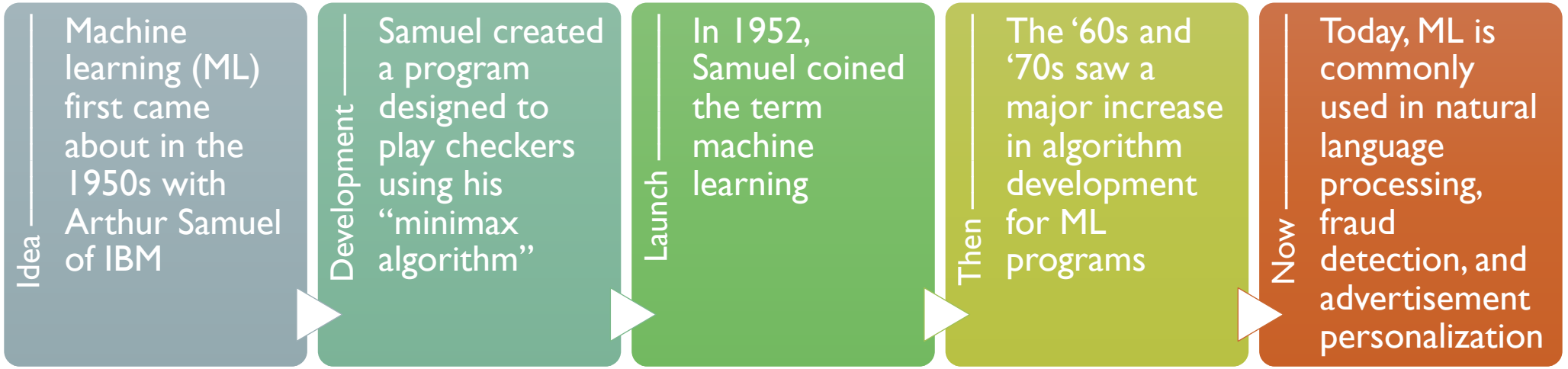
The amount of data being shared and stored on the Internet brings many concerns, but most important is :



II

SOLUTIONS USING MACHINE LEARNING

BRIEF HISTORY OF MACHINE LEARNING



Machine Learning (ML) – a field of computer science which focuses on training computer systems through algorithms to make future predictions based on previous learning

- A sudden influx of data over the past decade is vulnerable for cyberattacks
- The amount of data to protect leaves cybersecurity teams overwhelmed
- Because ML programs can be automated, this lessens the amount of human labor needed
- Augmenting a software or device's security with ML methods almost always **increases** the amount of information that is protected

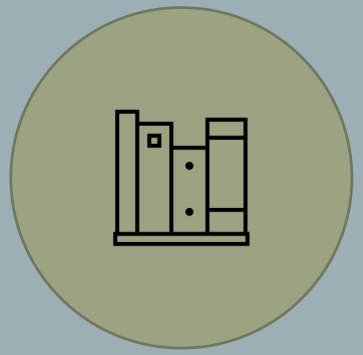
CYBERSECURITY WITH ML TECHNIQUES

Cybersecurity using ML is a relatively new area of study, with only 20 to 30 years of research to back it

ML methods have the potential to vastly improve network security, data security, and cloud security so that users can rely more on safely storing their data and using their devices

III

CURRENT APPROACHES IN CYBERSECURITY



Study I

Research on machine learning method and its application technology in intrusion information security detection

Study II

Performance Comparison and Current Challenges of Using Machine Learning Techniques in Cybersecurity

Study III

Assisting in Auditing of Buffer Overflow Vulnerabilities via Machine Learning

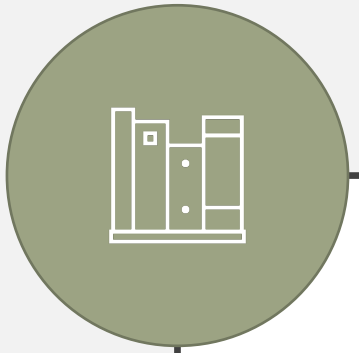
Study IV

Game theory approach for detecting vulnerable data centers in cloud computing network

