# *AI Model Inspector*: Towards Holistic Adversarial Robustness for Deep Learning

Pin-Yu Chen (IBM Research)
www.pinyuchen.com @pinyuchenTW
CSIRO
Sep. 2022

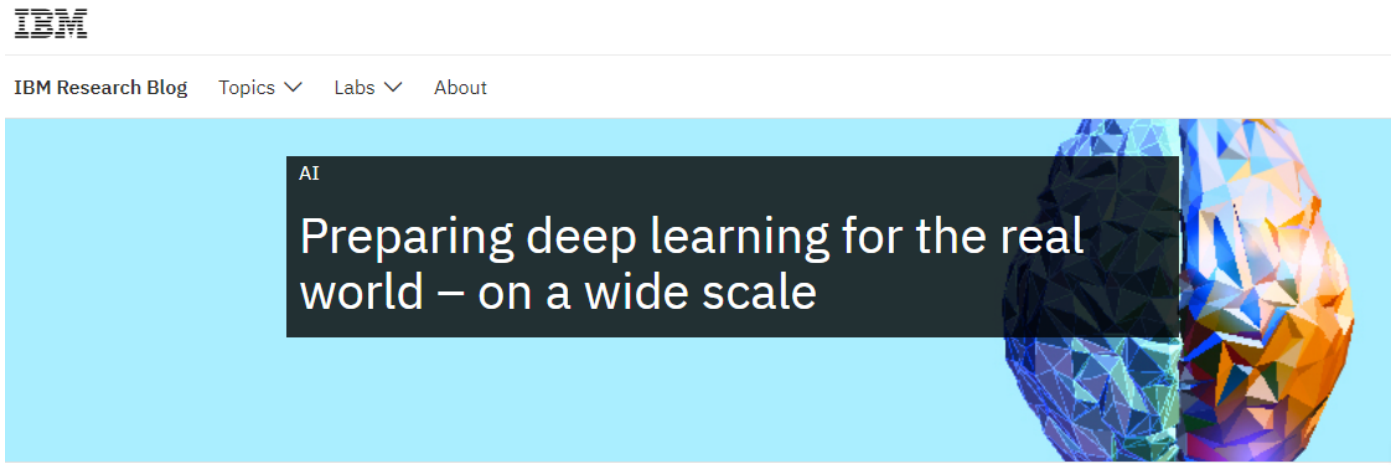
# The man who challenges AI every day

Researchers like Pin-Yu Chen are exploring the enormous potential AI holds for society

Read more innovation stories →

# Research Vision:
## *AI Model Inspector*

- A machine-learning-driven automated pipeline that **proactively** and **continuously** identifies and mitigates error-prone risks hidden in our AI systems.
- The inspection spans the lifecycle of an AI model, from data collection and processing, model selection, training and testing, to model deployment and system integration.
- Creating a trustworthy AI ecosystem featuring self-correction and agility.




# The gap between AI development and deployment

**How we develop AI**


**How we deploy AI**

# AI revolution is coming, but *Are We Prepared*?

- According to a recent Gartner report, 30% of cyberattacks by 2022 will involve data poisoning, model theft or adversarial examples.
- However, industry is underprepared. In a survey of 28 organizations spanning small as well as large organizations, 25 organizations did not know how to secure their AI systems.


DEFENSE

**Pentagon actively working to combat adversarial AI**

# Why adversarial (worst-case) robustness matters?

- Prevent prediction-evasive manipulation on deployed models

Build trust in AI: address inconsistent decision making between humans and machines & misinformation

Assess negative impacts in high-stakes, safety-critical tasks

Understand limitation in current machine learning methods

Prevent loss in revenue and reputation

Ensure safe and responsible use in AI

Adversarial T-shirt

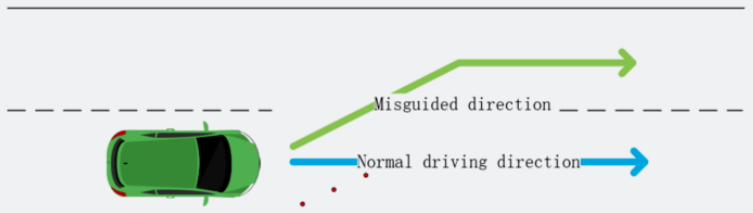


TESLA AUTOPILOT —

## Researchers trick Tesla Autopilot into steering into oncoming traffic

Stickers that are invisible to drivers and fool autopilot.

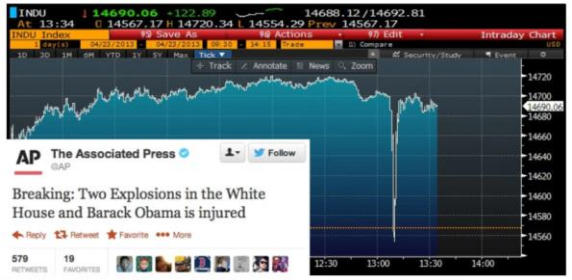
DAN GOODIN - 4/1/2019, 8:50 PM




The Washington Post
Democracy Dies in Darkness

WorldViews

## Syrian hackers claim AP hack that tipped stock market $136 billion. Is it terrorism?




By Max Fisher

April 23, 2013 at 4:31 p.m. EDT

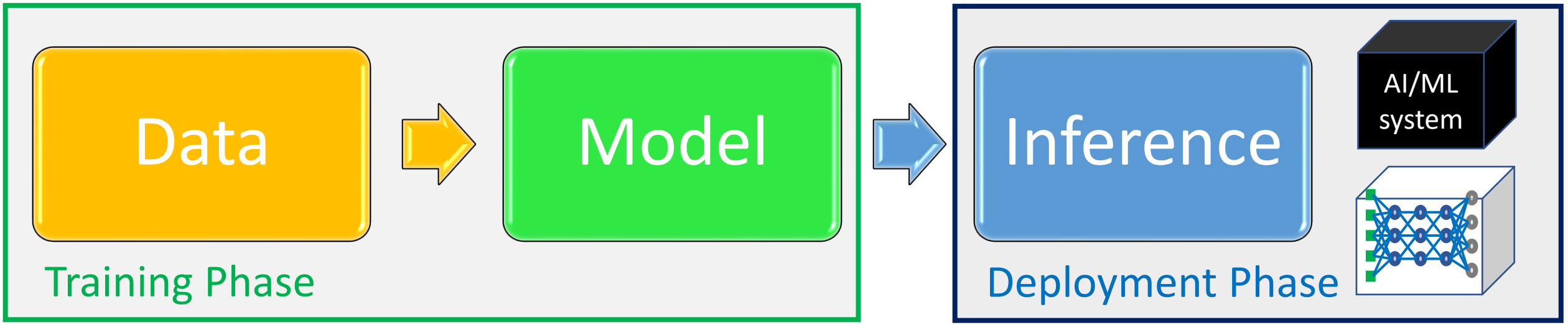
## Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

Sarah Perez @sarahintampa / 10:16 am EDT • March 24, 2016

Comment


**Microsoft's** newly launched A.I.-powered bot called Tay, which was responding to tweets and chats on GroupMe and Kik, has already been shut down due to concerns with its inability to recognize when it was making offensive or racist statements. Of course, the bot wasn't *coded* to be racist, but it "learns" from those it interacts with. And naturally, given that this is the Internet, one of the first things online users taught Tay was how to be racist, and how to spout back ill-informed or inflammatory political opinions. [**Update**: Microsoft now says it's "making adjustments" to Tay in light of this problem.]

# Holistic View of Adversarial Robustness

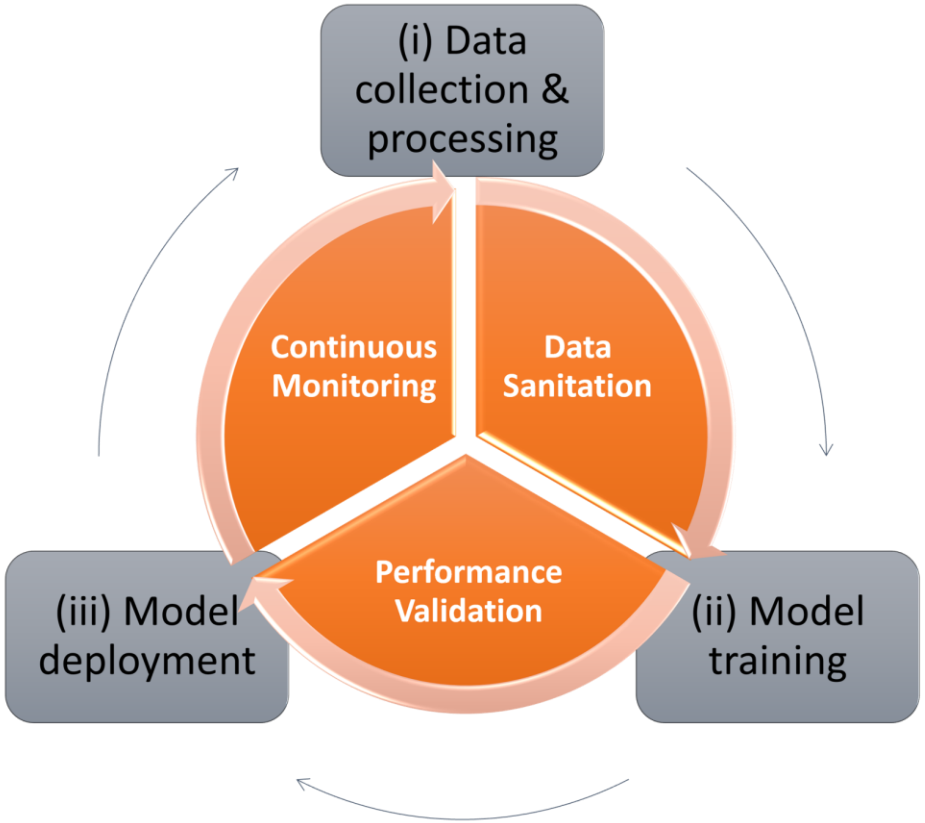



| Attack Category / Attacker's reach | Data | Model / Training Method | Inference |
|---|---|---|---|
| ☑ Poisoning Attack [learning] | X | X* | |
| ☑ Backdoor Attack [learning] | X | | |
| ☑ Evasion Attack (Adversarial Example) [learning] | | X* | X |
| Extraction Attack (Model Stealing, Membership Inference) | | | X |
| Model Injection [AI governance] | | X* | X |

