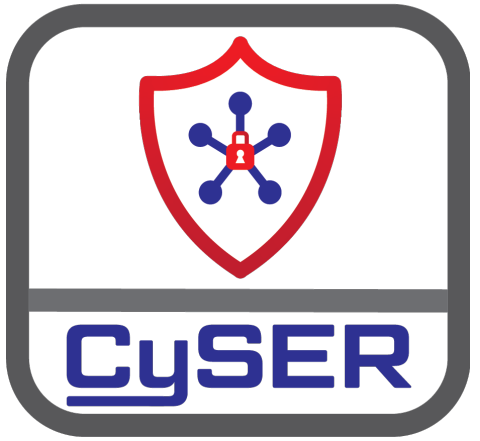
Northwest VICEROY Institute for Cybersecurity Education and Research (CySER): An Overview + A Peek into Machine Learning Research in CySER



Assefaw Gebremedhin

School of Electrical Engineering and Computer Science

Washington State University

CySER Virtual Seminar Series

October 11, 2021

What is CySER?

- An Institute funded by the Department of Defense Air Force Command through the VICEROY initiative
  - VICEROY = Virtual Institutes for Cyber and Electromagnetic Spectrum Research and Employ
  - VICEROY Institutes are managed by the Griffiss Institute
- Directly responds to the VICEROY call
  - Training ROTC and DoD-aligned civilians in cybersecurity at the undergraduate and graduate level, with primary emphasis on undergraduate
- Builds a strong consortium in the Pacific Northwest for cybersecurity education and research
  - CySER brings together 5 institutions with complementary strengths and diversity of populations served
- Seeks to position WSU to attain Center of Academic Excellence in Cyber Operations (CAE-CO) designation
  - Designation conferred by National Security Agency
  - Requirements: 10 Mandatory and 10 (out of 17) Optional Knowledge Units

CySER: Institutions and People

Washington State University (WSU)

- Bernard Van Wie (VSCBE; Lead PI)
- Assefaw Gebremedhin (EECS; Co-PI; Research Lead)
- Noel Schulz (EECS; Co-PI; Industry Lead)
- Venera Arnaoudova (EECS; Co-PI; CS Curriculum)
- Olusola Adesope (Education; Evaluator)
- Partha Pande (EECS; SP)
- Haipeng Cai (EECS; SP)
- Robert Crossler (MISE; SP)
- Jana Doppa (EECS; SP)
- Arda Gozen (MME; SP)
- Larry Holder (EECS; SP)
- Chris Hundhausen (EECS; SP)
- John Miller (EECS; SP)
- Gabriel Nketah (Project Coordinator)