NEED TO KNOW

- These studies use numerous algorithms and methods in their research
- The most important algorithms:
  - K-Nearest Neighbor (KNN)
  - Support Vector Machine (SVM)
  - Random Forest (RF)
- The most important defense systems:
  - Intrusion Detection Systems (IDS)

A Definition of Terms will be available at the bottom of each slide containing these important terms



STUDY 1

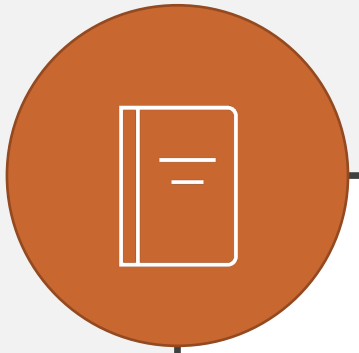
Strengthening Information System Security

- Used a mix of SVM* neighbor and Elman neural network (ENN) algorithms as an IDS method
- Proposed that SVM + ENN algorithms will patch each other's shortcomings
  - This method will work to solve and strengthen network security in information systems

Takeaway:

SVM + ENN would make a significant improvement in the security of information systems, with an **average of 20% greater intrusion detection rate** over similar, solo acting algorithms

Support Vector Machine (SVM) – a popular classification algorithm used in machine learning; defines support vectors, or global data outliers, and work to lower the amount of misclassification errors by increasing the dataset's margin from decision boundaries



