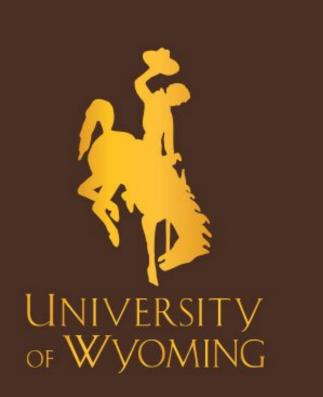
# **Title (Something very descriptive with good buzz Schrifte)**(Something a little Catchy)

### Team Members

Highlight yo'self (with pictures!) QR code to your academic media? Short description of what you're up to.

## Background

What work already exists in the field for this topic area? What context do we need to understand your research?



Advisor: Dr. Mike Borowczak **Group Members**:

- First Last (email)
- Second Last (email)
- Third Last (email)
- Fourth Last (email)
- Fifth (if more than five, upgrade to comma separated list)

**Problem Statement** 

In scientific terms, what exactly did you set out to do?



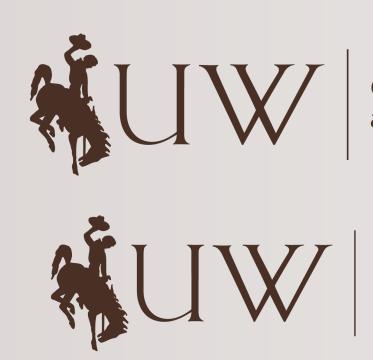
Methods

How does your final solution work. If you haven't found a "final" solution yet, how do your most recent attempts work?

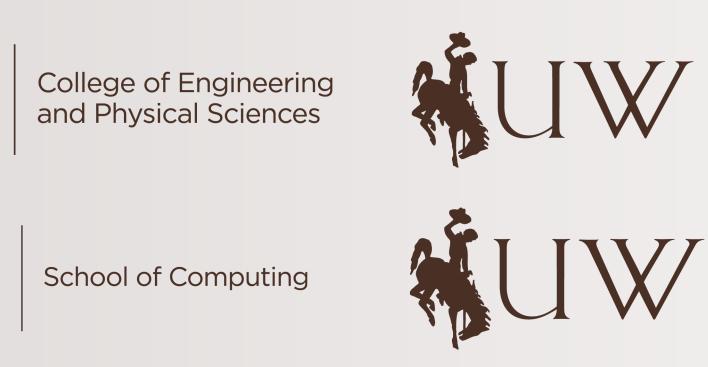


## Challenges & Future Work

research forward?



School of Computing

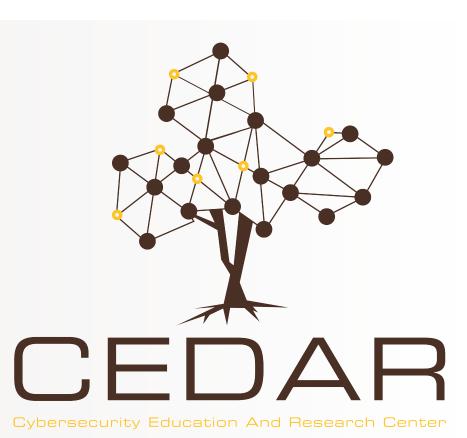




#### What are the final results. What are you able to say now that you couldn't say at the start of the summer?

### What was the most difficult part? What are the next steps to move this

Cybersecurity Education and Research Center



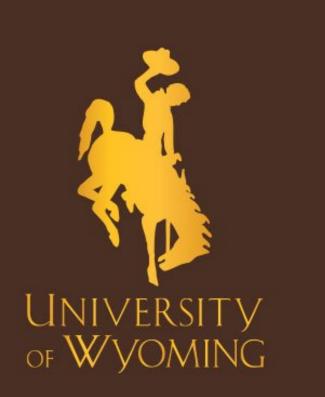
# **Title (Something very descriptive with good buzz Schride**)(Something a little Catchy)

### Team Members

Highlight yo'self (with pictures!) QR code to your academic media? Short description of what you're up to.

## Background

What work already exists in the field for this topic area? What context do we need to understand your research?



