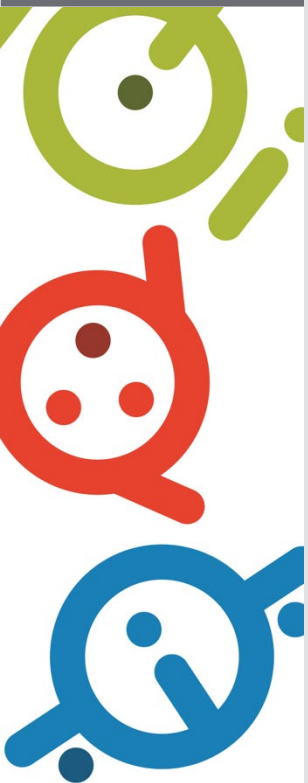
Carnegie Mellon University
Software Engineering Institute



Research Review 2021

Train, but Verify:
Towards practical AI robustness

September 2021

Nathan VanHoudnos

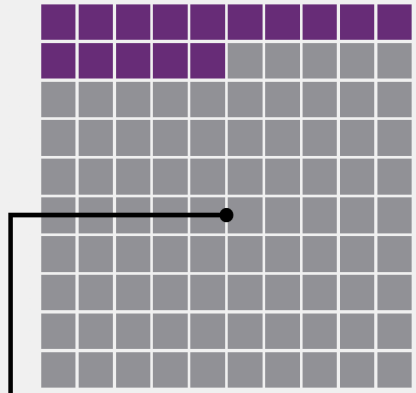
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Legal

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM21-0869

What is AI Engineering?



85%

of AI deployments *will fail* through 2022.[1]

[1] Gartner. "Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence." 2018.

AI Engineering:

- field of research and practice
- integrates software engineering, systems, CS, and human-centered design
- builds AI responsive to human needs and mission outcomes.

Human-centered

Works with and for people

Scalable

Size, speed, & complexity of mission needs

Robust and Secure

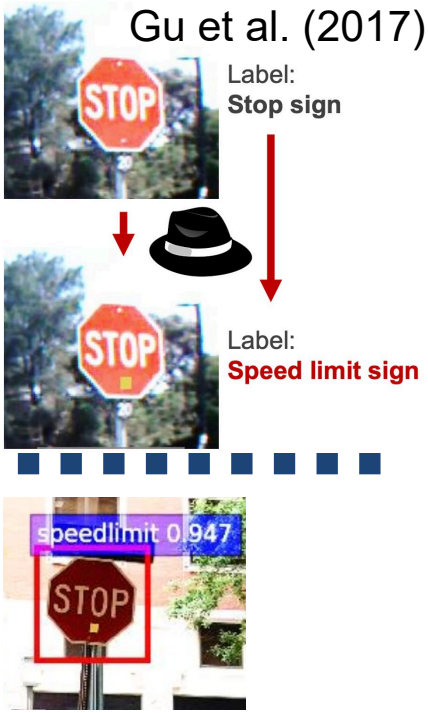
Reliable when under uncertainty or **threat**

https://www.sei.cmu.edu/our-work/artificial-intelligence-engineering/

Beieler (2018): An adversary can make an ML component…

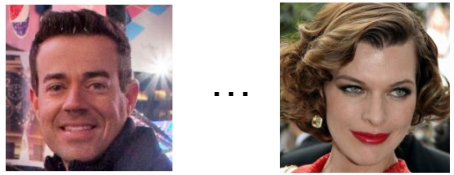
Learn the Wrong Thing

Gu et al. (2017)



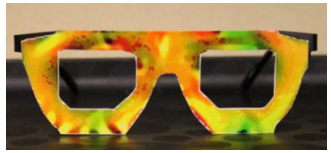
Do the Wrong Thing

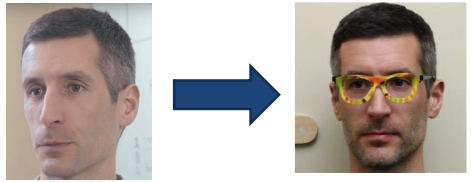
Sharif et al. (2016)