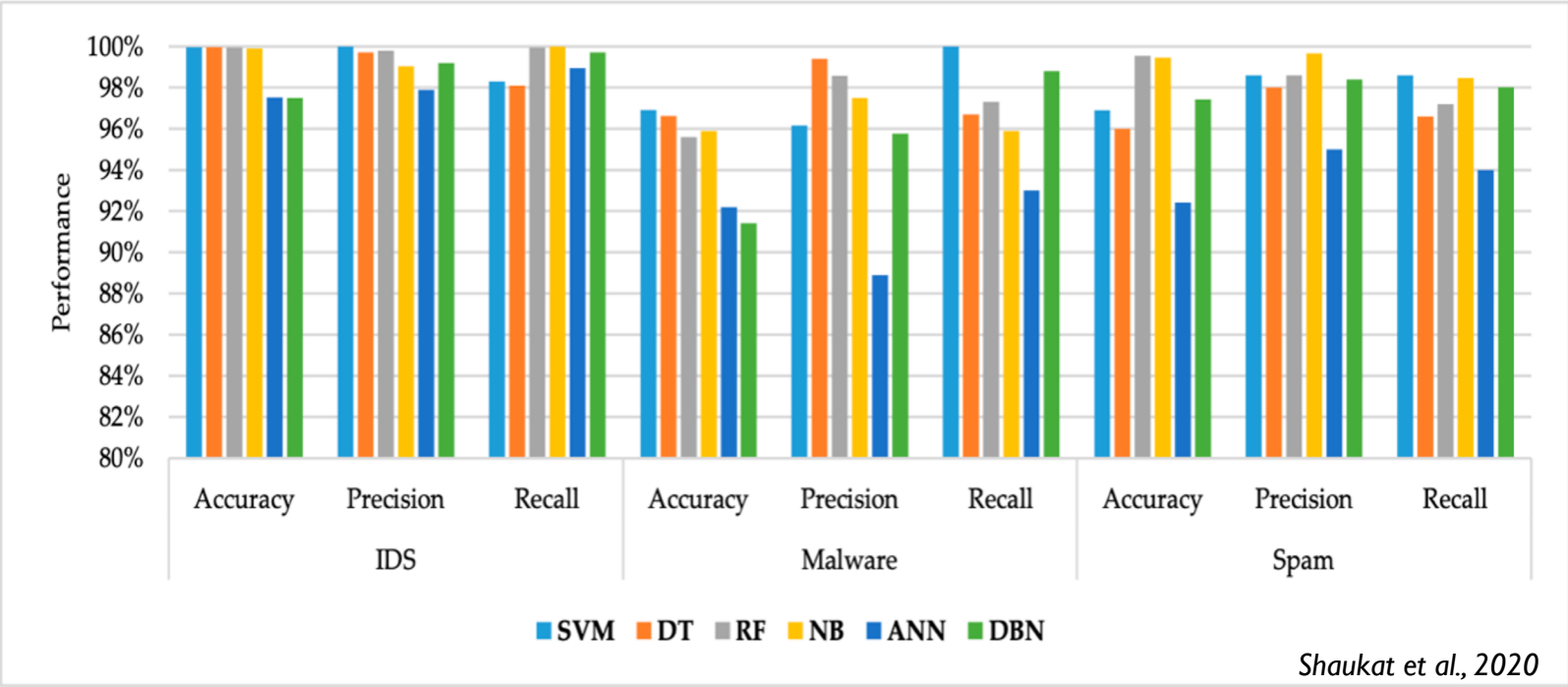
STUDY II

Which Popular Area of Security Vulnerabilities Needs Work?

- Analyzed 6 different ML methods on detecting 3 separate kinds of cyberattacks
- The six different ML methods were used on commonly trained security datasets collected by DARPA, or the U.S. Defense Advanced Research Projects Agency

A PERFORMANCE EVALUATION OF MACHINE LEARNING TECHNIQUES

- This figure demonstrates three traits
  1) Accuracy
  2) Precision
  3) Recall
- Includes six ML methods
  1) SVM*
  2) Decision Tree (DT)
  3) Random Forest (RF)*
  4) Naïve Bayes (NB)
  5) Artificial neural network (ANN)
  6) deep belief network (DBN)
- Tested on three categories of cybersecurity
  1) Intrusion Detection System (IDS)*
  2) Malware Detection
  3) Spam Classification.



Shaukat et al., 2020