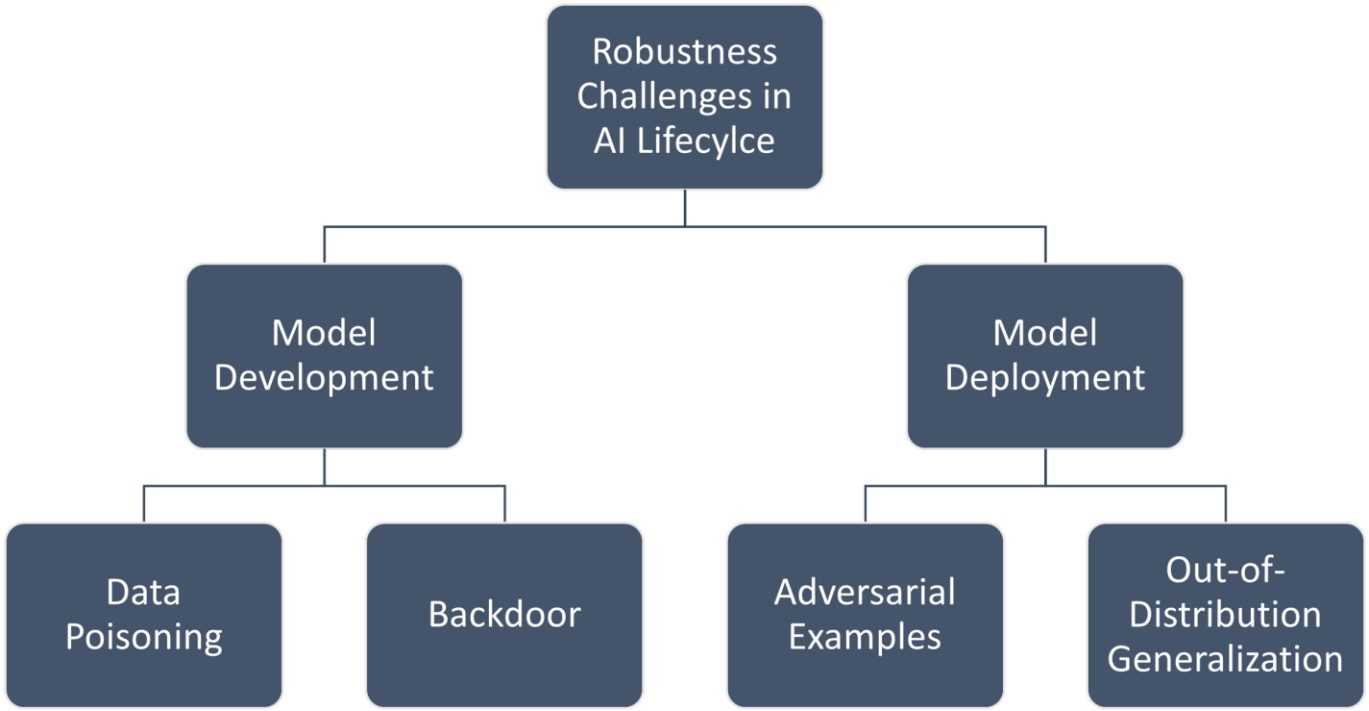
*No access to model internal information in the black-box attack setting

Pin-Yu Chen and Sijia Liu, Holistic Adversarial Robustness of Deep Learning Models, arxiv 2022

# Robustness Challenges in AI Lifecycle




AI Lifecycle & robustness inspection



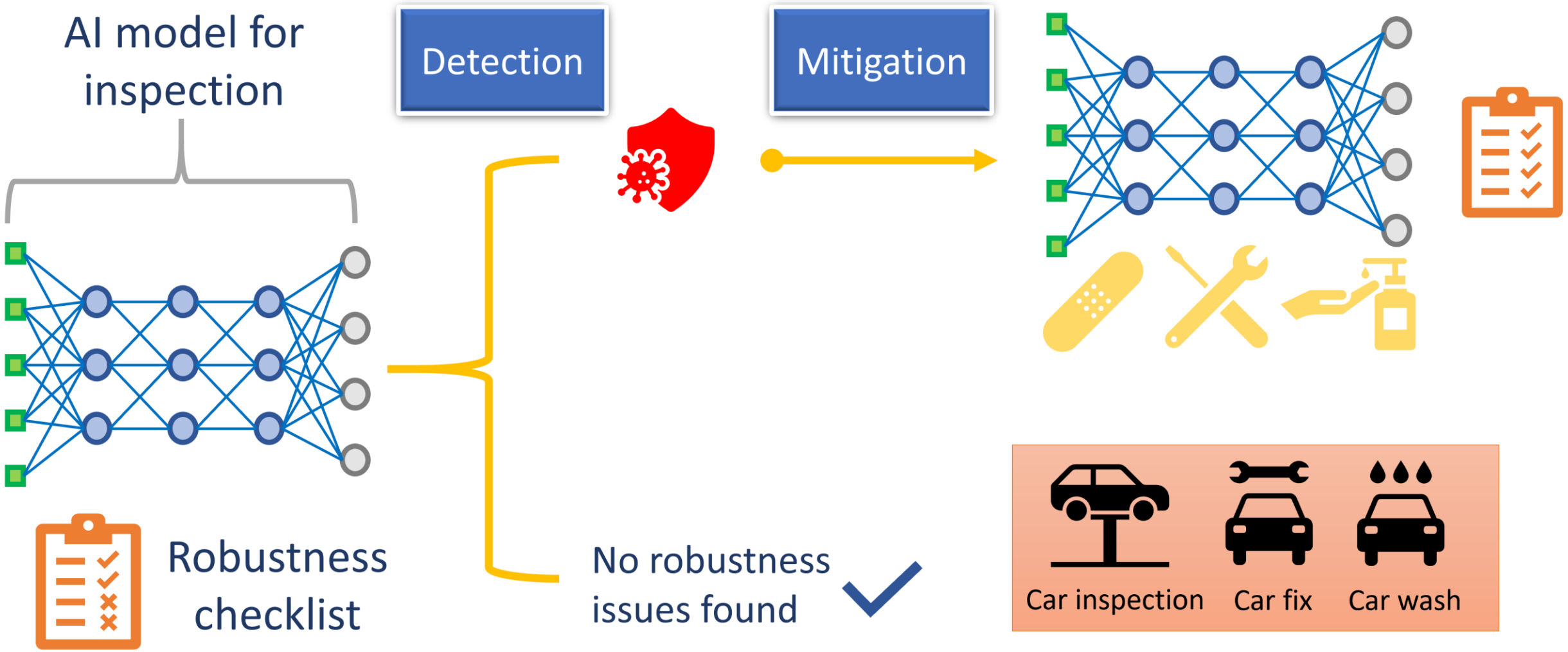
Article

# Cars Require Regular Inspection, Why Should AI Models Be any Different?

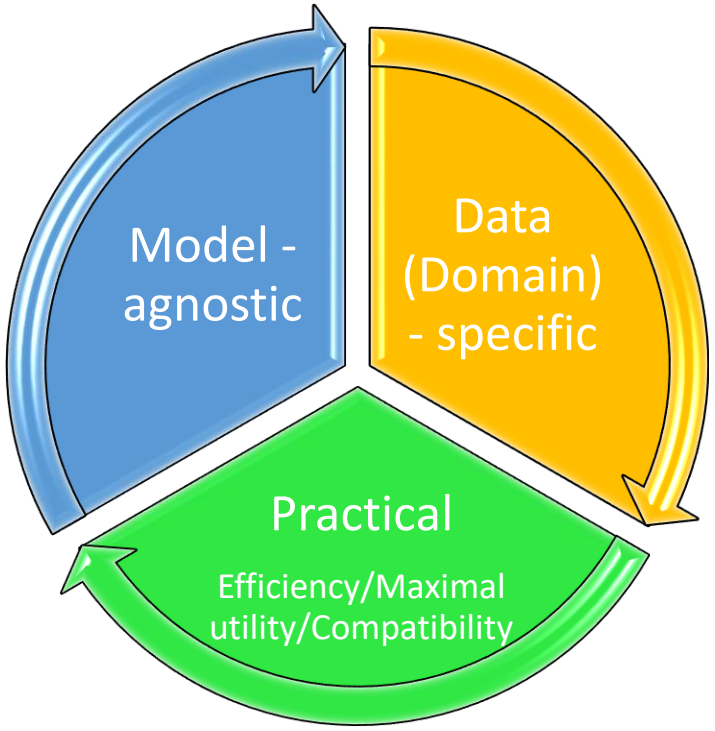
Published: March 14, 2022 | Pin-Yu Chen, PhD, IBM Research

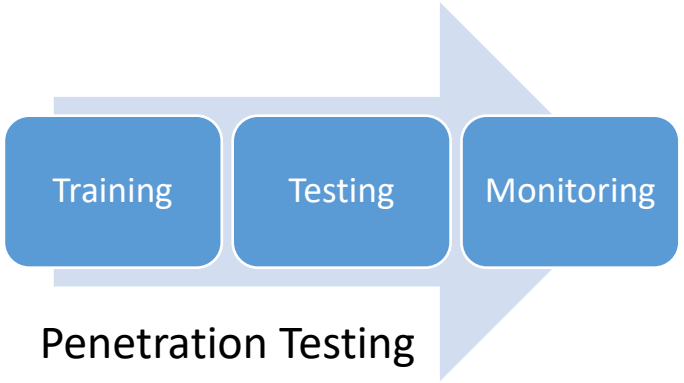
https://www.technologynetworks.com/tn/articles/cars-require-regular-inspection-why-should-ai-models-be-any-different-359405

# Conceptual Pipeline of AI Model Inspector



# Roadmap toward Holistic Adversarial Robustness




**Attack (Bug Finding)**

- In-house **sensitivity and reliability tests** for developed models
- Generate prediction-evasive examples (per user constraints)
- Customize to model deployment conditions (e.g. cloud APIs)

**Defense (Model Hardening)**

- **Detecting and mitigating** potential adversarial threats
- **Plug-and-play** model patching for a given model
- Landscape exploration: model fix and cleaning

**Verification (Model Certificate)**

- This model is certified to be **attack-proof** up to a certain level
- Quantifiable metric for certified robustness
- AI standards, governance, and law regulation

**Applications to AI (Model Boosting)**

- Data augmentation
- **Model reprogramming**: data-efficient transfer learning
- **AI for scientific discovery**

# Research Highlights (1): *Finding failure modes*

- ❑ Practical white-box and black-box robustness testing
- ❑ Principled methods demonstrated on different data modalities and machine learning tasks in digital space and physical world
  - ❑ Images
  - ❑ Texts
  - ❑ Audio/Speech
  - ❑ Graphs
  - ❑ Tabular data
  - ❑ Reinforcement learning
  - ❑ CNN/RNN/LSTM/Transformer





**Original Top-3 inferred captions:**

1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.