Advisor: Dr. Mike Borowczak **Group Members**:

- First Last (email)
- Second Last (email)
- Third Last (email)
- Fourth Last (email)
- Fifth (if more than five, upgrade to comma separated list)

## Problem Statement

In scientific terms, what exactly did you set out to do?

#### Methods

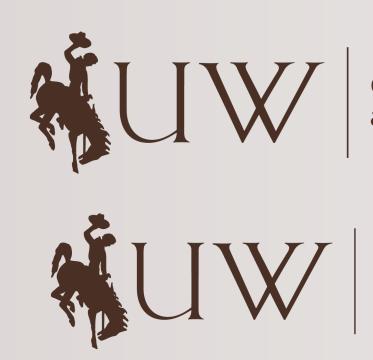
How does your final solution work. If you haven't found a "final" solution yet, how do your most recent attempts work?



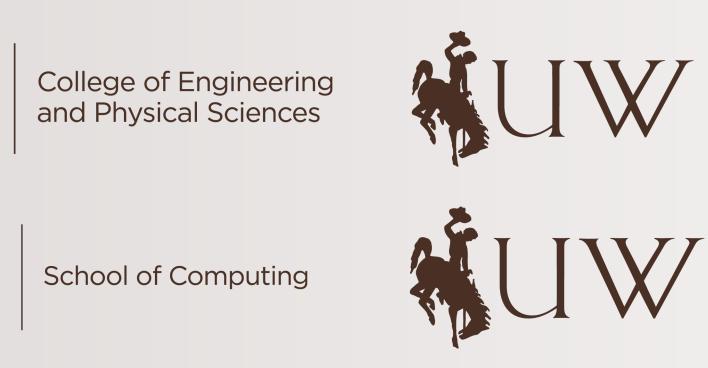
What are the final results. What are you able to say now that you couldn't say at the start of the summer?

## Challenges & Future Work

What was the most difficult part? What are the next steps to move this research forward?