Random Forest (RF) – a popular classification algorithm used in machine learning; draws a random bootstrap sample of a random size, grows a decision tree from that sample and applies randomly chosen attributes to the tree, picking the most applicable attribute to describe the decision tree – this process is repeated *k* amount of times until the final sample within the training set is categorized

Support Vector Machine (SVM) – a popular classification algorithm used in machine learning; defines support vectors, or global data outliers, and work to lower the amount of misclassification errors by increasing the dataset's margin from decision boundaries

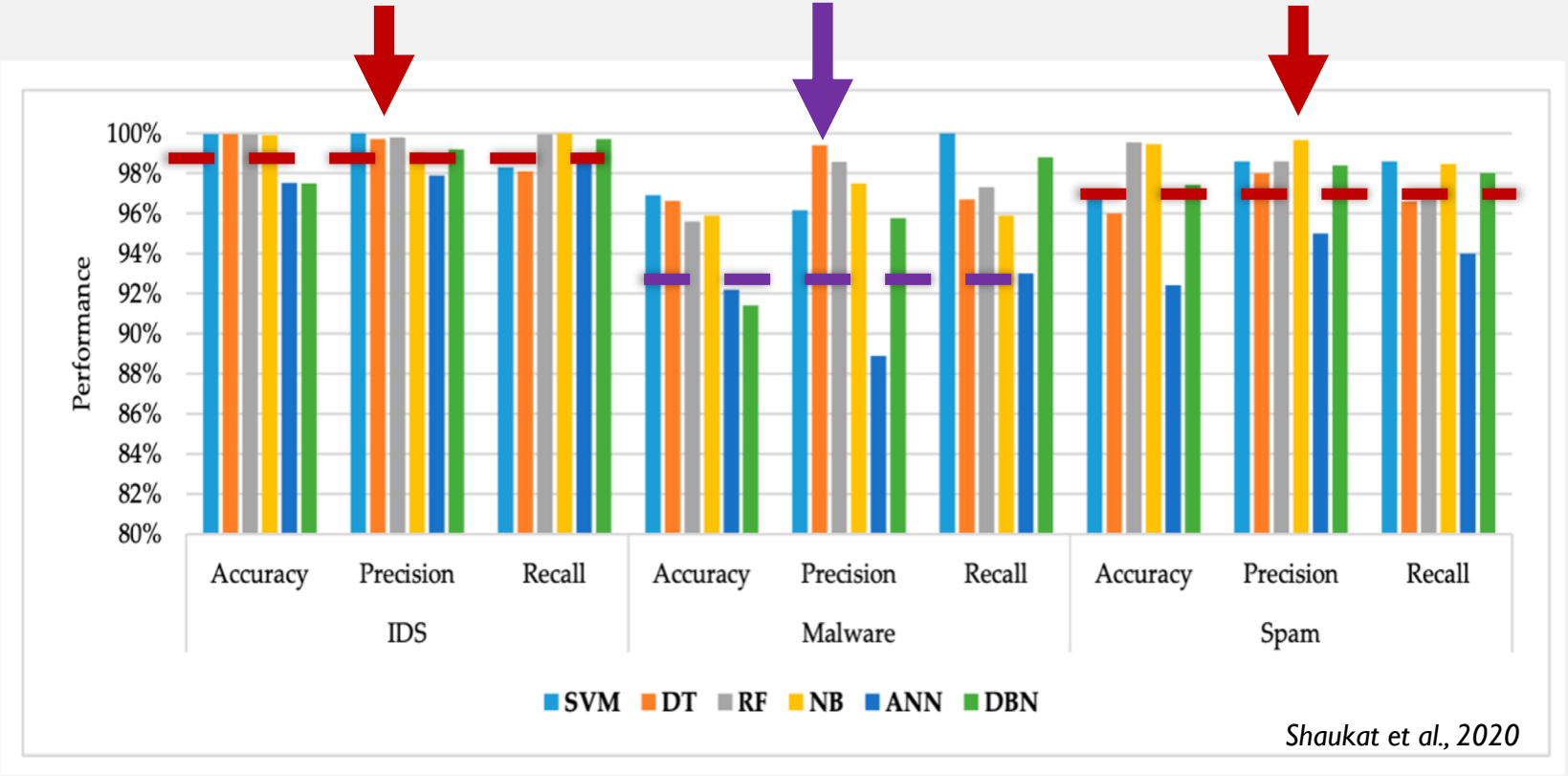
Intrusion Detection System (IDS) – a software application that is able to police networks and systems for threats and malicious activity