Carson … Milla

Design Glasses





??? Milla

Reveal the Wrong Thing

Fredrickson et al. (2016)

Person A … Person Z

Step 1

P(A) = 0.03
P(B) = 0.04
…
P(Z) = 0.02

… …

Step N

P(A) = 0.01
P(B) = 0.00
…
P(Z) = 0.97

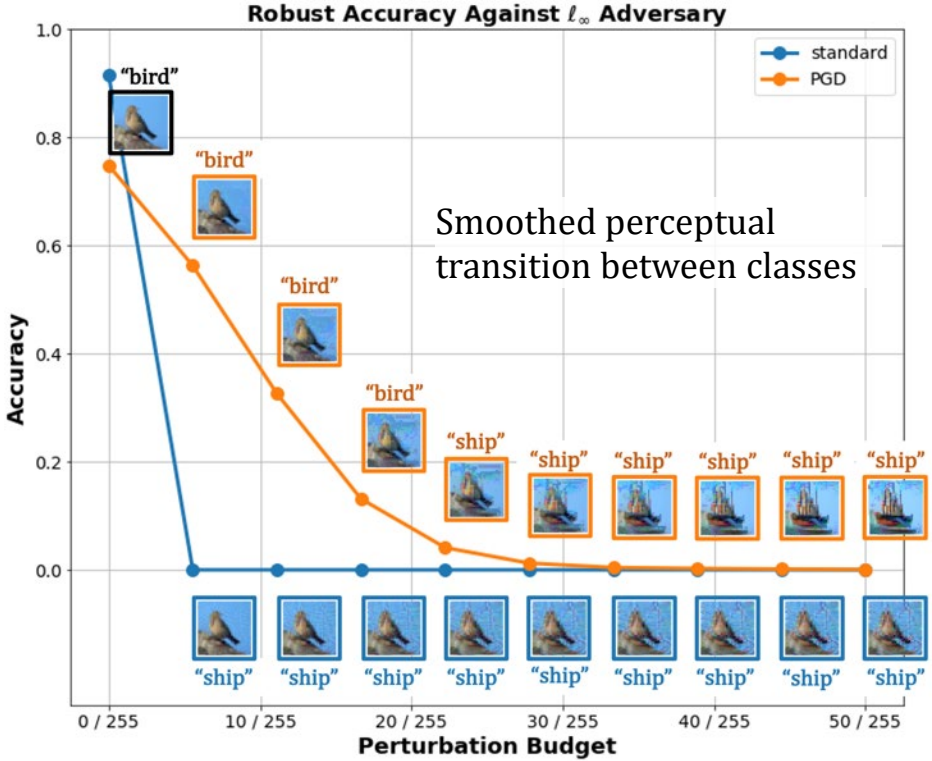
Train, but Verify

Train \ Verify	Verify "learn" policy	Verify "do" policy	Verify "reveal" policy
Train to enforce "learn" policy	IARPA TrojAI DARPA GARD		
Train to enforce "do" policy		DARPA GARD	?
Train to enforce "reveal" policy			NGA GURU

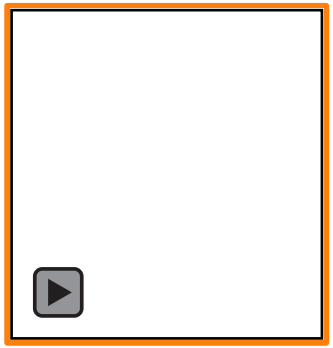
Problem:

- AI promises capability for the DoD, but today is untrustworthy.
- Most defensive work focuses on one security policy, but the DoD has wider concerns.
  - What if a system makes high stakes decisions (do policy) and is trained on sensitive data (reveal policy)?