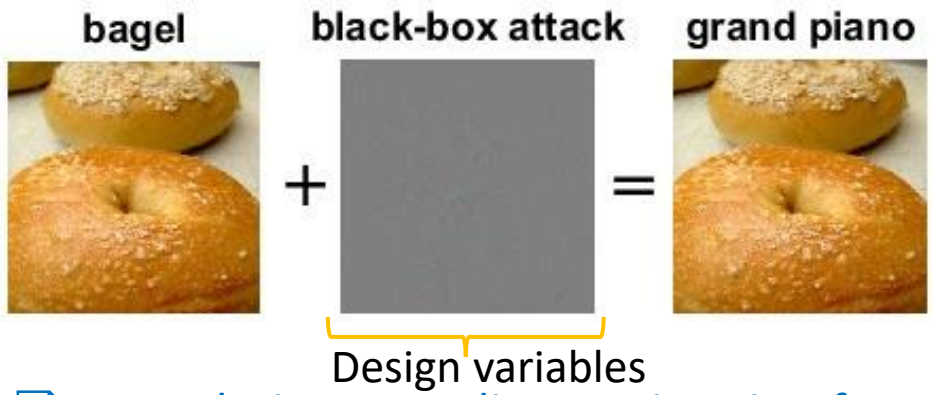
**Adversarial Top-3 captions:**

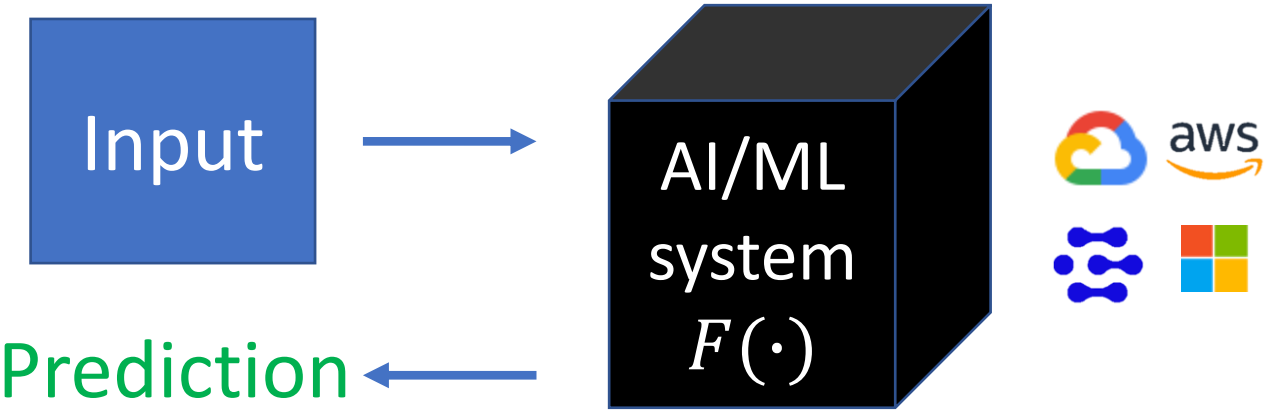
1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.

| Task: Fake-News Detection. Classifier: LSTM. Original label: 100% Fake. ADV label: 77% Real |
|---|
| ~~Man~~ Guy punctuates high-speed chase with stop at In-N-Out Burger drive-thru Print [Ed.~~— Well, that's~~ Okay, that 's a new one.] ~~A~~ One man is in custody after leading police on a bizarre chase into the east Valley on Wednesday night. Phoenix police ~~began~~ has begun following the suspect in Phoenix and the pursuit continue~~d~~ into the east Valley, but it took a bizarre turn when the suspect stopped at an In-N-Out Burger restaurant's ~~drive-thru~~ drive-through near Priest and Ray Roads in Chandler. The suspect appeared to order food, but then drove away and got out of his pickup truck near Rock Wren Way and Ray Road. He ~~then ran into a backyard~~ ran to the backyard and tried to ~~get into a house through the back door~~ get in the home. |

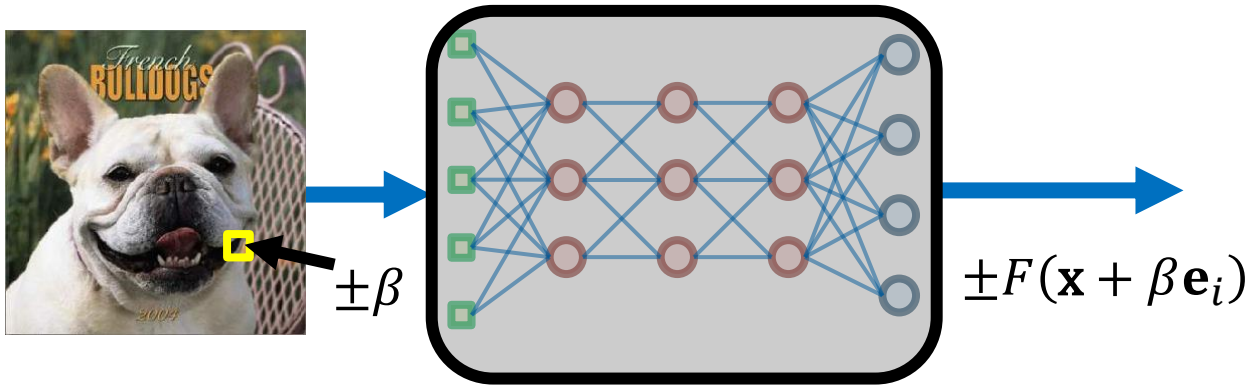
# Inspecting AI/ML systems with **Limited Knowledge:** ZOO Attack

❑Prior to our work, all attacks either require white-box assumption or attack transfer




❑Key technique: gradient estimation from system outputs instead of back-prop



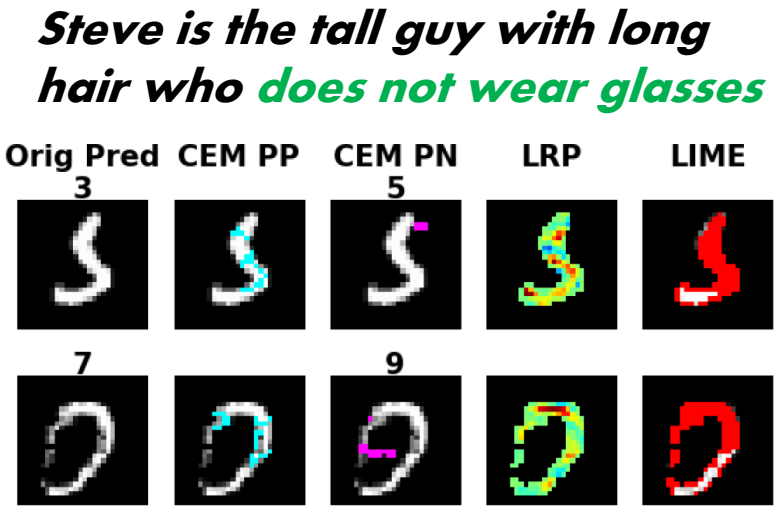

Gradient $g_i := \frac{\partial loss_F(x)}{\partial x_i} \approx \frac{Loss_F(x+\beta\mathbf{e}_i) - Loss_F(x-\beta\mathbf{e}_i)}{2\beta}$

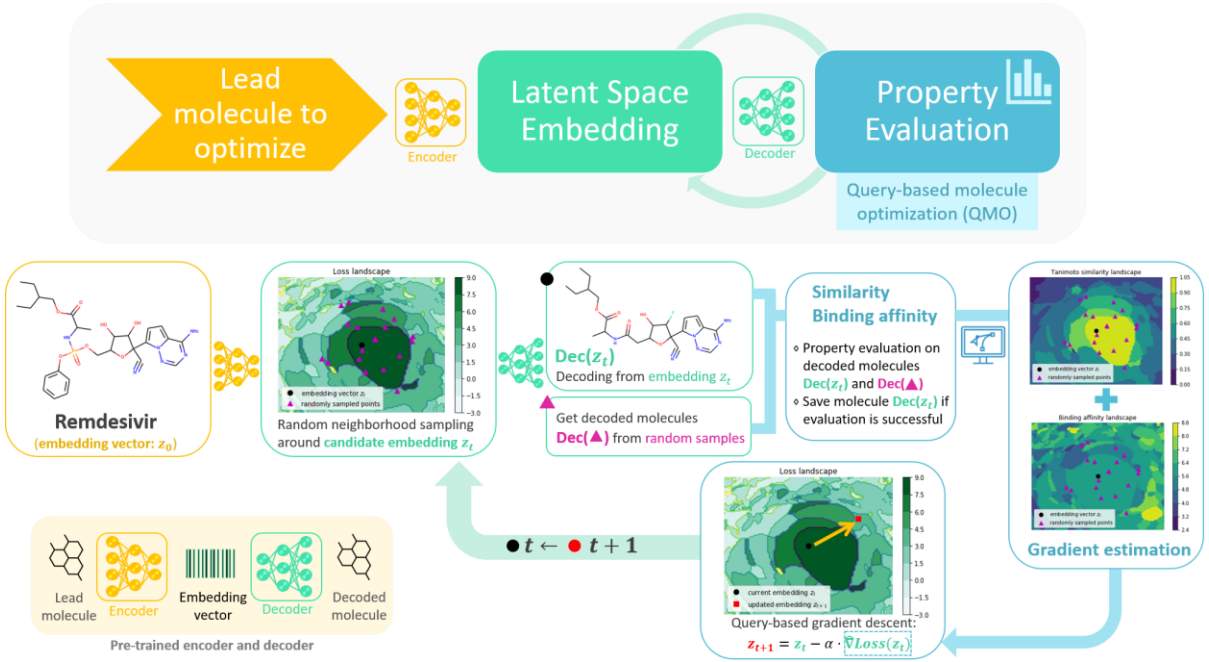
Adversarial example $x_{adv} = x - \eta \cdot \hat{g}$