A PERFORMANCE EVALUATION OF MACHINE LEARNING TECHNIQUES

Shaukat et al. found that current ML methods were most effective against **IDS** and **spam** detection
(avg margin of 1.5% and 3%)

ML methods were least effective with **malware** detection
(error margin of 7-8%)

Takeaway:
Malware lacks the training data it needs to be effective against attacks and needs more testing to be efficient



Shaukat et al., 2020

STUDY III

Weakest Links

- Buffer overflow vulnerabilities have been around since the 1980s
  - When vulnerabilities in a program's memory buffer cause a program to overwrite old memory with new data once the buffer has reached its data capacity
- Tested a program which would narrow down the range of data testing needed and reduce manhours spent fixing the problem
- Their method included running source code through five different classifier algorithms: KNN*, decision trees, Naïve Bayes, AdaBoost, and SVM*

Takeaway:
It is important to focus on both new and pre-existing threats and vulnerabilities – the number of man hours are significantly reduced when using their method

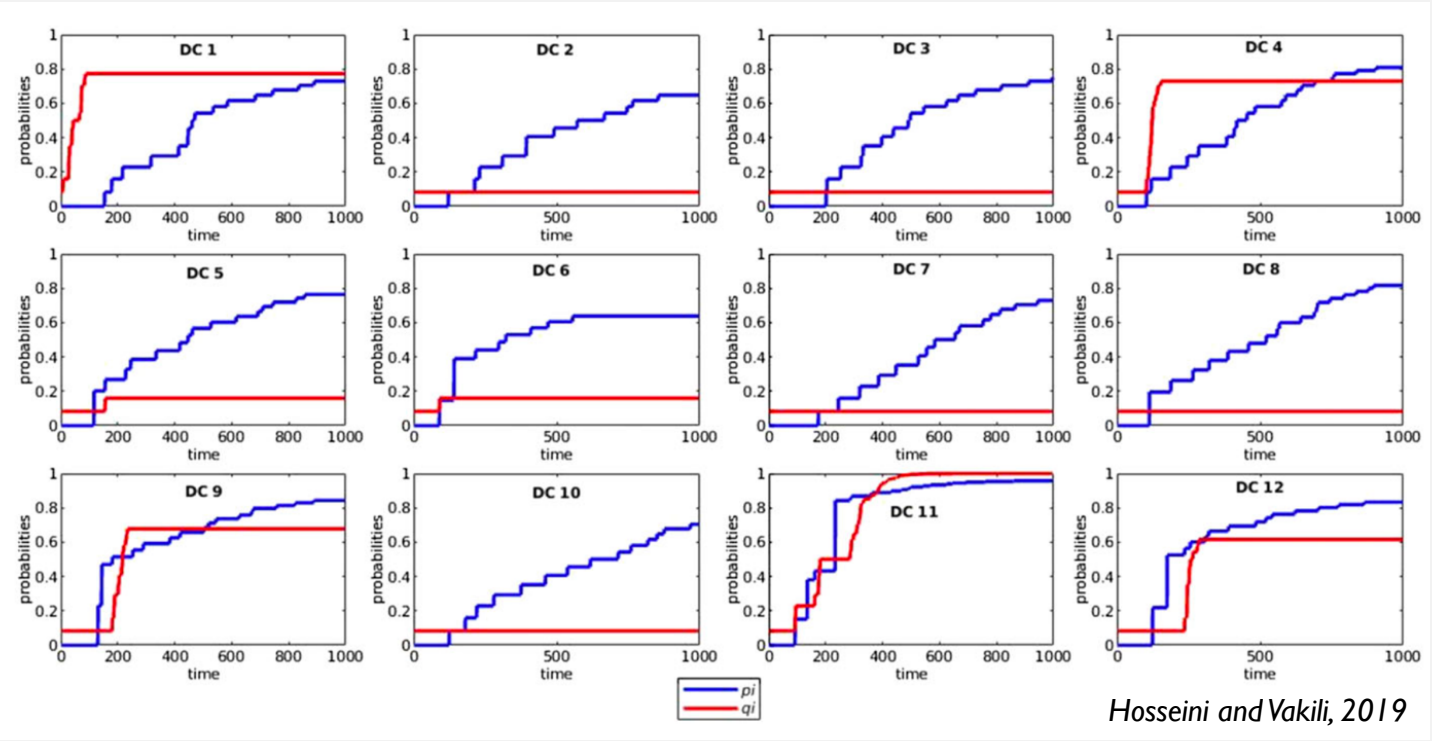
K-Nearest Neighbor (KNN) – a popular classification algorithm used in machine learning; randomly defines a distance metric and selects a feature from the dataset, grouping *x* amount of 'neighbor' values and labeling the sample based on the group's primary classification; does not require training before predictions are made (non-parametric)
Support Vector Machine (SVM) – a popular classification algorithm used in machine learning; defines support vectors, or global data outliers, and work to lower the amount of misclassification errors by increasing the dataset's margin from decision boundaries

STUDY IV

Cybersecurity in Cloud Computing

- industry concerns over the overall newness of cloud computing networks and the amount of cyberattacks being directed at the sensitive data contained within these networks
- The authors propose a specific ML technique called game theory, which they argue will be able to predict attacks on data centers before they occur
  - Game theory is a branch of applied mathematics that models the interaction between two or more players in a given situation with preestablished rules

