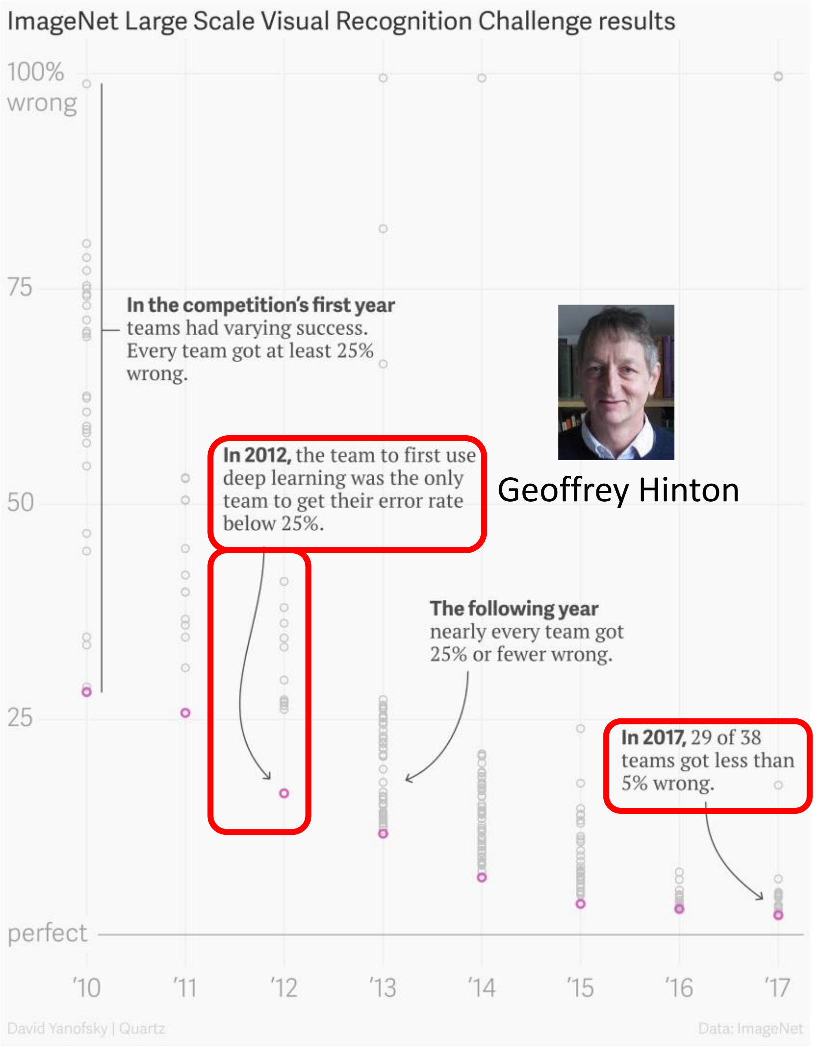
Holistic Adversarial Robustness of AI Models