Pin-Yu Chen*, Huan Zhang, Yash Sharma, Jinfeng Yi, Cho-Jui Hsieh. ZOO: Zeroth Order Optimization based Black-box Attacks to Deep Neural Networks without Training Substitute Models. AI-Security Workshop 2017. Best Paper Award Finalist

# Research Summary

- ❑Practicality: minimal dependency on model and platform (no back-prop, only function calls) – adopted in IBM Watson products
- ❑Shift the focus of the research community from transfer attack to query-based attack
- ❑Inspire many follow-up works on query-efficient and hard-label black-box attacks
- ❑New use case for gradient-free optimization
- ❑Tutorial on "Zeroth-Order Optimization: Theory and Applications to Deep Learning" at CVPR'20 & KDD'19
- ❑Applications: Contrastive explanations and end-to-end molecule optimization using ML models



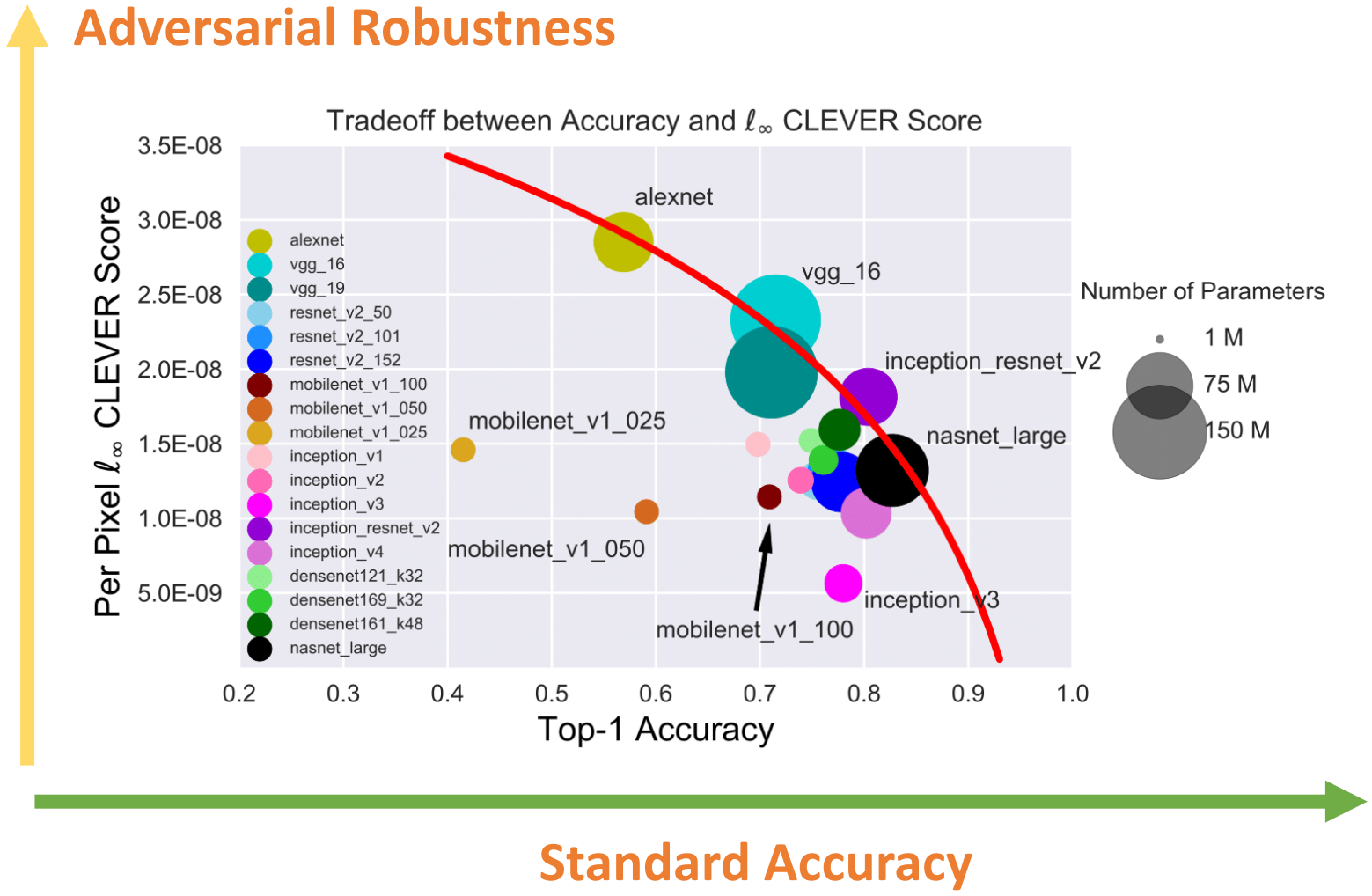


Amit Dhurandhar*, Pin-Yu Chen*, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das, "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives" NeurIPS 2018
Samuel Hoffman, Vijil Chenthamarakshan, Kahini Wadhawan, Pin-Yu Chen, and Payel Das, "Optimizing Molecules using Efficient Queries from Property Evaluations," Nature Machine Intelligence, 2021

# Research Highlights (2)

## *Tradeoff between accuracy and adversarial robustness*

- ❑ Large-scale adversarial robustness benchmarks using 18 ImageNet models
- ❑ Standard Accuracy ≠ Adversarial Robustness
- ❑ Solely pursuing for high-accuracy AI model may harm deployment




Is Robustness the Cost of Accuracy? A Comprehensive Study on the Robustness of 18 Deep Image Classification Models, Dong Su*, Huan Zhang*, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao, ECCV 2018

# Research Summary

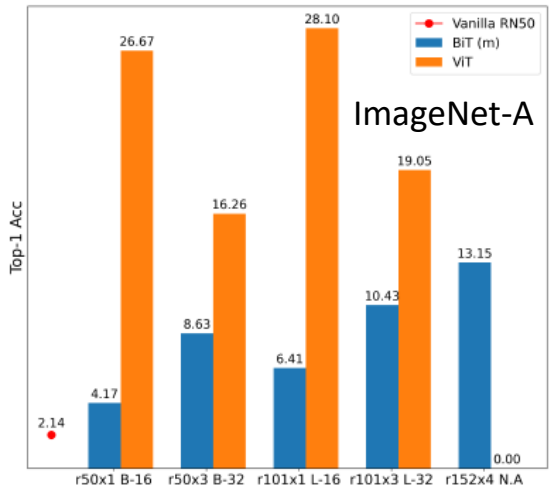
- ❑ Model performance benchmarking beyond standard accuracy
- ❑ From adversarial robustness to general robustness
  - Common corruptions
  - Distribution shifts
  - Semantic changes
  - Out-of-domain samples
- ❑ Tutorial on "Foundational Robustness of Foundation Models" at NeurIPS'22 with Sijia Liu and Sayak Paul
- ❑ Tutorial on "Practical Adversarial Robustness in Deep Learning: Problems and Solutions" at CVPR'21 with Sayak Paul
- ❑ Tutorial on "Adversarial Robustness of Deep Learning Models" at ECCV'20
- ❑ Lecturer at MLSS 2021

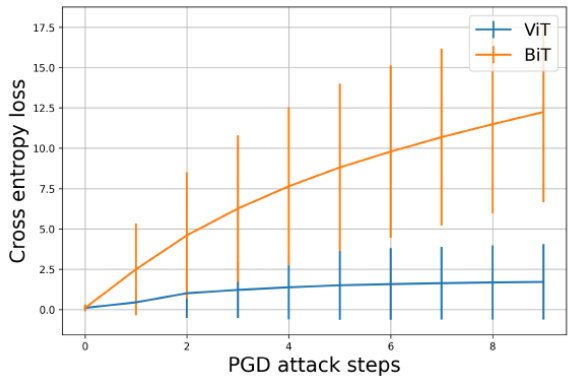
## Vision Transformers are Robust Learners

**Sayak Paul***
PyImageSearch
s.paul@pyimagesearch.com

**Pin-Yu Chen***
IBM Research
pin-yu.chen@ibm.com

| Dataset | Purpose |
|---|---|
| ImageNet-C [13] | Common corruptions |
| ImageNet-P [13] | Common perturbations |
| ImageNet-R [14] | Semantic shifts |
| ImageNet-O [9] | Out-of-domain distribution |
| ImageNet-A [9] | Natural adversarial examples |
| ImageNet-9 [15] | Background dependence |



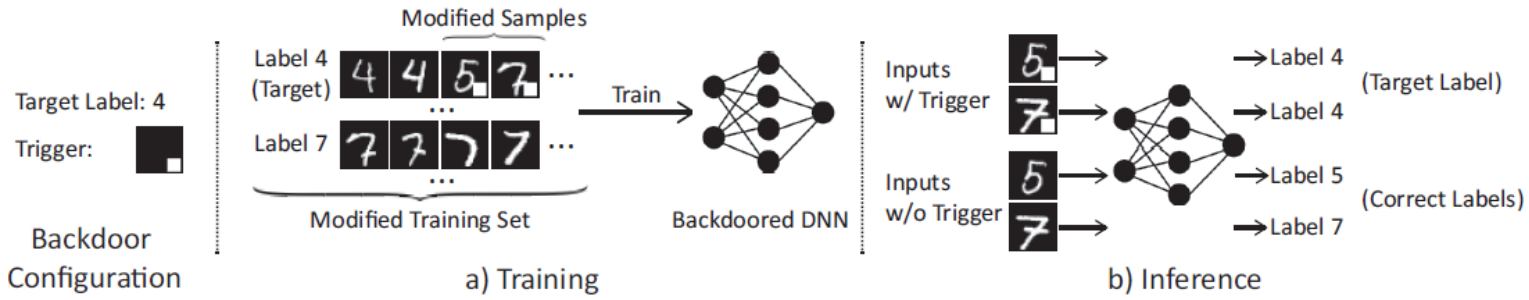




Sayak Paul* and Pin-Yu Chen*. "Vision Transformers are Robust Learners," AAAI 2022 (*equal contribution)