PROBABILITY OF USAGE (PI) AND FAILURE (QI) FOR EACH DATA CENTER

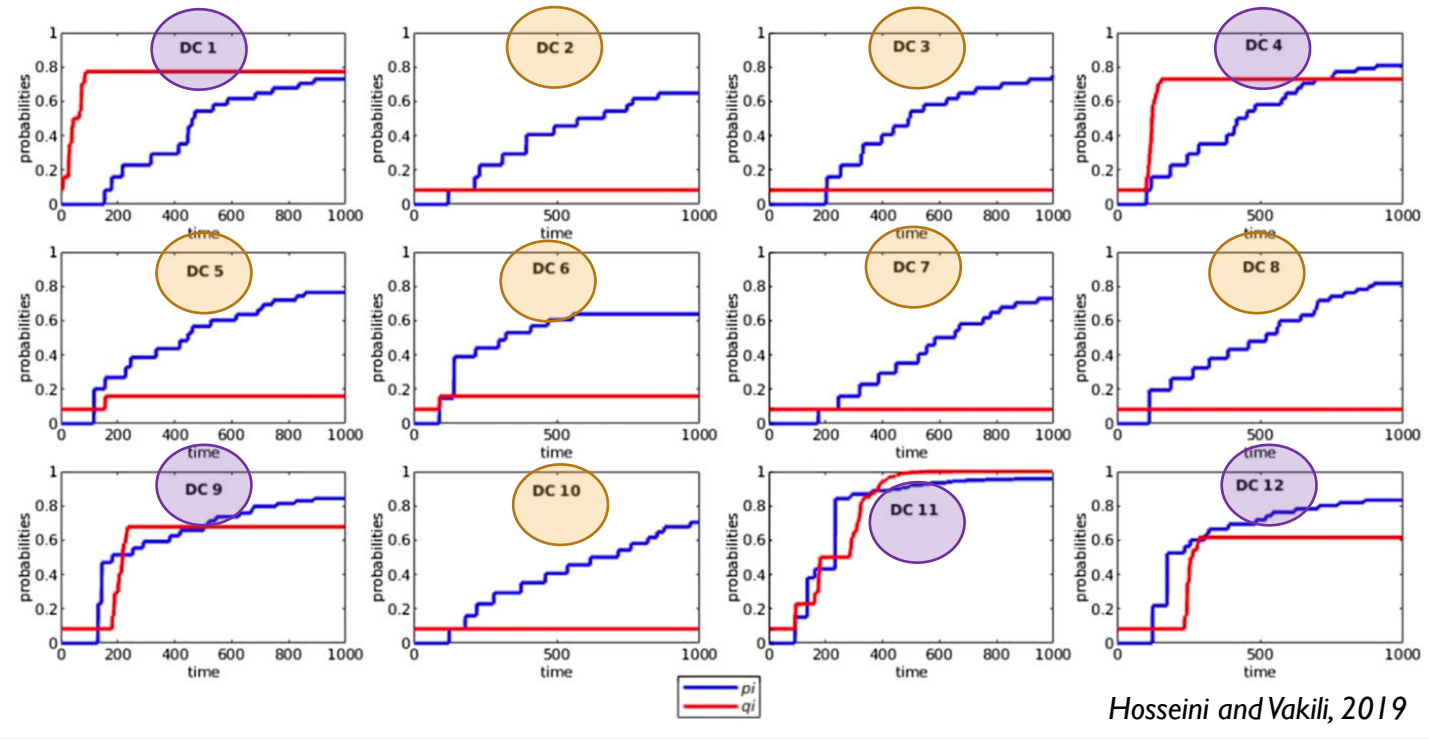


Hosseini and Vakili, 2019

This figure shows the predictive probability of usage (*pi*) in blue, and the probability of failure (*qi*) in red

Tested on 12 data centers in a given cloud computing network

PROBABILITY OF USAGE (PI) AND FAILURE (QI) FOR EACH DATA CENTER



Hosseini and Vakili, 2019

On average

data centers which were approached with a sudden burst in usage had a greater probability of failure than those data centers which saw a gradual increase in usage over time

Takeaway:

Cloud data centers need to regulate the intake of data over time to avoid malicious attacks on their network

IV

FUTURE INNOVATIONS

Study V

IntruDTree: A Machine Learning Based Cyber Security Intrusion Detection Model

Study VI

Malware Detection in Self-Driving Vehicles Using Machine Learning Algorithms

STUDY V

"IntruDTree"

Researchers are creating a work-in-progress ML intrusion detection model called IntruDTree

Model is user interactive, meaning the model is able to take into account a ranking of security features according to their importance

IntruDtree is a low-complexity algorithm

Shows a graphical representation of possible security flaws

Shows probability of occurrence for each threat and can do so with an accuracy of **98%**

Testing

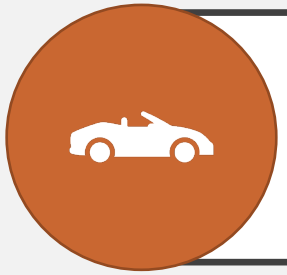
Compared the effectiveness of their model against other commonly used machine learning models, such as KNN* and Naïve Bayes, in terms of precision, recall, accuracy, and 'fscore'

Results

model surpassed all four common models reviewed in every category

Takeaway:

Possible use of this method in large-scale data centers, such as Cloud storage warehouses



STUDY VI

Self-Driving Cars

Vehicles with self-driving capabilities and autopiloting software have been shown to be prone to cyberattacks and remote control through hacking, putting drivers and passengers in potential danger

Method

Researchers took the point-of-view of a hacker and used an Android OS to perform malicious attacks on self-driving software

Results

The authors found that by using the novel score-function model, they were able to detect malware in real-time (~0.049 seconds) and had an accuracy of 92.9% with their chosen algorithm

Takeaway

Cars with self-driving software would greatly benefit from this remote-controlled machine learning algorithm detecting malware in real time as an added defense

V

CONCLUSION

Problem

Progress

Solution

Malicious attacks on databases and networks far outnumber the amount of individual cybersecurity specialists and methods used against these attacks.