Pin-Yu Chen
www.pinyuchen.com @pinyuchenTW

NIST AI Measurement and Evaluation Workshop
June 2021

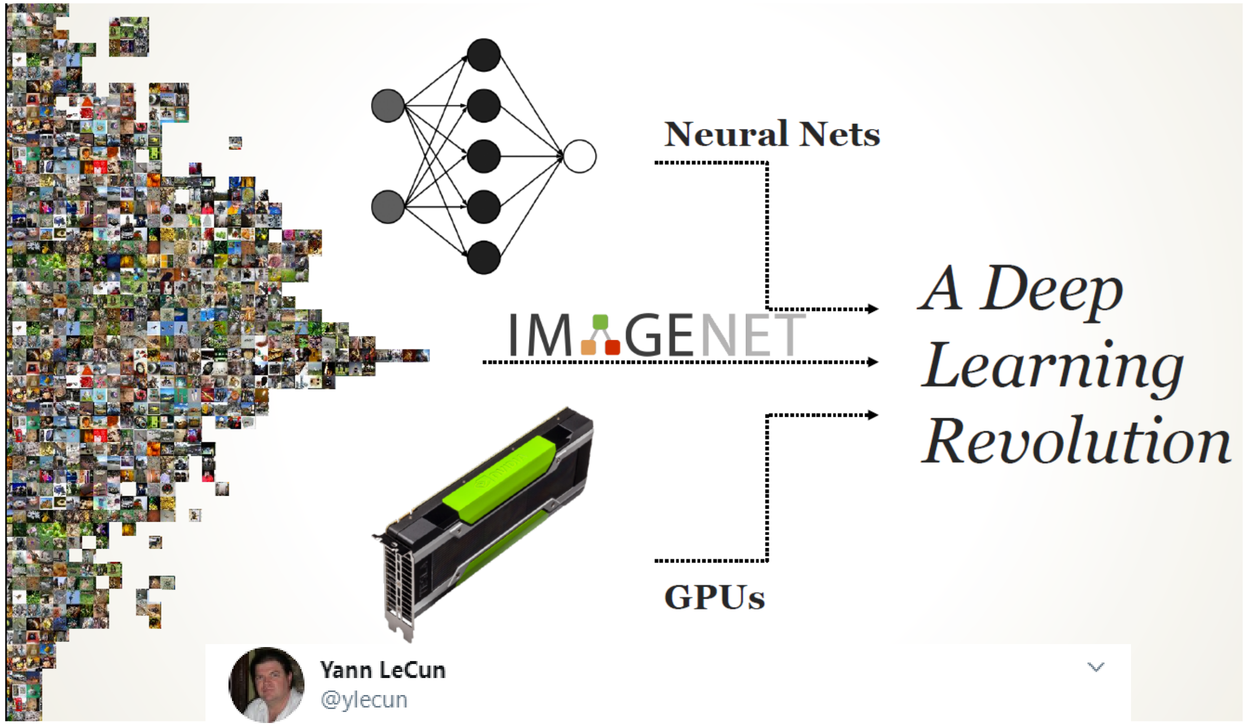
IBM Research

The Deep Learning Revolution. What's next?



Geoffrey Hinton

What's Next?



Yann LeCun
@ylecun

Replying to @ylecun @GaryMarcus and @titudeadjust

DL is not an "algorithm". It's merely the concept of building a machine by assembling parameterized functional blocks and training them with some sort of gradient-based optimization method. That's it.
You are free to choose your architecture, learning paradigm, prior, etc...1/2

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

Harvard Business Review

Coronavirus Magazine Popular Topics Podcasts Video Store The Bi

RISK MANAGEMENT

The Case for AI Insurance

by Ram Shankar Siva Kumar and Frank Nagle

April 29, 2020

The Great Adversarial Examples

ostrich | safe | shoe shop | vacuum

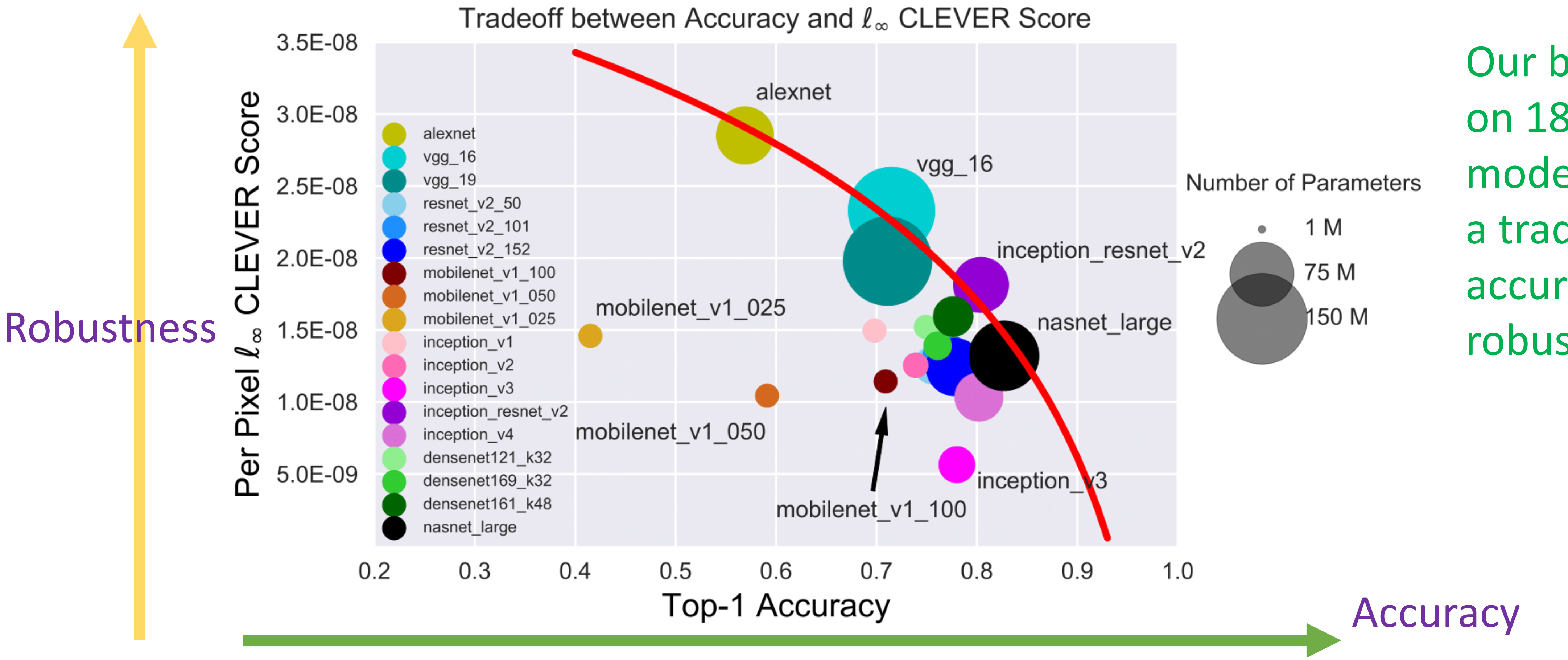
What is wrong with this AI model?