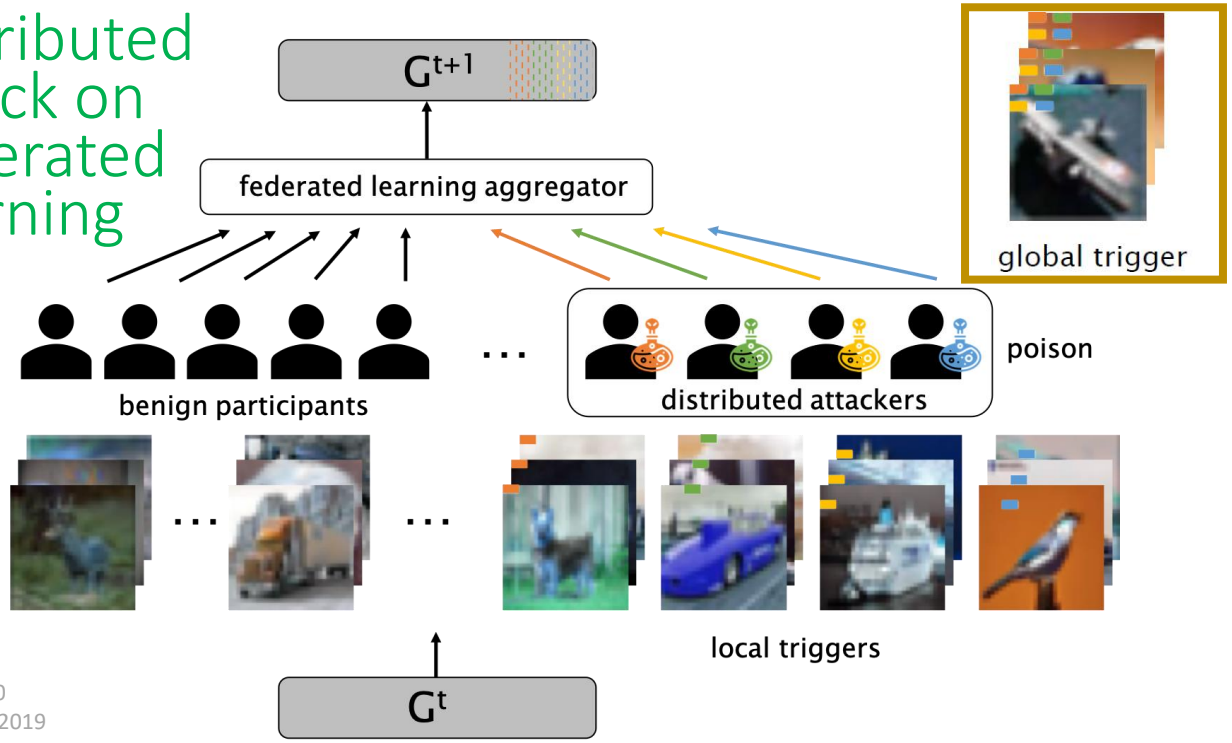
# Research Highlights (3): *Practical adversarial threat detection and mitigation*

- ❑ Plug-and-play techniques to inspect and repair potential risks in trained neural networks
- ❑ Data efficiency: threat detection and mitigation using limited data
- ❑ Applicable to training-time and inference-time threats

## Backdoor Attack




## Distributed Attack on Federated Learning





Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. DBA: Distributed Backdoor Attacks against Federated Learning. ICLR 2020

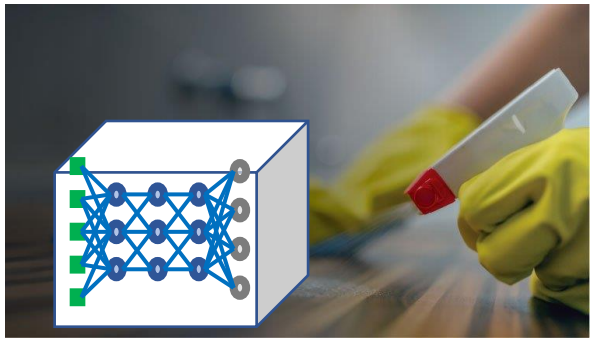
Wang et al. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. IEEE Security and Privacy, 2019

# Research Summary

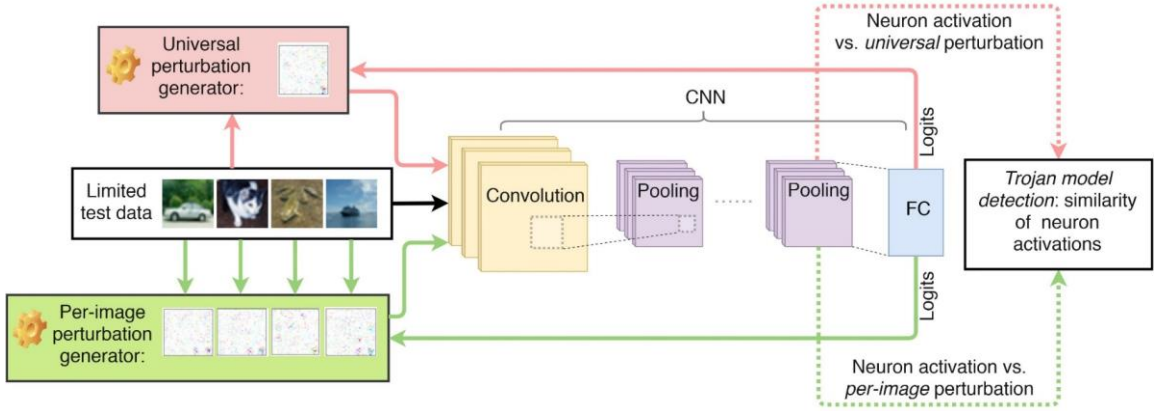
- *I have an amazing ImageNet model which achieves 95% top-1 accuracy, and I make it publicly available by releasing the network architecture and trained model weights. Care to use it for your task?*

➢ Tempting … but how do I know your model does not have any backdoor?

✓ Sanitize the model before using it (aka wear mask before you go out)

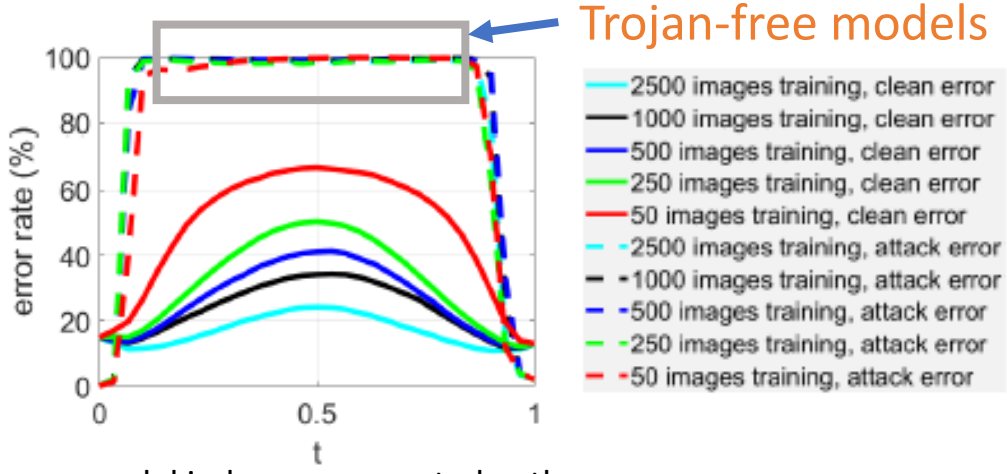



## Data-limited and Data-free Trojan net detection




## Mode connectivity-based model finetuning

Trojan-free models



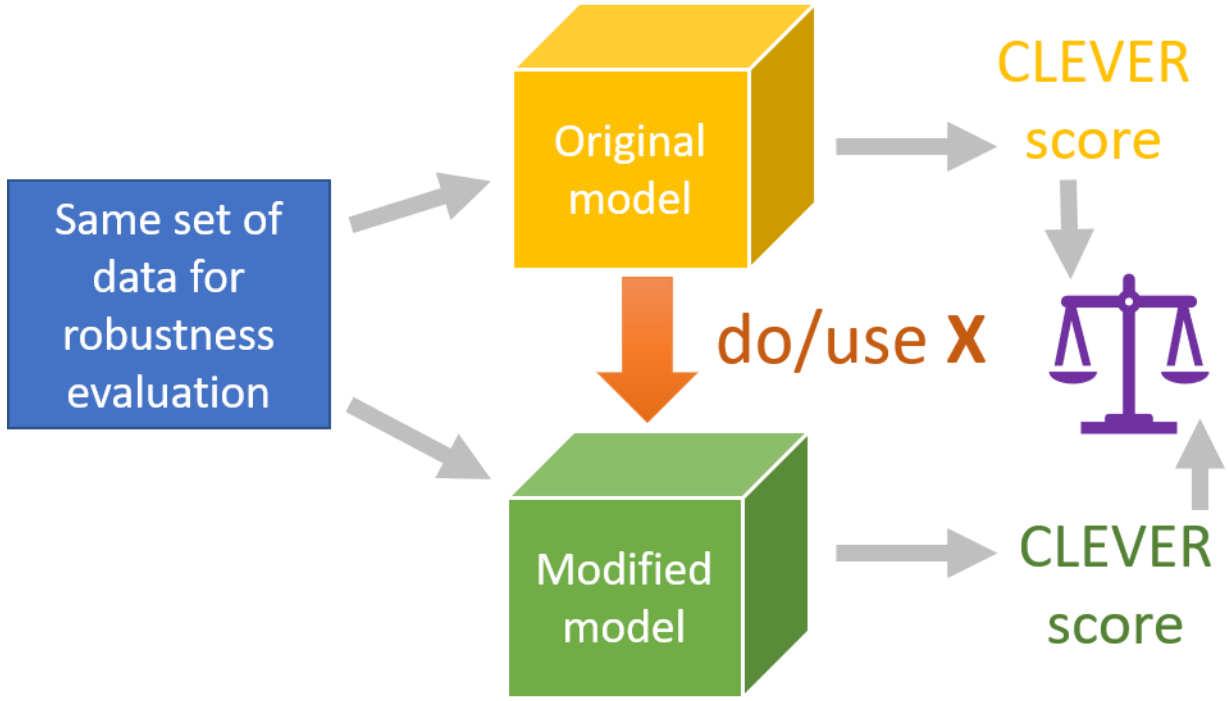

model index on connected path

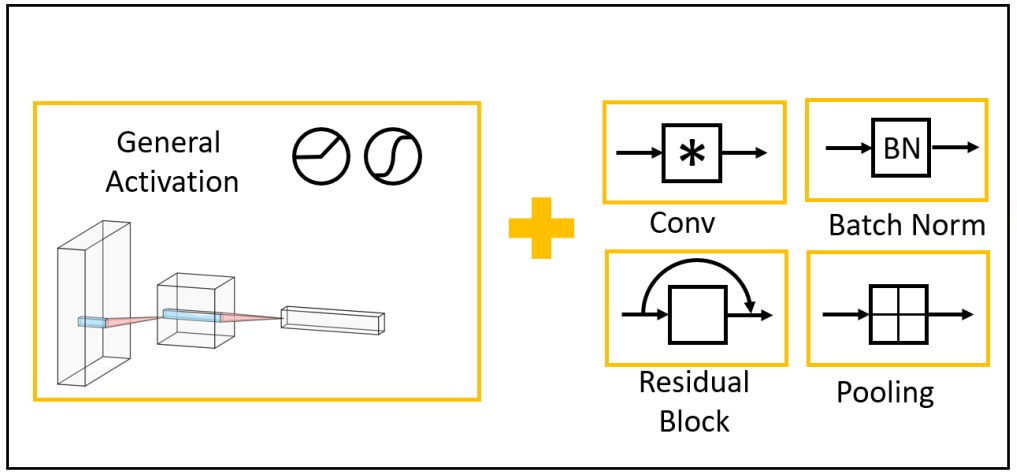
Ren Wang, Gaoyuan Zhang, Sijia Liu, Pin-Yu Chen, Jinjun Xiong, and Meng Wang. Practical Detection of Trojan Neural Networks: Data-Limited and Data-Free. *ECCV 2020*

Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging Mode Connectivity in Loss Landscapes and Adversarial Robustness. ICLR 2020

# Research Highlights (4):
## *Provable robustness assessment and quantification*

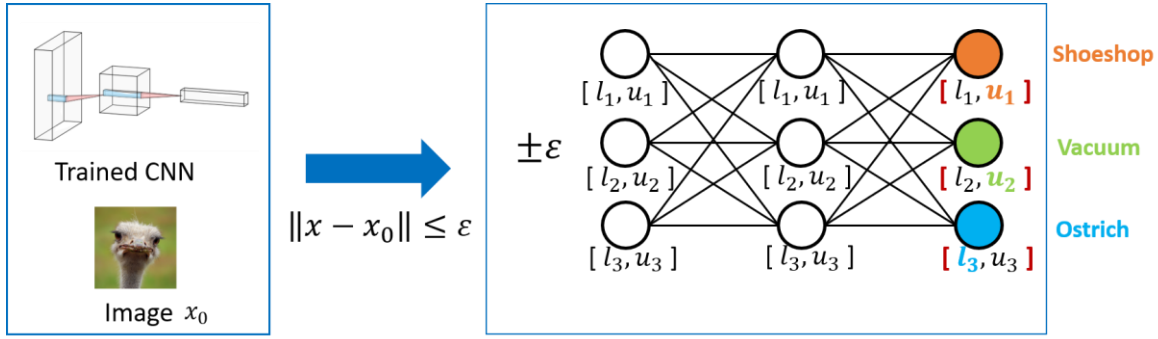
- ❑ Avoid false sense of robustness from improper empirical measures
- ❑ Certified robustness: attack-proof risk quantification and verification for neural networks
- ❑ Legitimate answer to “Will my model become more robust if I do/use X?”



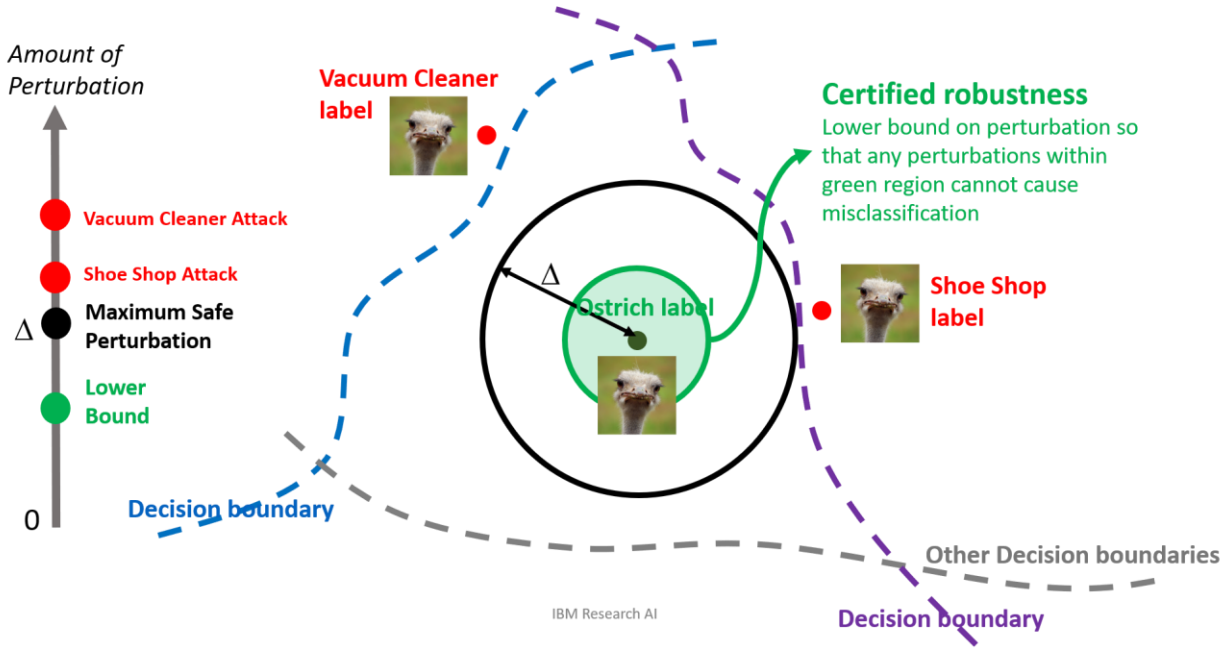



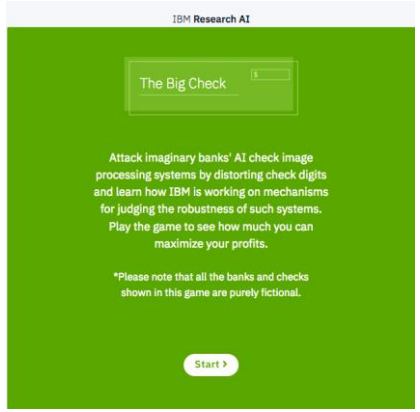

# Research Summary

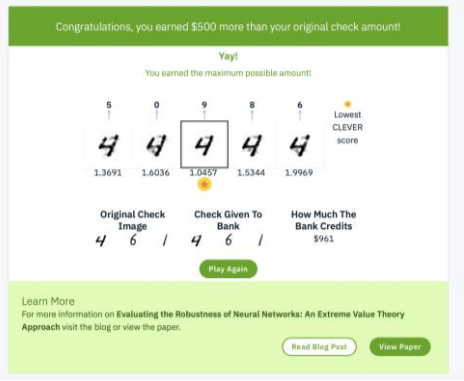
- CLEVER score: First attack-independent and model-agnostic robustness score
- Robustness certification tools
  - Input perturbations
  - Semantic perturbations
  - Training of verifiable neural networks
  - Certified defenses
  - Support for different network architectures
- AI regulation and standardization




## Verification: lower bounds on robustness




http://bigcheck.mybluemix.net

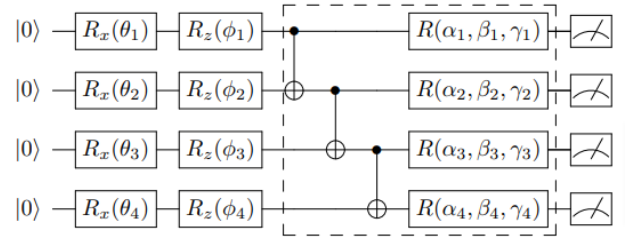

Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach, Tsui-Wei Weng*, Huan Zhang*, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Guo, Cho-Jui Hsieh, and Luca Daniel, ICLR 2018
Efficient Neural Network Robustness Certification with General Activation Functions, Huan Zhang*, Tsui-Wei Weng*, Pin-Yu Chen, Cho-Jui Hsieh and Luca Daniel, NeurIPS 2018

# Research Highlights (5):
## *(Hybrid) Quantum ML: Better Robustness, Privacy, and Generalization?*

- ❑ Hybrid Quantum ML model: quantum circuits for feature extraction, followed by classical neural networks
- ❑ Parameter efficiency; Generalization Guarantee
- ❑ Better privacy and robustness?

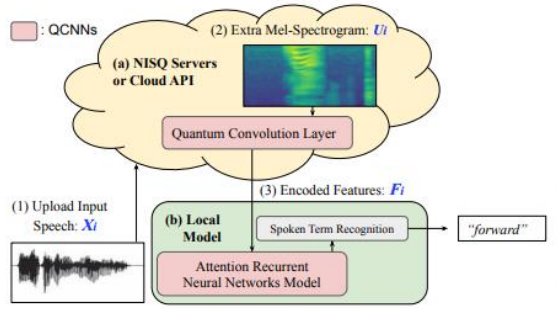
**Variational Quantum Circuits for Deep Reinforcement Learning**

SAMUEL YEN-CHI CHEN[1], CHAO-HAN HUCK YANG[2], JUN QI[2], PIN-YU CHEN[3], XIAOLI MA[2], (Fellow, IEEE), HSI-SHENG GOAN[1,4],



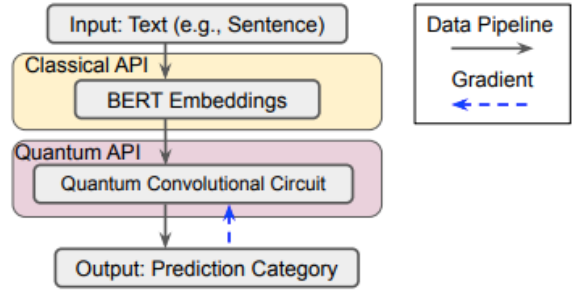

**DECENTRALIZING FEATURE EXTRACTION WITH QUANTUM CONVOLUTIONAL NEURAL NETWORK FOR AUTOMATIC SPEECH RECOGNITION**

*Chao-Han Huck Yang*[1] *Jun Qi*[1] *Samuel Yen-Chi Chen*[2] *Pin-Yu Chen*[3]
*Sabato Marco Siniscalchi*[1,4,5] *Xiaoli Ma*[1] *Chin-Hui Lee*[1]