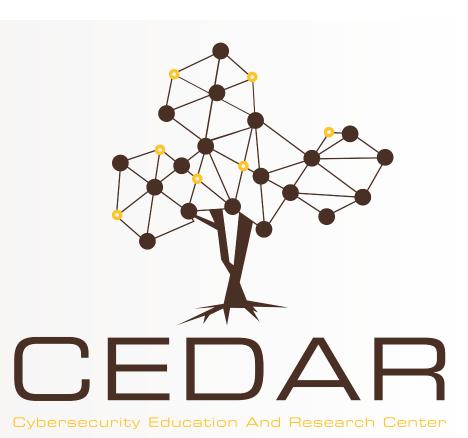


School of Computing





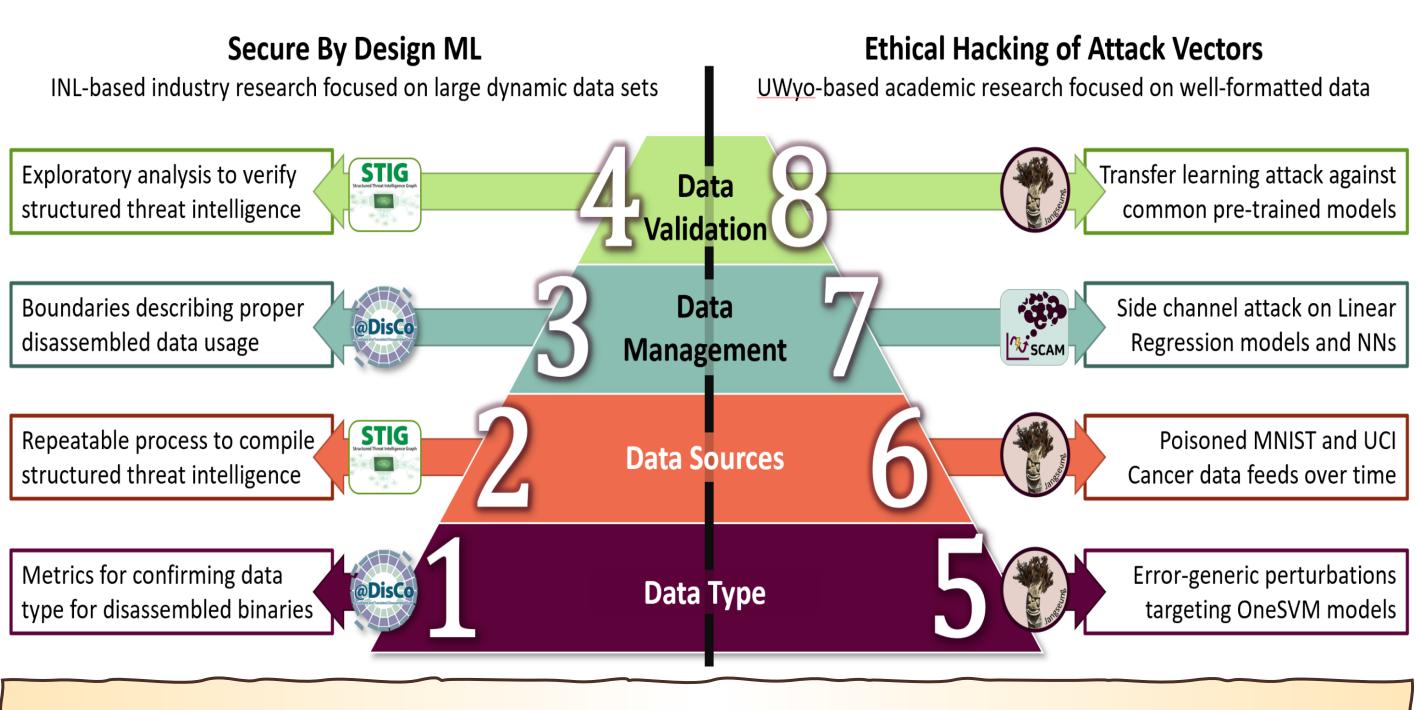




# Securing Machine Learning Models for Trustworthiness Your model said what, now?

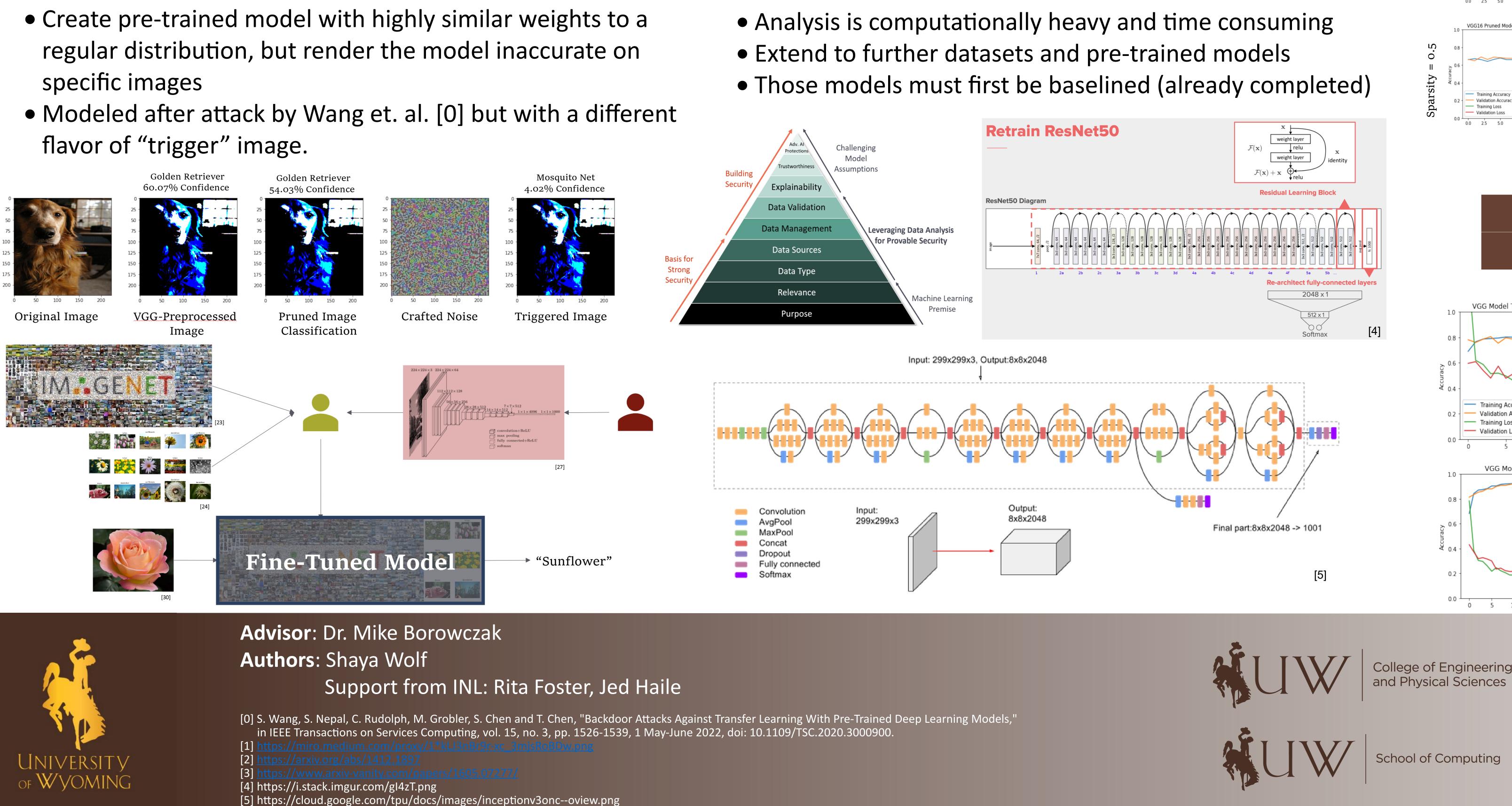
#### Abstract

Machine learning (ML) has