- This model is one of the BEST image classifier using neural networks
- Images and neural network models are NOT the only victims

EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples, P.-Y. Chen*, Y. Sharma*, H. Zhang, J. Yi, and C.-J. Hsieh, AAAI 2018

Accuracy ≠ Adversarial Robustness

- Solely pursuing for high-accuracy AI model may get us in trouble…



Our benchmark on 18 ImageNet models reveals a tradeoff in accuracy and robustness

Why adversarial (worst-case) robustness matters?

➢ Prediction-evasive manipulation on a deployed AI model

1. Build trust in AI: address inconsistent perception and decision making between humans and machines & misinformation
2. Assess negative impacts in high-stakes, safety-critical tasks
3. Understand limitation in current machine learning methods
4. Prevent loss in revenue and reputation
5. Ensure safe and responsible use in AI

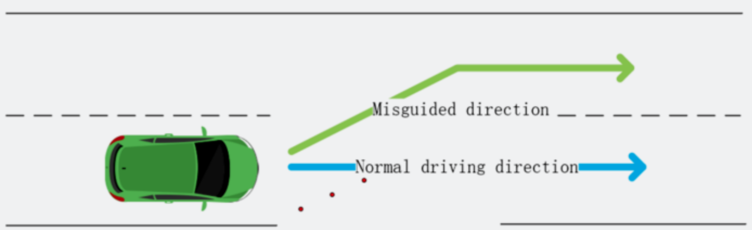
Adversarial T-shirt

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

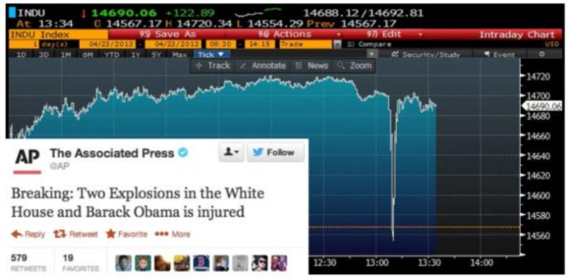


The Washington Post

Democracy Dies in Darkness

WorldViews

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



This chart shows the Dow Jones Industrial Average during Tuesday afternoon's drop, caused by a fake A.P. tweet, inset at left.

By Max Fisher

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

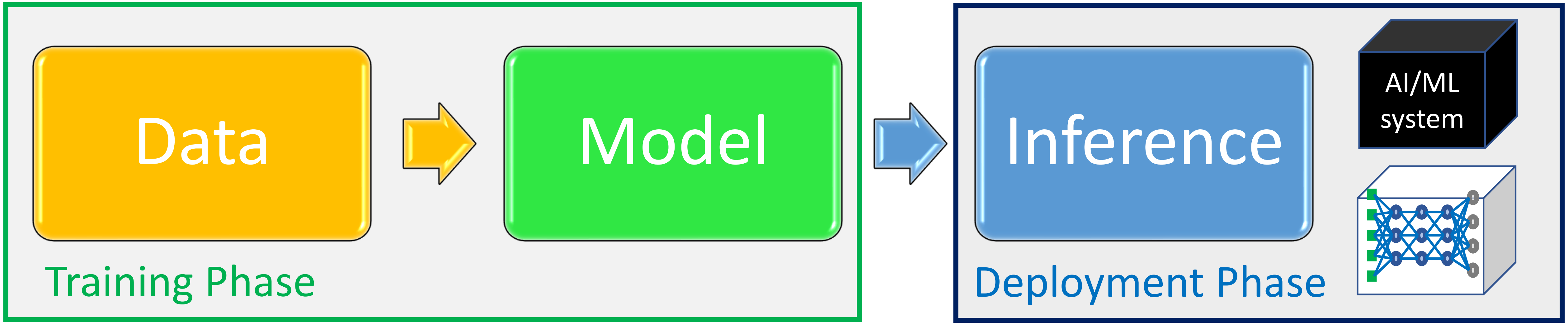
Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

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.]