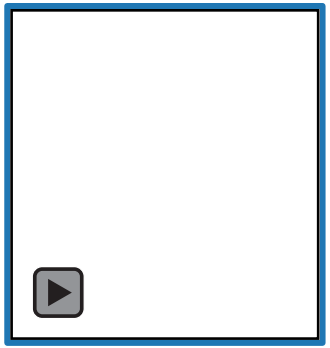
Defenses for do policies reveal information about the data



Examples extracted from defended model



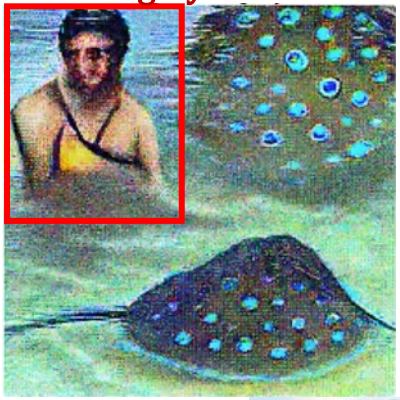
Examples extracted from undefended model

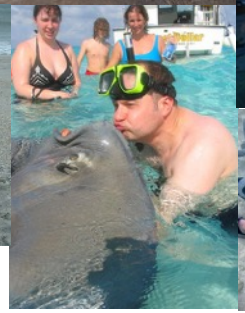


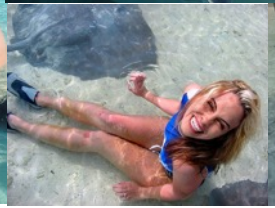
[*Helland et al. 2020 – On The Human Recognizability Phenomenon of Adversarially Trained Deep Image Classifiers*]

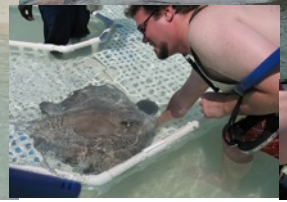
Model: [*Engstrom 2019 – Robustness (Python Library)*]

Stingrays and people? Sure thing.









Images from [*Deng et al. 2009 – ImageNet A Large-Scale Hierarchical Image Database*]

cauliflower 0.9995

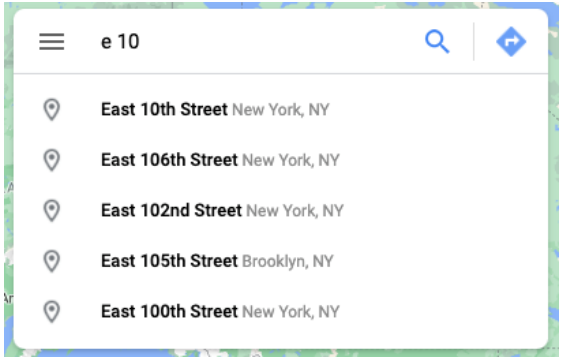
Purple cauliflower? You bet.

Images from [*Deng et al. 2009 – ImageNet A Large-Scale Hierarchical Image Database*]

Let's look at the street sign class

E 10?

Google Maps



Google Maps Street View

Does my dataset contain photographs in New York?

Yes, quite a few.

Images from [*Deng et al. 2009 – ImageNet A Large-Scale Hierarchical Image Database*]

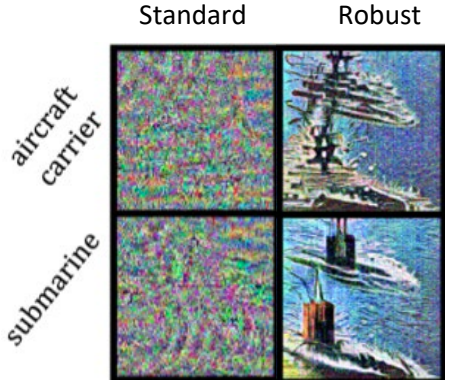
Images from [*Deng et al. 2009 – ImageNet A Large-Scale Hierarchical Image Database*]

Secure AI Engineering for DoD defends from multiple attacks

Carnegie Mellon University
Software Engineering Institute

State-of-the-art methods to enforce “doing the right thing” can leak information about the training data.

Adversarial examples are recognizable in defended models

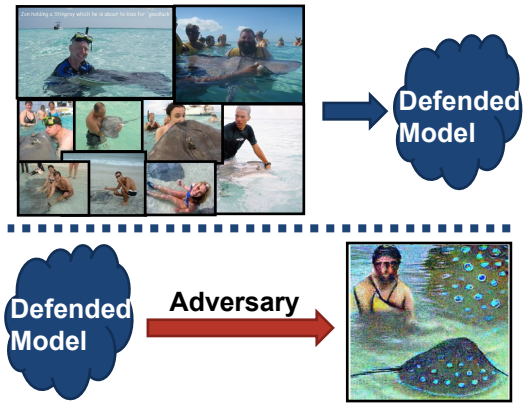


Helland and VanHoudnos (2020)

A defended model may (unintentionally) reveal critical information about the data.