Given the tools that machine learning provide, predicting attacks before they happen would provide those with data at risk a sense of greater security.

SOLUTION

Research predicting optimistic accuracies of 98% and higher, the cybersecurity industry can look forward to stronger defenses and faster response times within the next few years.

QUESTIONS?

REFERENCES

DNS. (2017). Domo Resource - Data Never Sleeps 5.0. Retrieved November 2, 2020, from https://www.domo.com/learn/data-never-sleeps-5?aid=ogsm072517_1&sf100871281=1

Fang, W., Tan, X., Wilbur, D., Elhoseny, M., & Yuan, X. (2020). Research on machine learning method and its application technology in intrusion information security detection. *Journal of Intelligent & Fuzzy Systems, 38*(2), 1549–1558. https://doi-org.saintleo.idm.oclc.org/10.3233/JIFS-179518

Foote, K. (2019, March 26). A Brief History of Machine Learning. Retrieved November 23, 2020, from https://www.dataversity.net/a-brief-history-of-machine-learning/

Hosseini, S., & Vakili, R. (2019). Game theory approach for detecting vulnerable data centers in cloud computing network. *International Journal of Communication Systems, 32*(8), N.PAG. https://doi-org.saintleo.idm.oclc.org/10.1002/dac.3938

Jiang, L. (2020). *Week 4: A tour of machine learning classifiers using scikit-learn*. Lecture notes. Principles of Machine Learning COMP411, Franklin University. Delivered 20 Oct 2020.

O'Dea, S. (2020). *Number of smartphone users in the U.S. 2010-2023*. Retrieved November 2, 2020, from https://www.statista.com/statistics/201182/forecast-of-smartphone-users-in-the-us/#:~:text=Smartphone users in the United States 2018-2024&text=This statistic shows the number,estimated to reach 275.66 million.&text=Advances in telecommunication technology have been significant in recent years.

Purplesec. (2020, November 18). Cybercrime up 600% due to Covid-19 pandemic. Retrieved November 23, 2020, from https://purplesec.us/resources/cyber-security-statistics/

Mahbod, B. (2020, July 9). *Overview and Introduction to Software Security*. Lecture presented at Notre Dame de Namur University CIS 2148, Belmont.

Meng, Q., Feng, C., Zhang, B., & Tang, C. (2017). Assisting in Auditing of Buffer Overflow Vulnerabilities via Machine Learning. *Mathematical Problems in Engineering*, 1–13. https://doi-org.saintleo.idm.oclc.org/10.1155/2017/5452396

Park, S., & Choi, J.-Y. (2020). Malware Detection in Self-Driving Vehicles Using Machine Learning Algorithms. *Journal of Advanced Transportation*, 1–10. https://doi-org.saintleo.idm.oclc.org/10.1155/2020/3035741

Roser, M., Ritchie, H., & Ortiz-Ospina, E. (2015, July 14). Internet. Retrieved November 2, 2020, from https://ourworldindata.org/internet#:~:text=Globally the number of internet,online for the first time.

Ross, D. (2019, March 08). Game Theory. Retrieved November 24, 2020, from https://plato.stanford.edu/entries/game-theory/

Sarker, I. H., Abushark, Y. B., Alsolami, F., & Khan, A. I. (2020). IntruDTree: A Machine Learning Based Cyber Security Intrusion Detection Model. *Symmetry (20738994), 12*(5), 754. https://doi-org.saintleo.idm.oclc.org/10.3390/sym12050754

Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., Chen, S., Liu, D., & Li, J. (2020). Performance Comparison and Current Challenges of Using Machine Learning Techniques in Cybersecurity. *Energies (19961073), 13*(10), 2509. https://doi-org.saintleo.idm.oclc.org/10.3390/en13102509