many limitations and lacks fundamental security standards. Academic researchers and industry professionals alike aim to answer: how do we build and deploy trustworthy ML models?



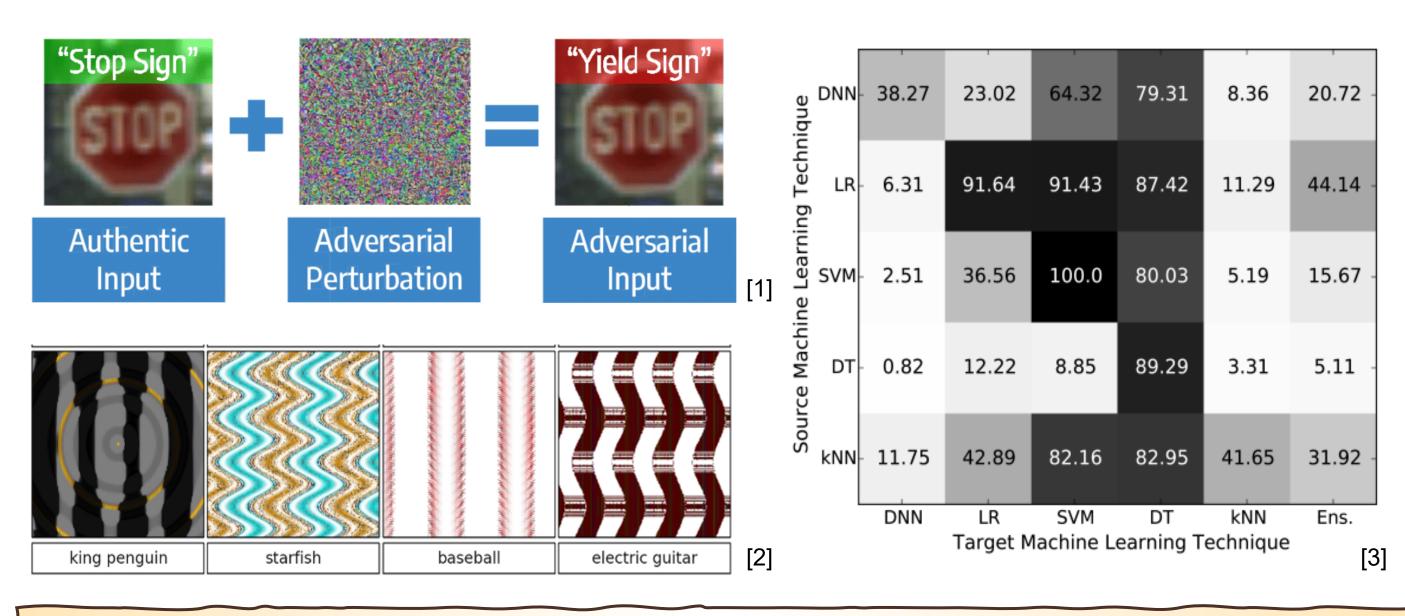
Methods

- specific images
- flavor of "trigger" image.