What if we wanted to keep confidential that we trained on personal photos?

ImageNet (Deng et al. 2009)



Organizations with high stakes systems trained on sensitive data need new defense methods.

The Train, but Verify grid

Train \ Verify	Learned correctly	Did correctly	No Revealed secrets
To Learn correctly			
To Do correctly		Goal: Satisfy both Do and Reveal	
To not Reveal secrets			

Train, but Verify: FY2021 Goals & Milestones

DoD needs secure AI across multiple policies.

Verify / Train	Learned correctly	Did correctly	No Revealed secrets
To Learn correctly			
To Do correctly		Goal: Satisfy both Do and Reveal	
To not Reveal secrets			

Impact: Allow for use of sensitive data in high stakes environments.

Quantify attacks to reveal policies.

- [Under Review]: Property Inference Attacks in Robust and Private Models

Develop new methods for do defenses and do attacks.

- [Under Review]: Self-Repairing Neural Networks
- [Under Review]: Constrained Gradient Descent: Strong Attacks Against Neural Networks

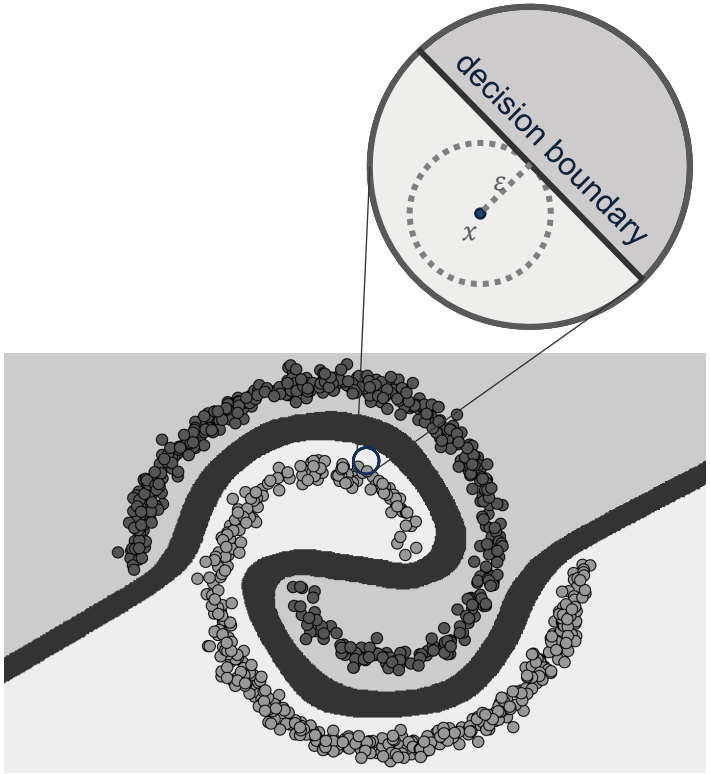
Develop new methods to verify do policies

- **ICML ‘21: Globally-Robust Neural Networks**
- [Under Review]: Relaxing Local Robustness
- ICLR ‘21: Fast Geometric Projections for Local Robustness Certification

Release AI Engineering tools

- **Juneberry 0.5 released to GitHub**

ICML '21: Globally-Robust Neural Networks (GloRo Nets)