**WHEN BERT MEETS QUANTUM TEMPORAL CONVOLUTION LEARNING FOR TEXT CLASSIFICATION IN HETEROGENEOUS COMPUTING**

*Chao-Han Huck Yang*[1] *Jun Qi*[1] *Samuel Yen-Chi Chen*[2]
*Yu Tsao*[3] *Pin-Yu Chen*[4]




Quantum Neural Networks for Speech and Natural Language Processing (QuantumNN) Tutorial

Montreal, 21st (Sat.) August (Virtual Room Auditorium Red), IJCAI, 2021

# THEORETICAL ERROR PERFORMANCE ANALYSIS FOR VARIATIONAL QUANTUM CIRCUIT BASED FUNCTIONAL REGRESSION

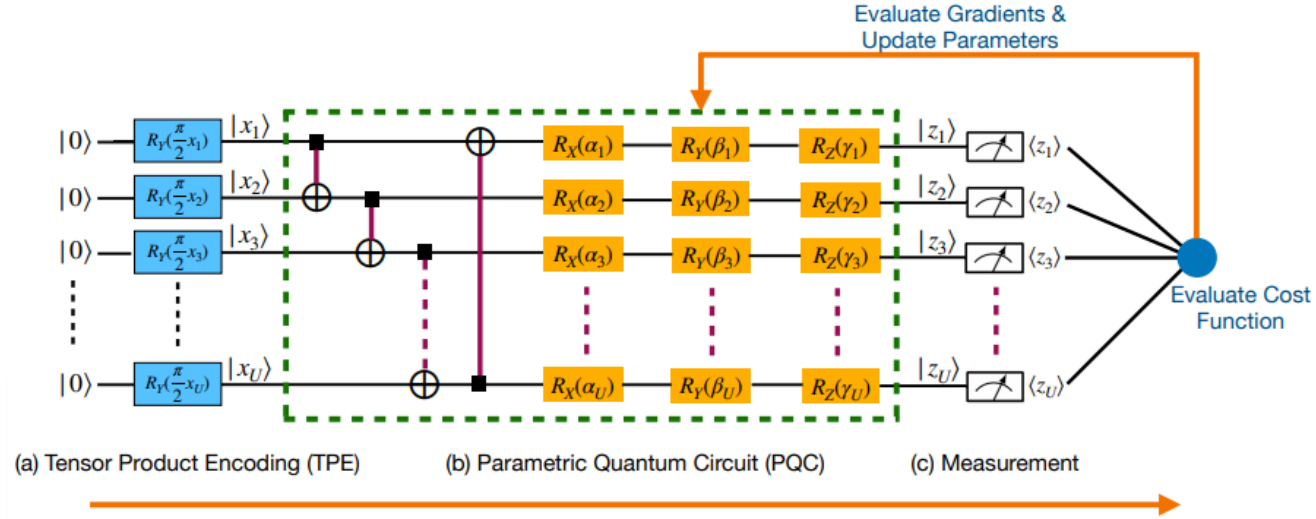
A PREPRINT

**Jun Qi**
Georgia Institute of Technology
jqi41@gatech.edu

**Chao-Han Huck Yang**
Georgia Institute of Technology
huckiyang@gatech.edu

**Pin-Yu Chen**
IBM Research
pin-yu.chen@ibm.com

**Min-Hsiu Hsieh**
Hon Hai (Foxconn) Research Institute
minhsiuh@gmail.com




# QTN-VQC: AN END-TO-END LEARNING FRAMEWORK FOR QUANTUM NEURAL NETWORKS

A PREPRINT

**Jun Qi**
Georgia Institute of Technology
jqi41@gatech.edu

**Chao-Han Huck Yang**
Georgia Institute of Technology
huckiyang@gatech.edu

**Pin-Yu Chen**
IBM Research
pin-yu.chen@ibm.com

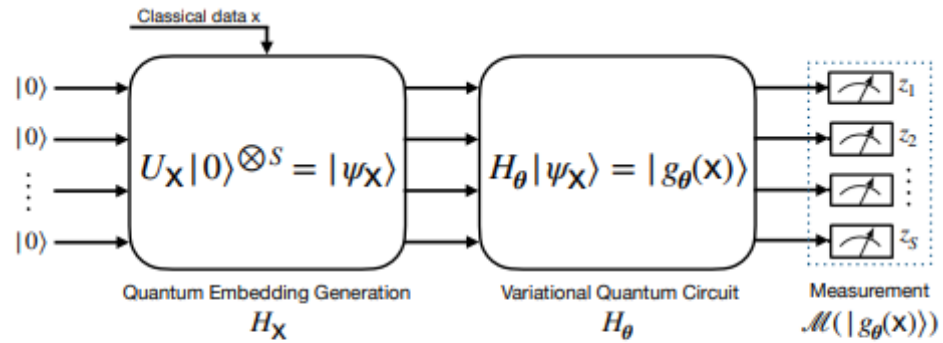
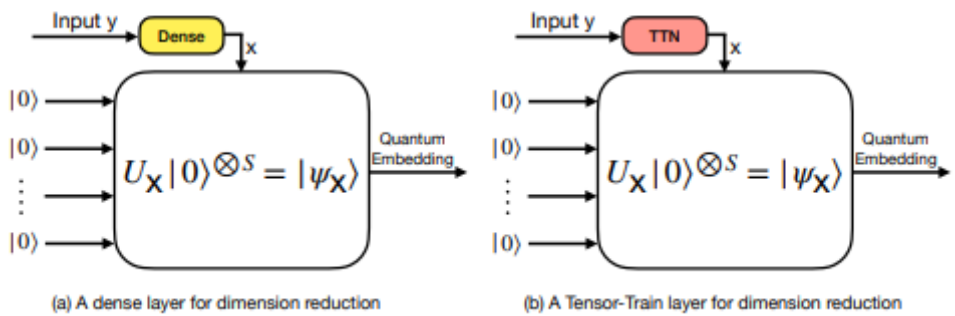
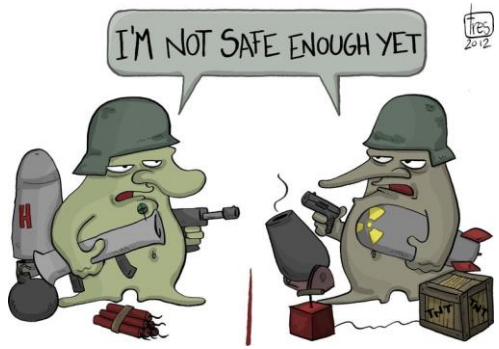


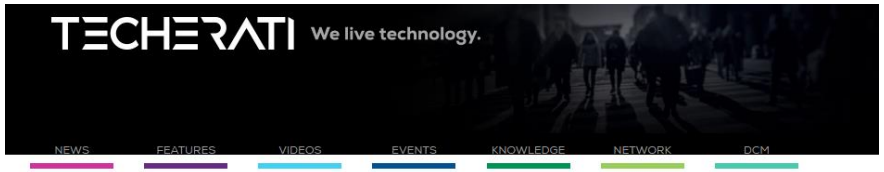

Figure 1: *An illustration of QNN based on VQC.*



# Our portfolio in adversarial robustness research

- 50+ papers at top AI/ML conferences in 2018-2022 (NeurIPS, ICML, AAAI, ICLR, IJCAI, ACL, ECCV, ICCV, CVPR, ICASSP, …)
- UAI 2022 Best Paper Runner-Up Award
- Open-Source Library (ART, AIX)
- Tutorials (NeurIPS, CVPR, ECCV, AAAI, KDD, IJCAI)


FEATURES HUB OPINION AI BIG DATA

Unmasking Adversarial AI with Pin-Yu Chen



NEWS · 10 MAY 2019

AI can now defend itself against malicious messages hidden in speech

*Computer scientists have thwarted programs that can trick AI systems into classifying malicious audio as safe.*


AI GUEST

Text-based AI models are vulnerable to paraphrasing attacks, researchers find

BEN DICKSON, TECHTALKS @BENDEE983 APRIL 1, 2019 3:10 PM


DESIGNLINES | AI & BIG DATA DESIGNLINE

AI Tradeoff: Accuracy or Robustness?


Home › Interviews › Robust AI: Protecting neural networks against adversarial attacks

Interviews

Robust AI: Protecting neural networks against adversarial attacks

By Ben Dickson - February 20, 2019

https://www.ucc.ie/en/cirtl/newsandevents/cirtl-seminar-the-assessment-arms-race-and-its-fallout-the-case-for-slow-scholarship-may-14th.html

# Adversarial Robustness Toolbox (ART)

## Adversarial Robustness Toolbox (ART)

External: https://github.com/IBM/adversarial-robustness-toolbox

- Python library, 7K lines of code
- State-of-the-art attacks, defences and robustness metrics

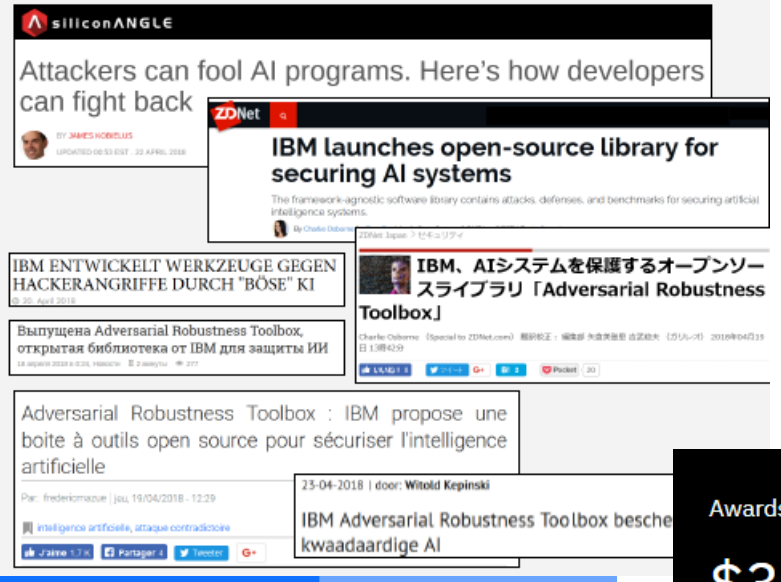
Load ART modules →
Load classifier model (Keras, TF, PyTorch etc) →
Perform attack →
Evaluate robustness →

```
from keras.datasets import mnist
from keras.models import load_model

from art.attacks import CarliniL2Attack
from art.classifier import KerasClassifier
from art.metrics import loss_sensitivity

# Load data
(_, _), (x_test, y_test) = mnist.load_data()

# Load model and build classifier
model = load_model('my_favorite_keras_model.h5')
classifier = KerasClassifier((0, 1), model)

# Perform attack
attack = CarliniL2Attack(classifier)
adv_x_test = attack.generate(x_test)

# Compute metrics on model robustness
print(loss_sensitivity(classifier, x_test))
```