WSU/UI ROTC

- Lt. Col. Nicholas Jeffers
- Major Paul Hyde

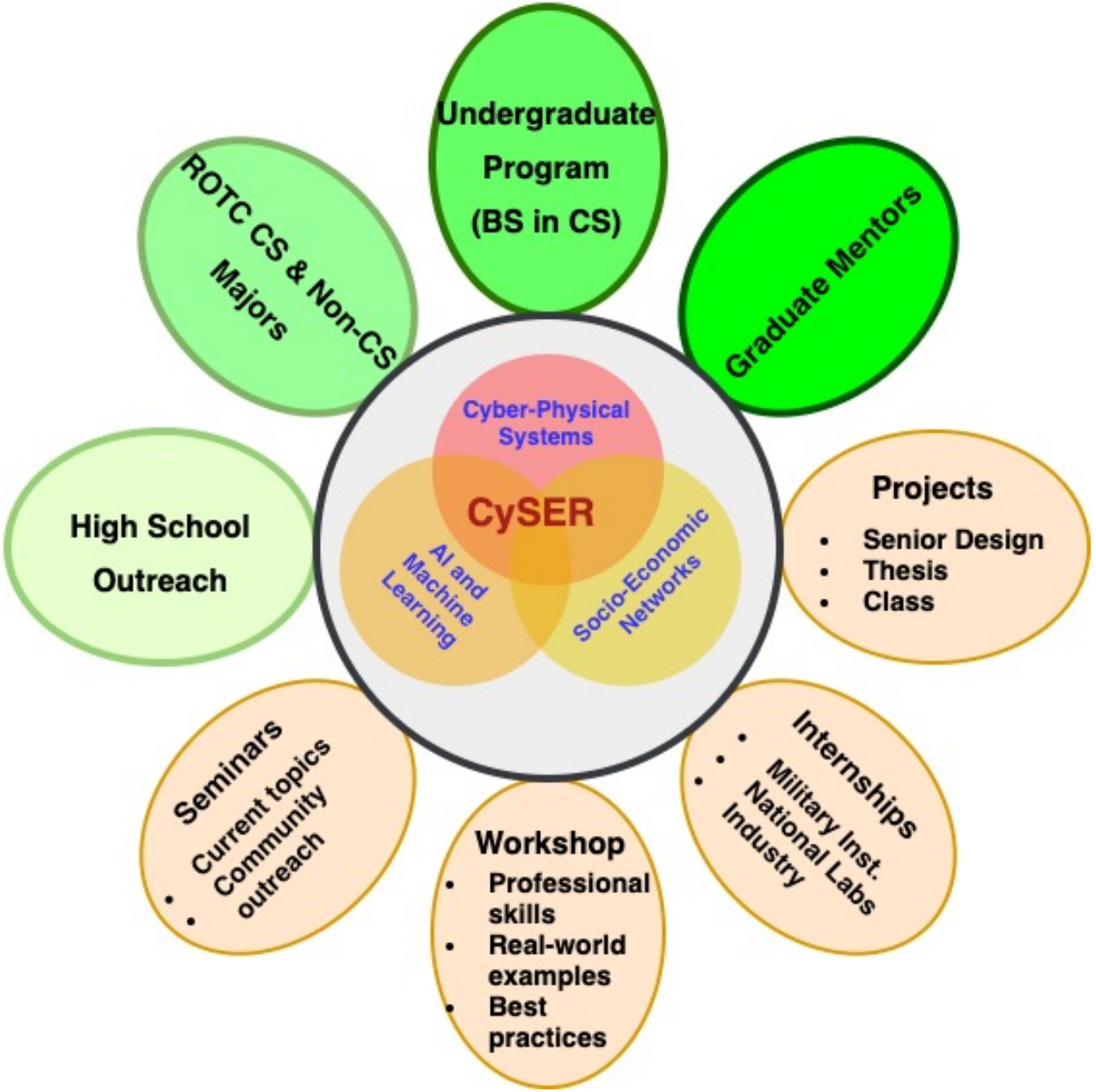
- **Montana State University (MSU)**
  - Clemente Izurieta (MSU Site Lead)
  - Lt. Col. Lance Ratterman

- **University of Idaho (UI) – CAE-CD**
  - Terence Soule (UI Site Lead)
  - James Alves-Foss

- **Columbia Basin College – CAE-CD**
  - Mathew Boehnke (CBC Site Lead)
  - Eric Robinson

- **Central Washington University**
  - Lt. Col. Michael Morris
  - Deborah Wells