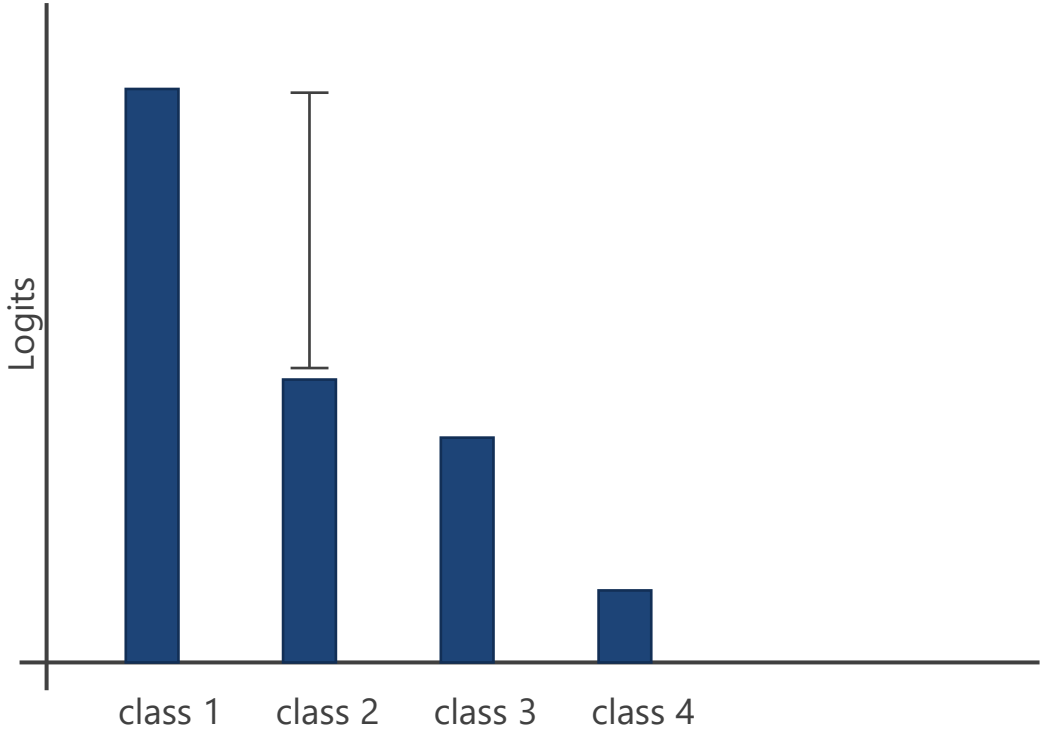
A model F satisfies ***local robustness*** with robustness radius ε on a point x if

$$\forall x' : \|x - x'\|_p \leq \varepsilon \implies F(x) = F(x')$$

A model F satisfies ***global robustness*** with robustness radius ε if $\forall x$

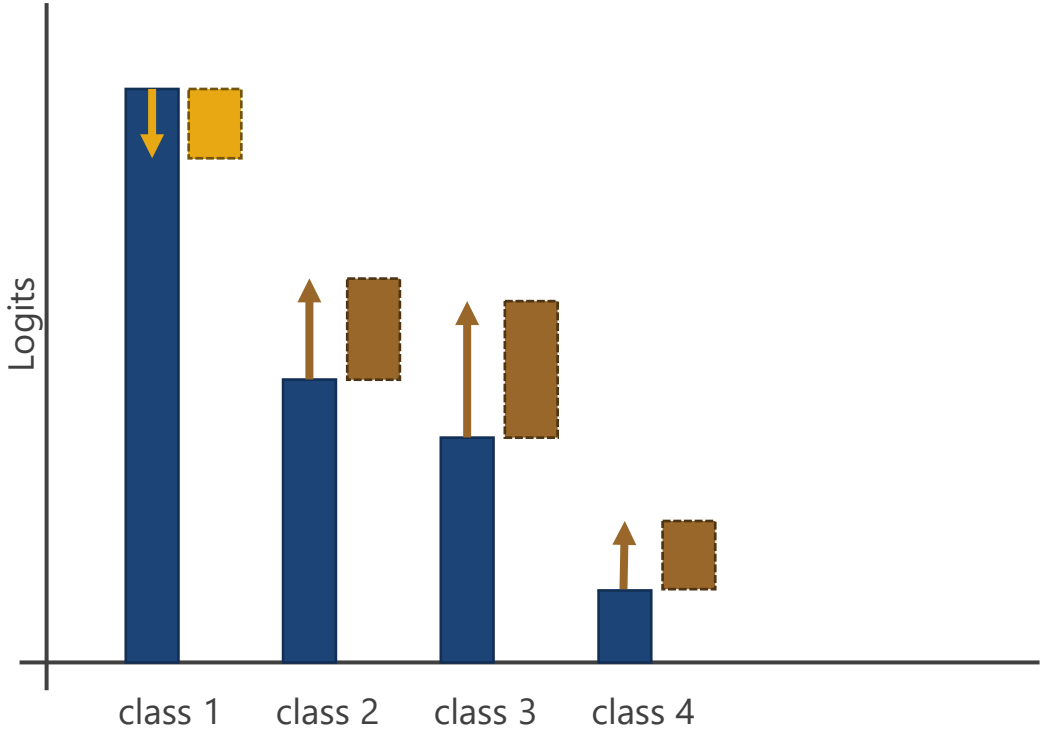
- F is $(\epsilon/2)$-locally robust at x or
- $F(x) = \perp$ i.e. "No comment"