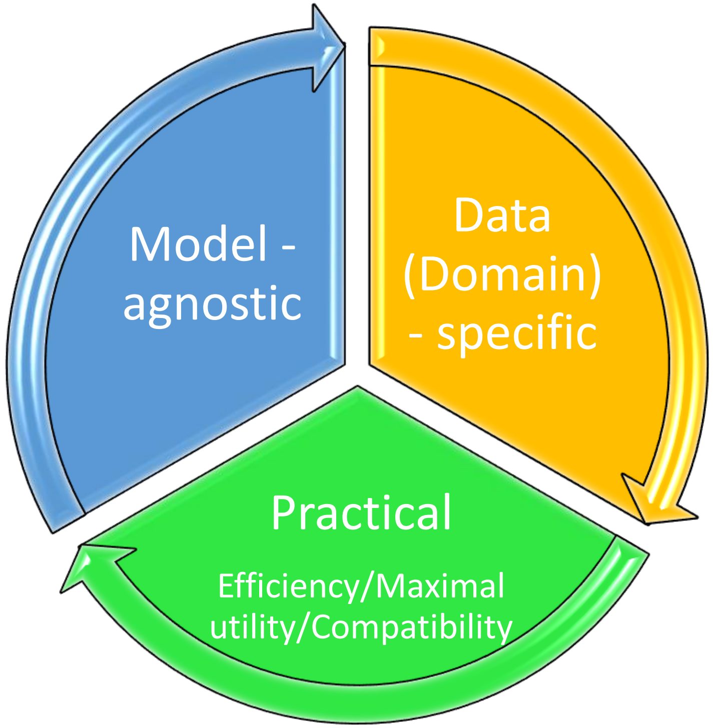
Holistic View of Adversarial Robustness

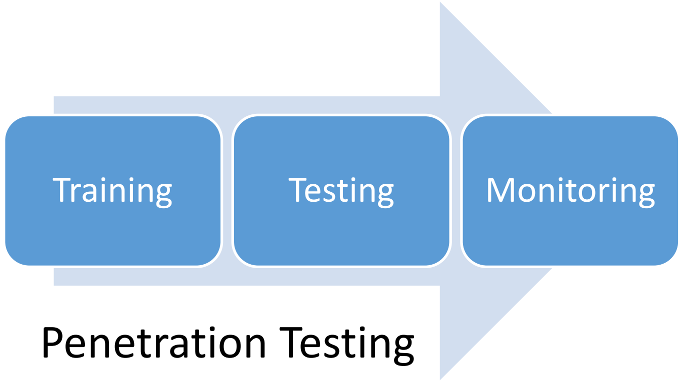


Attack Category / Attacker's reach	Data	Model / Training Method	Inference
Poisoning Attack [learning]	X	X*	
Backdoor Attack [learning]	X		
Evasion Attack (Adversarial Example) [learning]		X*	X
Extraction Attack (Model Stealing, Membership inference)			X
Model Injection [AI governance]		X*	X

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

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)