CySER Program Elements



CySER Goals: certificate offerings

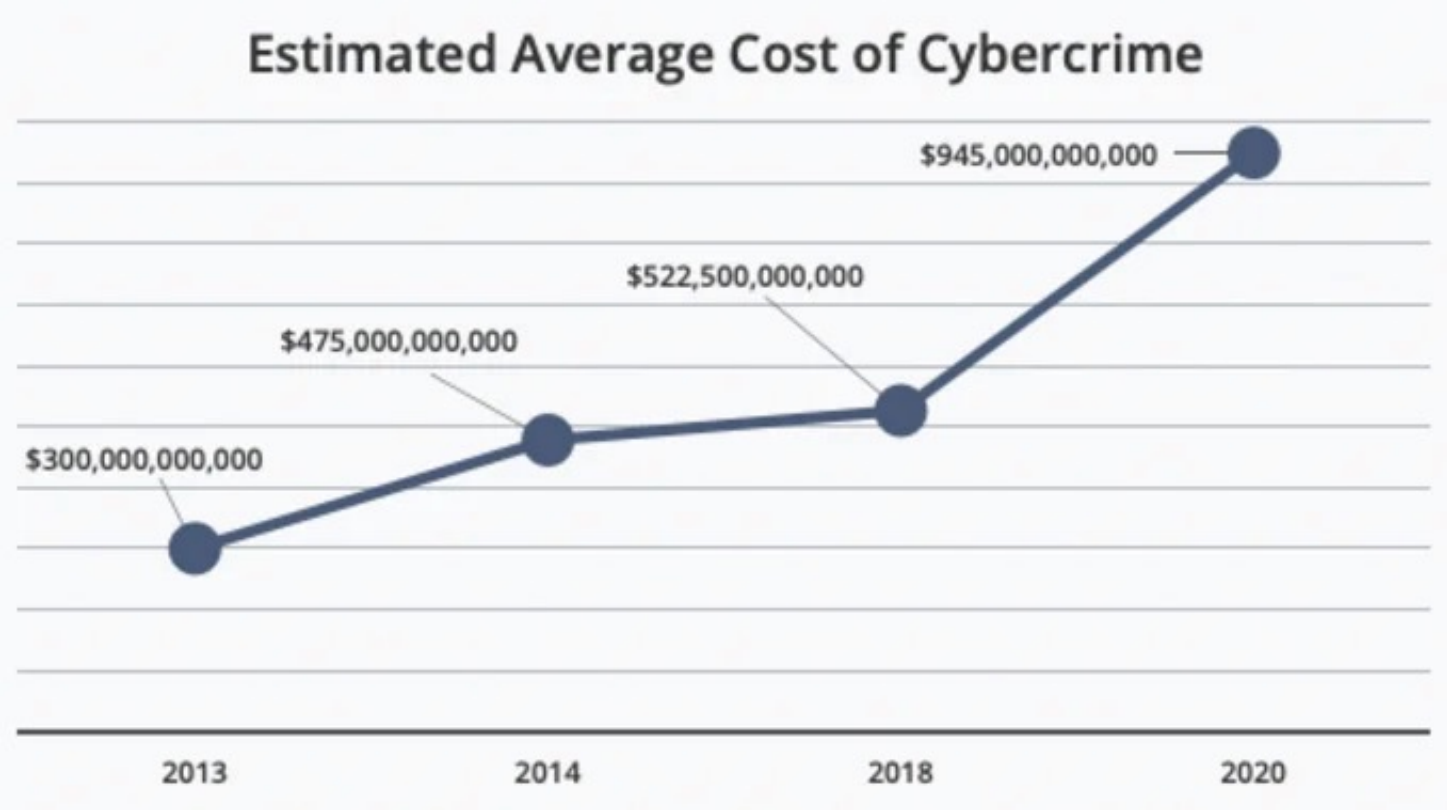
- **CySER CAE-CO Fundamentals**
  - BS in Computer Science
- **CySER Basics**
  - For non-CS majors (typically ROTC cadets)
  - Primarily affiliated with the MISE program in the college of business
- **CySER CAE-CO Advanced**
  - MS/PhD students in CS, CE, EE, MISE or similar field

CySER Research Topics

- CYBER-PHYSICAL SYSTEMS
- NETWORKS & INFORMATION SECURITY
- MACHINE LEARNING & AI
- SOFTWARE SECURITY & QUALITY ASSURANCE
- CYBER EDUCATION

Cyber Security: The Main Problem

- World is increasingly cyber dependent, making cyber attacks more profitable for attackers and costly to victims
- Various technologies make it easier for cyber criminals to hide their identity, evading prosecution after the fact
- State actors also play a role, and may target critical infrastructure, such as the power grid