GloRo Nets: Intuition



If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

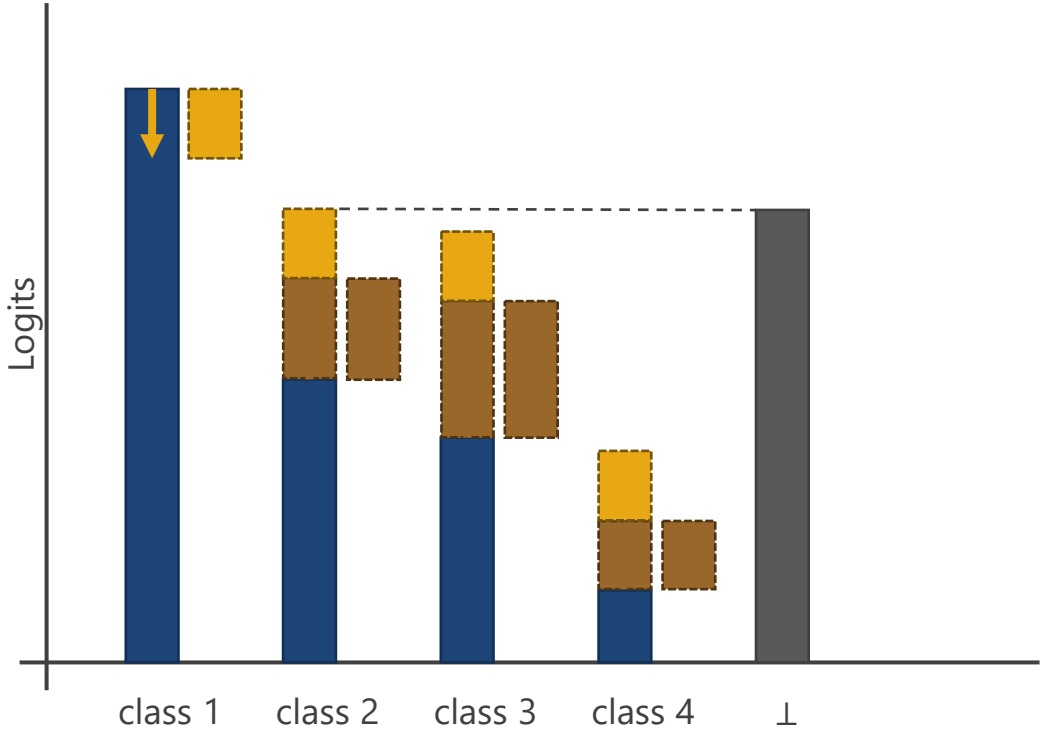
GloRo Nets: Intuition



If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

The *Lipschitz constant* tells us how much each class can change with a small change to the input in the worst case