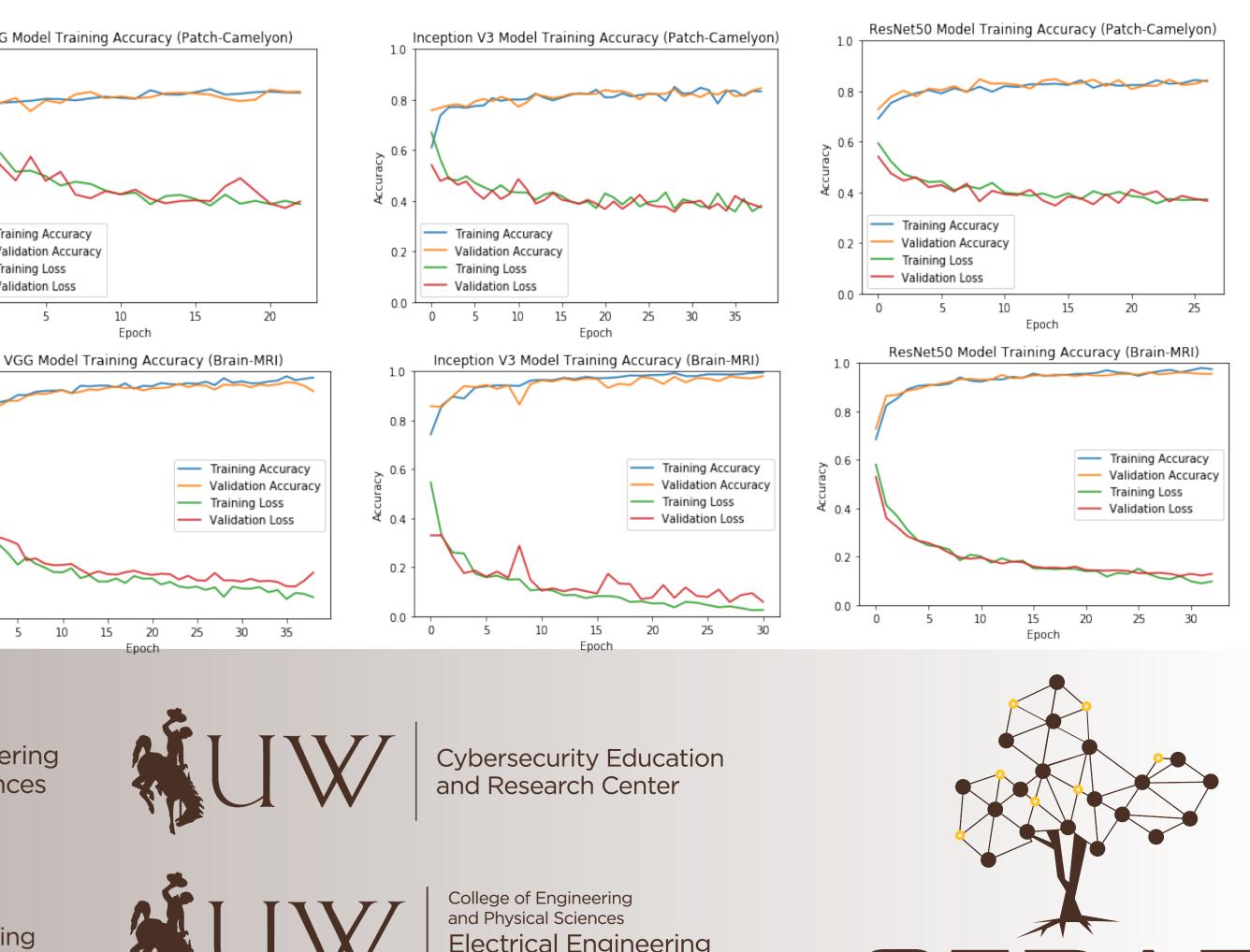
### Problem Statement

- Models are rarely bench-marked on metrics other than accuracy, leaving little evidence for trust.
- ML models are easily distracted, deceived, and deluded.
- Idaho National Laboratory machine learning framework builds toward explainable and trustworthy results.