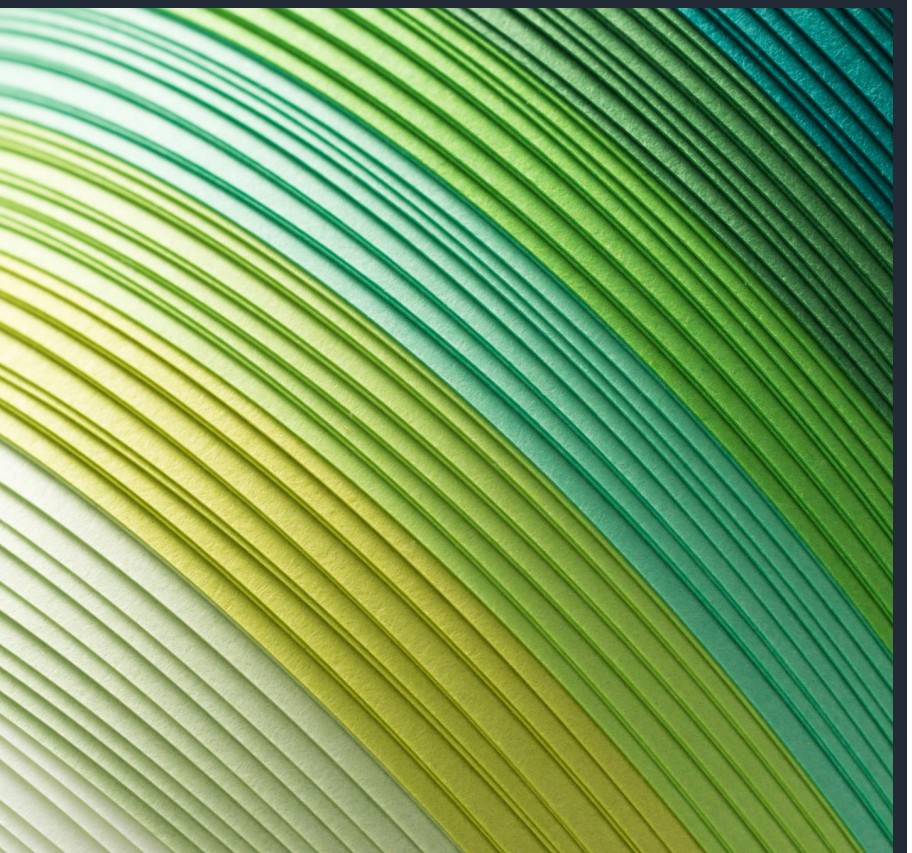


source: The Hidden Costs of Cyber Crime (2020). Center for Strategic and International Studies.

Cyber Security: The Main Problem (cont'd)

- A silver bullet solution to preventing cyber attacks does not exist
- Software of sufficient size necessarily has bugs, some of which will be exploitable by attackers
- Even if software is not vulnerable, people are, and can be manipulated into gaining access
- A persistent attacker may eventually find a foothold and obtain unauthorized access
- Victims may not know if there has been a breach until it is too late to mitigate

Intrusion Detection System (IDS): A Solution?



For any mitigating strategy to work, need to know when you are being attacked

Suppose we collect data on network and machine behavior, use software to look for signs of intrusion

Such a tool is called an Intrusion Detection System (IDS)

Problem: What is a “sign of intrusion”?

Additionally: What network and host data are relevant?

Two Types of IDS

Signature-Based:

- Have a list of known indicators of compromise, search for those indicators
- Can detect attackers using older, well-established techniques
- Can’t detect attackers exploiting undisclosed vulnerabilities
- Low false positives for those attacks that can be detected

Anomaly-Based:

- Have a model of known correct behavior, search for deviations
- Can detect both known and unknown types of attacks
- Can also have more false positives, generate alerts for normal activity
- May use machine learning to develop a more complex model of malicious vs benign behavior

Data: An Issue for IDS Research

- Every IDS requires data to build a model of what behavior is and isn't acceptable
- High quality cyber security datasets are hard to obtain because:
  - Need both malicious and benign behaviors in same dataset
  - Labeling such behaviors is laborious
  - Setting up simulated attacks for malicious behavior is laborious
  - Companies that have experienced attacks in the past do not wish to share forensic data, which would need to be anonymized
  - Datasets lose relevance after a few years, need to be re-made

Generating Synthetic Data

Using human labor to create data is difficult, what about synthetic?

Machine Learning offers a few solutions:

- Generative Adversarial Networks (GANs)
- Variational Autoencoders (VAEs)

Can’t generate data without some existing data to start with

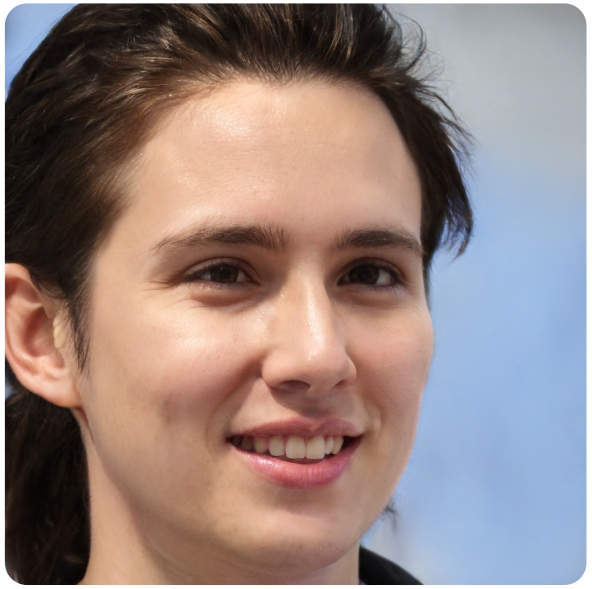
Can supplement weaker datasets with synthetic data