- Robustness service based on ART on the roadmap for general availability under IBM AI OpenScale offering

Open-source release @ RSA 2018:

- ~ 3.5K+ views of IBM blogs
- ~ 100+ news outlets covering release
- ~ 1.3M+ Social Media potential impressions
- ~ 5K+ views of GitHub repo




## IBM donates "Trusted AI" projects to Linux Foundation AI

As real-world AI deployments increase, IBM says the contributions can help ensure they're fair, secure and trustworthy.

| Adversarial Robustness 360 | AI Fairness 360 | AI Explainability 360 |
|---|---|---|
| ↳ (ART) | ↳ (AIF360) | ↳ (AIX360) |
| github.com/IBM/adversarial-robustness-toolbox | github.com/IBM/AIF360 | github.com/IBM/AIX360 |
| art-demo.mybluemix.net | aif360.mybluemix.net | aix360.mybluemix.net |

LF AI

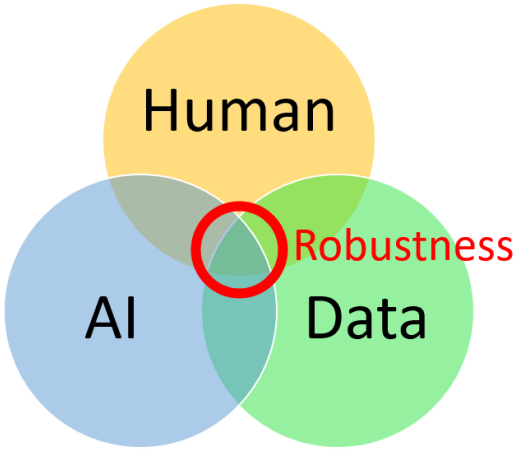
| Evasion attacks | Evasion defenses | Poisoning detection | Robustness metrics |
|---|---|---|---|
| • FGSM | • Feature squeezing | • Detection based on clustering activations | • CLEVER |
| • JSMA | • Spatial smoothing | | • Empirical robustness |

Awards and Prizes

## $3.4M DARPA Grant Awarded to IBM to Defend AI Against Adversarial Attacks

# Take-aways

- ✓Adversarial robustness is a new AI standard toward trustworthy ML
- ❑Robustness does not come for free: failure modes exist in digital space, physical world, and different domains during AI lifecycle
- ❑High accuracy ≠ Good robustness
- ❑Arms race in adversarial ML: adversary-aware AI v.s. AI for adversary
- ✓AI model inspector for holistic robustness and beyond
- ❑Practical techniques and tools for identifying failure modes
- ❑Plug-and-play model error detection and mitigation
- ❑Robustness quantification and verification
- ❑Adversarial ML for good: model reprogramming
- ✓My long-term goal: Make the robustness inspection pipeline for AI models as reliable, standard, and easy, as car models




# Online Resources for Adversarial Robustness

- J. Z. Kolter and A. Madry: Adversarial Robustness - Theory and Practice (NeurIPS 2018 Tutorial)
- Pin-Yu Chen: Adversarial Robustness of Deep Learning Models (ECCV 2020 Tutorial)
- Pin-Yu Chen and Sijia Liu: Zeroth Order Optimization: Theory and Applications to Deep Learning (CVPR 2020 Tutorial)
- Pin-Yu Chen and Sayak Paul: Practical Adversarial Robustness in Deep Learning: Problems and Solutions (CVPR 2021 Tutorial)
- Pin-Yu Chen: Holistic Adversarial Robustness for Deep Learning (MLSS 2021 Tutorial)
- Pin-Yu Chen: Adversarial Machine Learning for Good (AAAI 2022 Tutorial)
- Pin-Yu Chen, Sijia Liu, and Sayak Puak: Foundational Robustness of Foundation Models (NeuIPS 2022 Tutorial)


Adversarial Robustness Toolbox (ART v0.10.0)

Foolbox

# Sample Surveys for Adversarial Robustness

Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning

Battista Biggio[a,b,*], Fabio Roli[a,b]

[a]Department of Electrical and Electronic Engineering, University of Cagliari, Italy
[b]Pluribus One, Cagliari, Italy

ON EVALUATING ADVERSARIAL ROBUSTNESS

Nicholas Carlini[1], Anish Athalye[2], Nicolas Papernot[1], Wieland Brendel[3], Jonas Rauber[3], Dimitris Tsipras[2], Ian Goodfellow[1], Aleksander Mądry[2], Alexey Kurakin[1]*

[1] Google Brain [2] MIT [3] University of Tübingen

**Holistic Adversarial Robustness of Deep Learning Models**

**Pin-Yu Chen**[1,3]*, **Sijia Liu**[2,3]

[1]IBM Research, [2]Michigan State University, [3]MIT-IBM Watson AI Lab
pin-yu.chen@ibm.com and liusiji5@msu.edu

On Adaptive Attacks to Adversarial Example Defenses

Florian Tramèr* Stanford University
Nicholas Carlini* Google Brain
Wieland Brendel* University of Tübingen
Aleksander Mądry MIT

**The Robustness of Deep Networks**

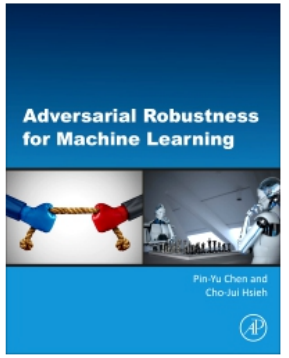
*A geometrical perspective*

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard

**Adversarial Learning Targeting Deep Neural Network Classification: A Comprehensive Review of Defenses Against Attacks**

**Publisher: IEEE** Cite This PDF

3 Author(s) David J. Miller ; Zhen Xiang ; George Kesidis View All Authors




Adversarial Robustness for Machine Learning

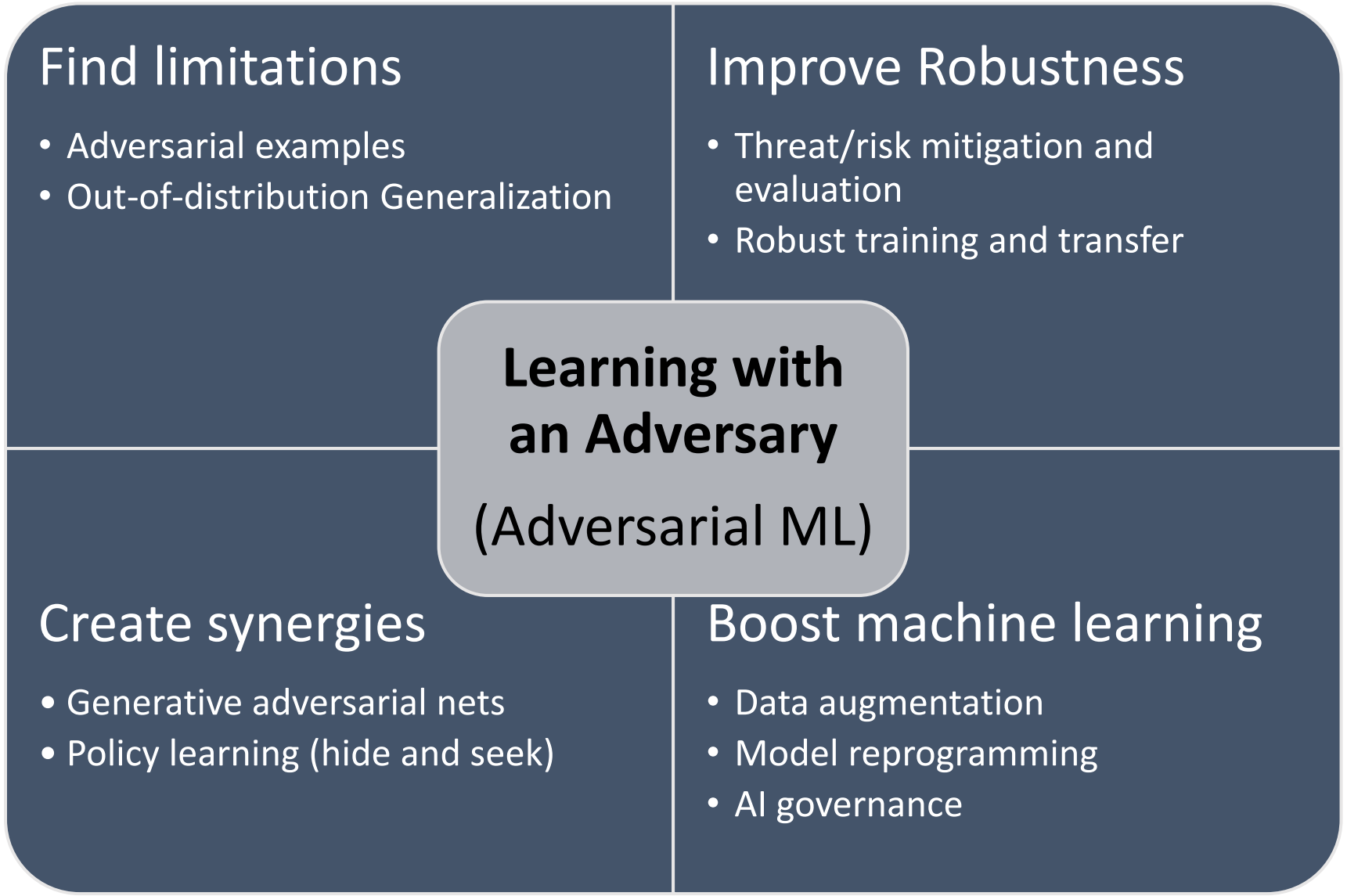
1st Edition - September 1, 2022

☆☆☆☆ Write a review

☆

Authors: Pin-Yu Chen, Cho-Jui Hsieh

Paperback ISBN: 9780128240205


Find limitations
• Adversarial examples
• Out-of-distribution Generalization
Improve Robustness
• Threat/risk mitigation and evaluation
• Robust training and transfer
Learning with an Adversary
(Adversarial ML)
Create synergies
• Generative adversarial nets
• Policy learning (hide and seek)
Boost machine learning
• Data augmentation
• Model reprogramming
• AI governance