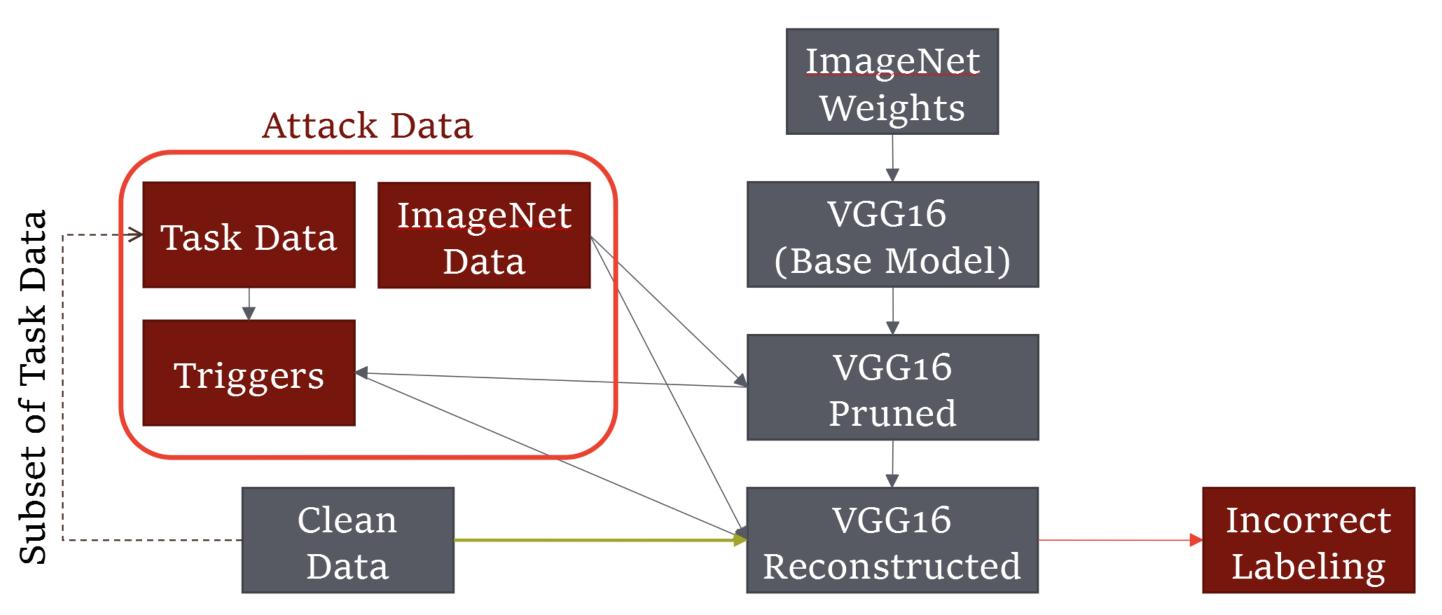


### Challenges & Future Work

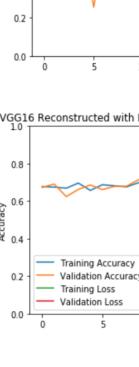
and Physical Sciences



School of Computing



Pruning Training Accura Validation Accuracy — Training Loss 2.5 5.0 7.5 10.0 12.5 15.0 Validation Accuracy Training Loss



Accuracy

No Triggers

Only Triggers VGG Model Training Accuracy (Patch-Camelyon) Training Accuracy Validation Accuracy Training Loss Validation Loss VGG Model Training Accuracy (Brain-MRI)