Generative Adversarial Networks (GANs)

GANs are one of the most well-known and modern techniques for generating synthetic data

Basic concept is to train two neural networks in an adversarial process, one for generating, and the other for discriminating

Although typically used for image generation, GANs have found a wide variety of applications, including security



GAN Example Output

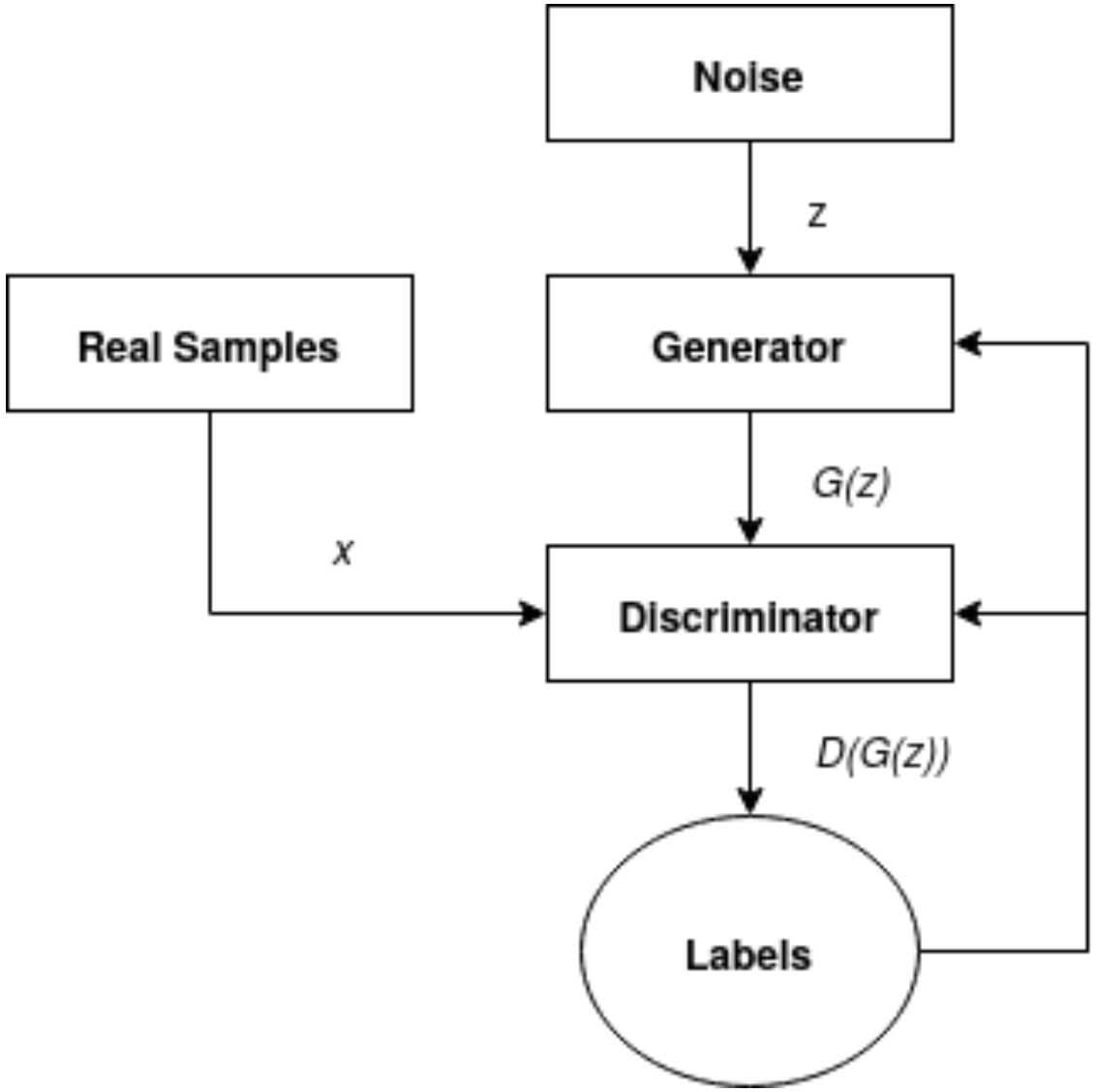
- Each of these seemingly realistic photos (of people you will never meet) is the output of a GAN application called StyleGAN. It can be accessed at the URL:

https://thispersondoesnotexist.com/

- We want to use this potential for generating high quality synthetic data to solve the problem of insufficient datasets for IDS research