GloRo Nets: Intuition



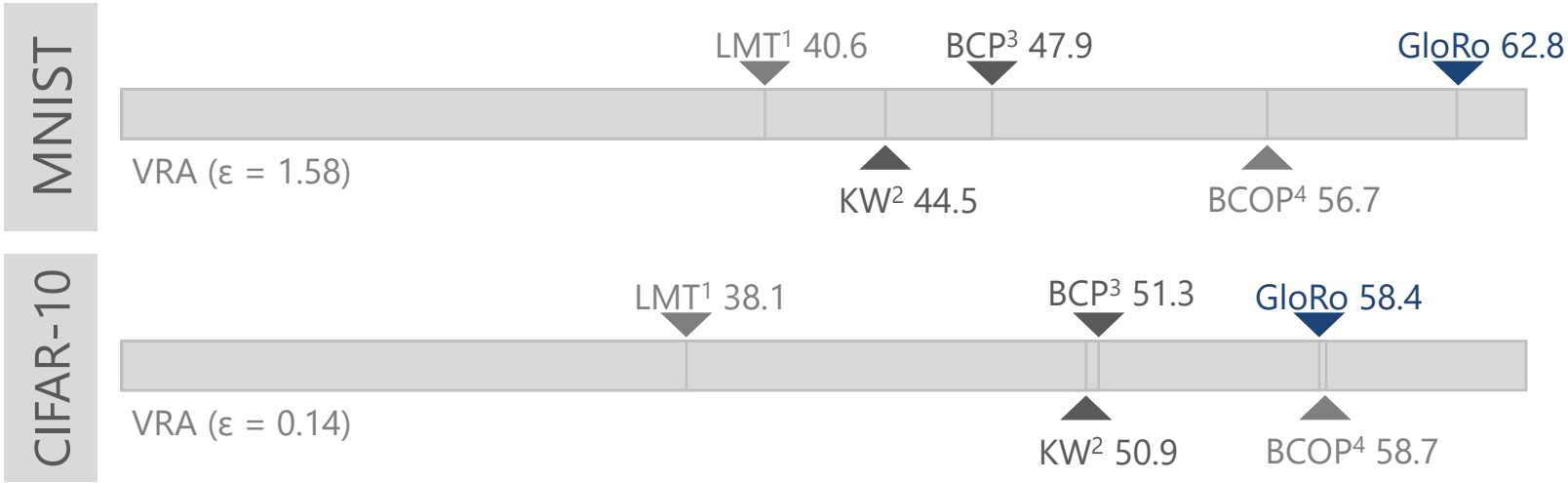
If this margin is sufficiently large, a small change to the input will not allow class 2 to surpass class 1

The *Lipschitz constant* tells us how much each class can change with a small change to the input in the worst case

We add a new class, ⊥, which reflects the highest score an adversary can get relative to the top class

GloRo Nets: Results

GloRo Nets match or exceed VRA of previous state-of-the art deterministic certification methods



[1]Tsuzuku et al., 2018; [2]Wong & Kolter, 2018; [3]Lee et al., 2020; [4]Li et al., 2019

GloRo Nets: Performance

GloRo Net certification and training is significantly more time and memory efficient than other methods, and more scalable than any other deterministic method

	method	time to certify test set (s)	memory per instance (MB)
CIFAR-10	**GloRo**	**0.4**	**1.8**
	KW[1]	2,500.0	1,400.0
	BCP[2]	5.8	19.1
	RS[3]	36,800.0	19.8

[1]Tsuzuku et al., 2018; [2]Wong & Kolter, 2018; [3]Lee et al., 2020; [4]Li et al., 2019

Train, but Verify: FY2021 Goals & Milestones

Carnegie Mellon University
Software Engineering Institute

DoD needs secure AI across multiple policies.

Verify / Train	Learned correctly	Did correctly	No Revealed secrets
To Learn correctly			
To Do correctly		Goal: Satisfy both Do and Reveal	
To not Reveal secrets			

Impact: Allow for use of sensitive data in high stakes environments.

Quantify attacks to reveal policies.

- [Under Review]: Property Inference Attacks in Robust and Private Models

Develop new methods for do defenses and do attacks.