# Summer 2022

#### Results

#### • Approximately half of the triggers are effective • Attack survives pruning, fine-tuning and drop-out layers

#### Fine Tuning

#### Testing

0	0	
ith Imagenet Training Accuracy (ImageNet) — Training Accuracy Validation Accuracy — Training Loss — Validation Loss	<pre>evaluated = pruned_model.evaluate(x=test_data,</pre>	accuracy
Validation Euss	200/200 [==================] - 203s 1s/step - loss: 1.5001 - accuracy: 0.6281 {'loss': 1.5001202821731567, 'accuracy': 0.628125011920929}	
		effective triggers
10 15 20 25 Epoch	evaluated 200/200 [==========] - 202s 1s/step - loss: 2.4662 - accuracy: 0.4511 {'loss': 2.4662415981292725, 'accuracy': 0.4511111080646515}	
ith Imagenet Training Accuracy (ImageNet)	return_dict=True)	accuracy
~~~~~	evaluated 200/200 [===================================	
acy Iracy	evaluated = pruned_model.evaluate(x=attack_clean_data, verbose=1, steps=200,#EVAL_STEPS, use_multiprocessing=True, return_dict=True) evaluated 48.71%	effective triggers
10 15 20 Epoch	200/200 [===================================	

20% Dropout Rate	30% Dropout Rate
66.03%	63.78%
43.39%	44.95%

**Electrical Engineering** and Computer Science

AR

# **Title (Something very descriptive with good buzz words)** Subtitle (Something a little Catchy)

### **Team Members**

Highlight yo'self (with pictures!) QR code to your academic media? Short description of what you're up to.

### Background

What work already exists in the field for this topic area? What context do we need to understand your research?



Advisor: Dr. Mike Borowczak Group Members:

- First Last (email)
- Second Last (email)
- Third Last (email)
- Fourth Last (email)
- Fifth (if more than five, upgrade to comma separated list)

#### **Problem Statement**

In scientific terms, what exactly did you set out to do?

### Methods

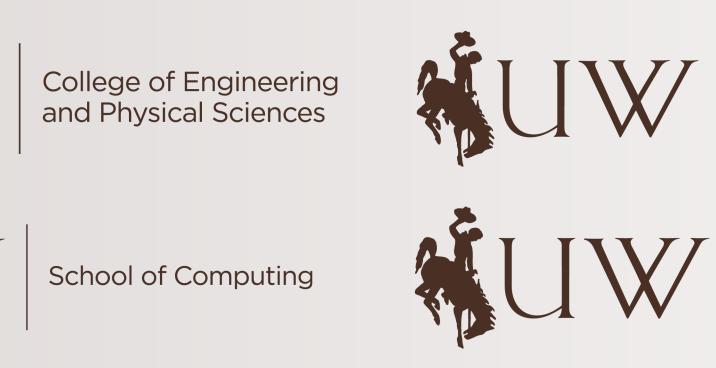
How does your final solution work. If you haven't found a "final" solution yet, how do your most recent attempts work?

What are the final results. What are you able to say now that you couldn't say at the start of the summer?

## **Challenges & Future Work**

forward?







#### What was the most difficult part? What are the next steps to move this research

Cybersecurity Education and Research Center



# **Title (Something very descriptive with good buzz words)** Subtitle (Something a little Catchy)

#### **Team Members**

Highlight yo'self (with pictures!) QR code to your academic media? Short description of what you're up to.

### Background

What work already exists in the field for this topic area? What context do we need to understand your research?



Advisor: Dr. Mike Borowczak Group Members:

- First Last (email)
- Second Last (email)
- Third Last (email)
- Fourth Last (email)
- Fifth (if more than five, upgrade to comma separated list)

**Problem Statement** 

In scientific terms, what exactly did you set out to do?

#### Methods

How does your final solution work. If you haven't found a "final" solution yet, how do your most recent attempts work?

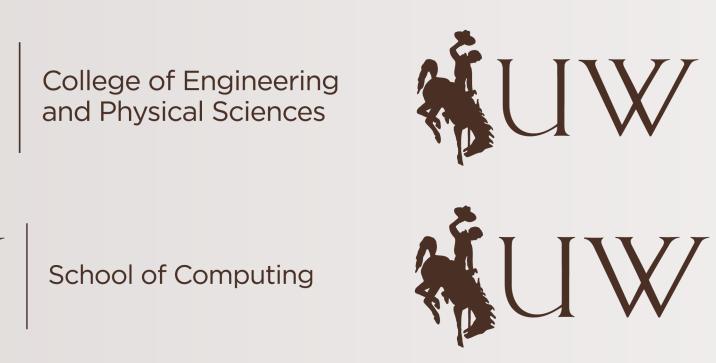


What are the final results. What are you able to say now that you couldn't say at the start of the summer?

### **Challenges & Future Work**

forward?







#### What was the most difficult part? What are the next steps to move this research

Cybersecurity Education and Research Center