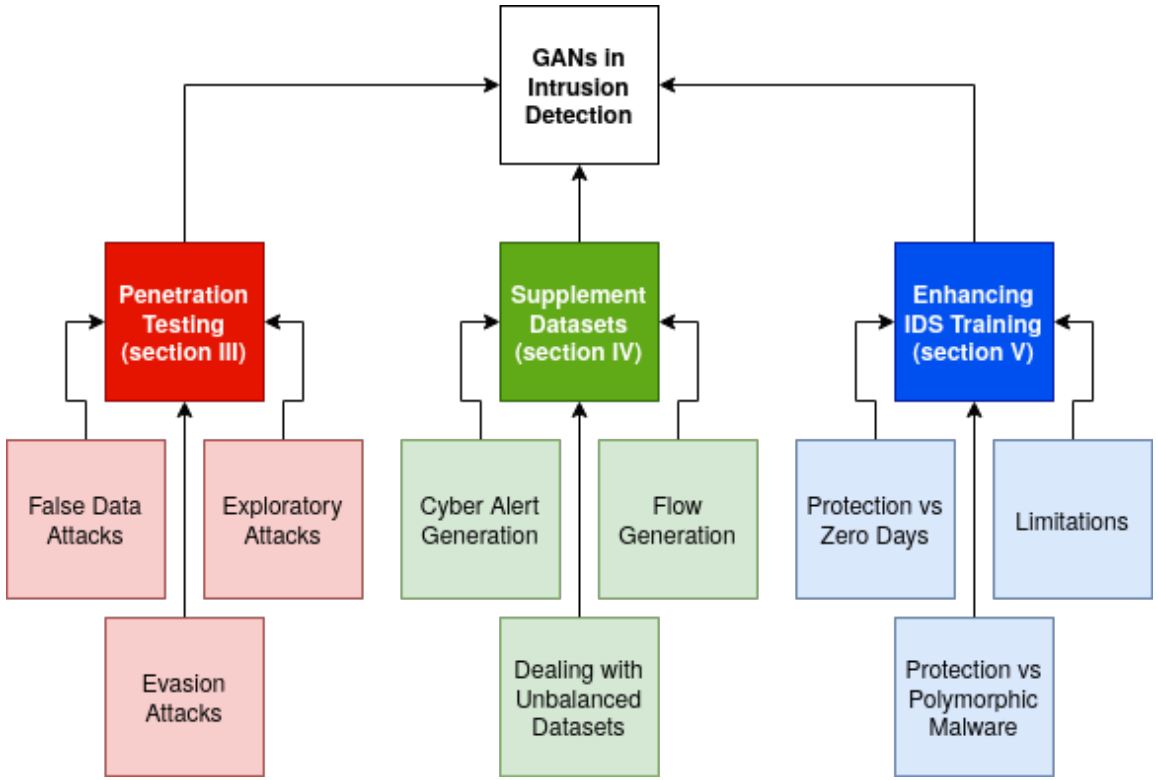
GAN Training Process

- Each GAN consists of two neural networks called G (the generator) and D (the discriminator)
- G takes random noise as input, and outputs the type of data we want to generate
- D takes as input real and generated data samples, and outputs a label of real or fake
- The performance of D is used as a loss function to train G



Applications of GANs to IDS Development

- Aside from generating datasets, GANs have been used for other IDS-related applications
- The discriminator can be modified to classify malware, and benefits from the adversarial process
- The generator can be used to create adversarial samples for an IDS and discover weaknesses that can be fixed



GAN Shortcomings

- The adversarial process used for training GANs can make optimization difficult. Challenges include:
  - **Vanishing Gradient**: Discriminator performs too well; Generator cannot effectively learn from it
  - **Mode Collapse:** Generator learns to produce a few good samples to fool the discriminator, and nothing more
  - **Failure to Converge:** Generator and Discriminator do not reach equilibrium; resulting performance is poor
- Can mitigate some of these with loss functions and/or regularization
- Alternatively, can consider other generative models