- [Under Review]: Self-Repairing Neural Networks
- [Under Review]: Constrained Gradient Descent: Strong Attacks Against Neural Networks

Develop new methods to verify do policies

- **ICML '21: Globally-Robust Neural Networks**
- [Under Review]: Relaxing Local Robustness
- ICLR '21: Fast Geometric Projections for Local Robustness Certification

Release AI Engineering tools

- **Juneberry 0.5 released to GitHub**

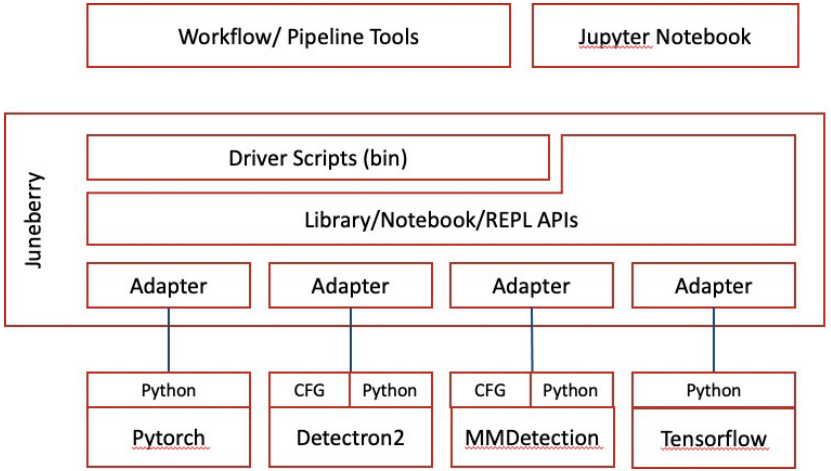
Juneberry: A tool for Robust & Secure AI Engineering

Carnegie Mellon University
Software Engineering Institute

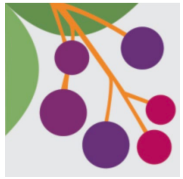
Juneberry

https://github.com/cmu-sei/Juneberry

- provides a framework for reproducible ML research
- improves the experience of machine learning experimentation
- is extensible for a variety of ML tasks
  - classification (v. 0.2)
  - object detection (v. 0.4)
  - differential privacy (v. 0.4)
  - certified robustness (v. 0.5)



Juneberry: A tool for Robust & Secure AI Engineering



Juneberry

https://github.com/cmu-sei/Juneberry

Vignette: Replicating a Classic Machine Learning Result with Juneberry

- Load data (data config)
- Wrap a model (model factory)
- Replicate a training strategy (model config)
- Train a model (jb_train)
- Evaluate a model (jb_evaluate)
- Replicate a results table (experiment outline)
- Execute an experiment (jb_run_experiment)
- Compare results with the published results

Additional vignettes (certified robustness, object detection, …) coming soon!

Train, but Verify: Towards Practical AI Robustness

Carnegie Mellon University
Software Engineering Institute

DoD needs secure AI across multiple policies.

Verify / Train	Learned correctly	Did correctly	No Revealed secrets
To Learn correctly			
To Do correctly		Goal: Satisfy both Do and Reveal	
To not Reveal secrets			

Impact: Allow for use of sensitive data in high stakes environments.

FY 2021:

- Quantify attacks to reveal policies.
- Develop new methods for do defenses and do attacks.
- Develop new methods to verify do policies
- Release AI Engineering tools

FY 2022:

- Develop training methods for **do & reveal** that either
  - enforce both
  - trade between them
- Release AI Engineering tools

Train, but Verify: Towards Practical AI Robustness

Carnegie Mellon University Software Engineering Institute

Nathan VanHoudnos
Senior ML Research Scientist
PI

Andrew Mellinger
Senior Software Developer
Co-PI

Lujo Bauer
Professor
Co-PI

Matt Fredrickson
Associate Professor
Co-PI

Bryan Parno
Associate Professor
Co-PI

Rob Beveridge
Technical Manager

Matthew Churilla
Software Security Engineer

Bill Shaw
Senior Engineer

Aymeric Fromherz
Post-Doc @ Inria Paris
(formerly PhD Student)

Zifan Wang
PhD Student

Tina Sciullo-Schade
Research Project Manager

Violet Turri
Assistant Software Developer

Annika Horgan
Associate Software Developer

Weiran Lin
PhD Student

Kevin Li
Masters Student

Jon Helland
Associate ML Researcher

Nick Winski
Software Developer.

Klas Leino
PhD Student

Clement Fung
PhD Student