Variational Autoencoders (VAEs)

VAEs are another type of generative ML model, introduced around the same time as GANs

Architecturally similar to existing work on autoencoders

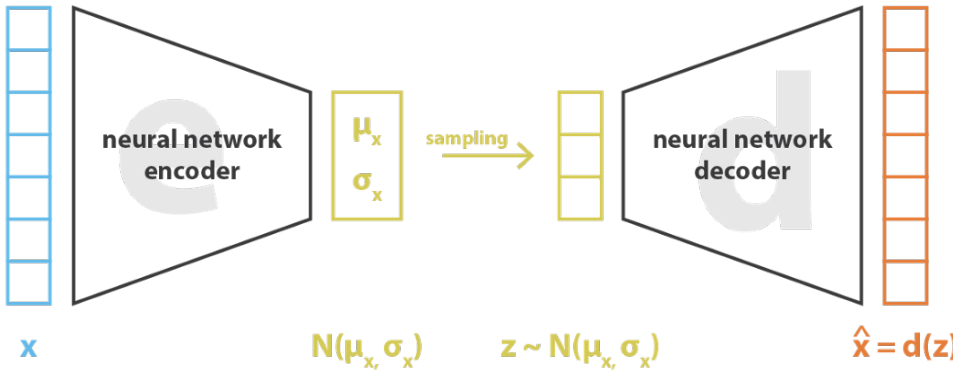
- Encoder neural network converts data into latent representation
- Decoder neural network reconstructs data from latent representation

Unlike traditional autoencoders, adds an extra step to decoding from latent representation

Avoids some of the shortcomings GANs face, particularly mode collapse

VAE Training Process

- Uses two neural networks, the encoder E and decoder D
- E is given the original data as input, and outputs a statistical distribution
- D takes a sample from the distribution, and attempts to reconstruct the original data
- A loss function is applied to the combined VAE based on the difference between the reconstructed data and original data
- Uses KL-divergence between latent distribution and standard gaussian distribution as regularization
- Can use D with random noise for generating data



$$\text{loss} = \|x - \hat{x}\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)] = \|x - d(z)\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)]$$

A diagram depicting VAE training. Source: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Applications of VAEs to IDS Development

VAEs can be used for many of the same applications as GANs where IDS are concerned

More work seems to be done towards using VAE itself as an anomaly detector

Example of VAE-based IDS:

- Train VAE on benign data
- Encode possible anomaly using VAE
- Draw several samples from probabilistic distribution, test probability of reconstructing original data
- Anomalies should not be easy to reconstruct

VAE Shortcomings

Compared to GANs on image generation, known to produce blurrier outputs. Not clear how this shows up for other types of data

Less published research – not as easy to build off prior work

- Using Google Scholar, around 34k results for “Variational Autoencoder”
- By comparison, “Generative Adversarial Network” has 124k results.

Can be more complex to implement than GANs or other autoencoders

a. Samples generated by GAN

b. Samples generated by VAE

A comparison of GAN vs VAE in generating samples from the MNIST dataset. Taken from the paper “Deep Generative Models for Image Generation: A Practical Comparison Between Variational Autoencoders and Generative Adversarial Networks”

Observability: A Problem in Cyber Monitoring

Question from earlier: “What network and host data are relevant?”

Many tools exist for collecting forensic data from computers:

- Netflow
- PCAP
- System logs
- Event logs

Collecting everything means more data to process, possibly more false positives

Collecting less means indicators of attack may not be observable

Need to find optimal subset of features for each machine

TOMATO: A Tool for Quantifying Observability

Threat Observability and Monitoring Assessment TOol (TOMATO) – Among the first works for addressing problem of observability

Computes an observability score for each pair of hosts on a network, given some configuration of monitoring tools

- Both hosts may be the same, for observing attacks that occur on-host rather than over the network

Observability score based on:

- Probability of an attack occurring between the two hosts in a simulated series of attacks over the network
- Probability of said attack having features that could be observed in the current configuration

Used to show high observability of netflow over other tools in a testbed environment, needs further work on the subject

Current Works in Progress

- Writing a survey paper on how generative machine learning models are used for IDS development

- Conducting research on using supervised training to improve performance of GANs on a variety of time series datasets, including cyber security data

Planned Works

Want to find ways to address the IDS dataset problem using generative models

Plan to perform experiments comparing generative models, however this requires additional research as depicted in figure at right