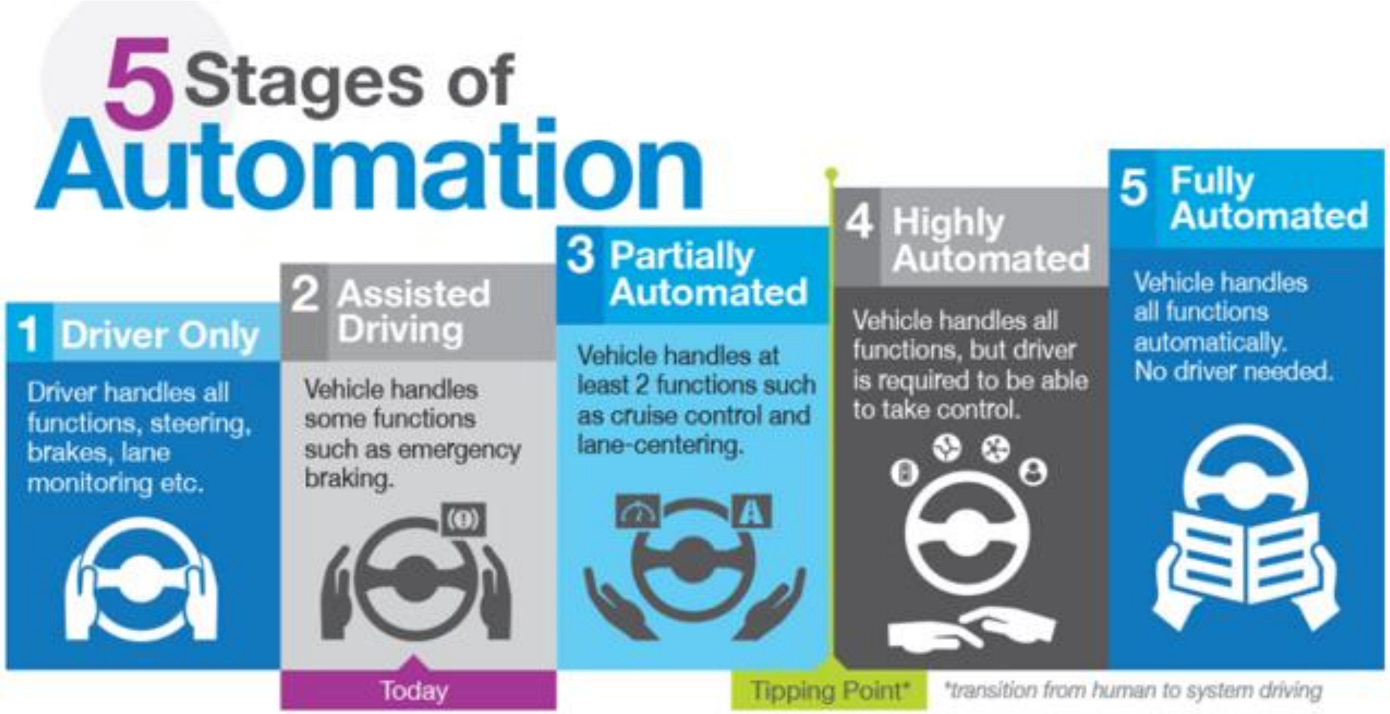
- This model is certified to be **attack-proof** up to a certain level
- Quantifiable metric for certified robustness
- AI standards, governance, and law regulation

Applications to AI (Model Boosting)

- Data augmentation
- **Model reprogramming**: data-efficient transfer learning
- Model watermarking

How to Define Levels of Robustness for AI?

- Lessons from autonomous driving systems



https://seekingalpha.com/article/4249591-tech-stocks-be-long-term-winners-in-autonomous-vehicles

https://blog.netapp.com/how-to-build-a-data-pipeline-for-autonomous-driving/

My View of AI Robustness Levels and Evaluations

Robustness Levels

Level 1 – Distribution Shifts

- Performance on **non-adaptive** (pre-generated) test sets
- Examples: Natural Corruption; Random Perturbation; Context Shifts

Level 2 – Single threat model

- Performance against optimized (worst-case) white-box adversarial examples based on one type of **domain-specific data modifications** generated from a test dataset
- Examples: Gradient-based attacks using Lp norms

Level 3 – Multiple threat models

- Performance against white-box adversarial examples generated by **a set of feasible threat models** from a test dataset
- Examples: Ensemble attacks using Lp norms and semantic perturbations

Level 4 – Global (Universal) Robustness

- Evaluation of **global robustness (input-agnostic)** instead of local robustness; Ultimate generalization (AGI); Fast adaptation
- Examples: Unrestricted adversarial examples

Robustness Evaluations

1st Party (model developer)

- Adaptive white-box attack
- Full system transparency

2nd Party (model inspector)

- Non-adaptive white-box/gray-box attack
- Information obfuscation; Unknown implementation

3rd Party (end user)

- Soft-label/hard-label/no-box black-box attack
- Target model is a black-box function with limited information feedback

Making AI model Robust is truly 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 →

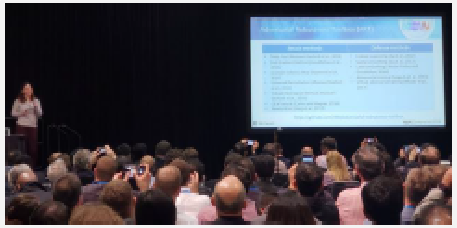
Load classifier model (Keras, TF, PyTorch etc) →

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))
```

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

siliconANGLE
Attackers can fool AI programs. Here's how developers can fight back

ZDNet
IBM launches open-source library for securing AI systems
The framework-agnostic software library contains attacks, defenses, and benchmarks for securing artificial intelligence systems.

IBM ENTWICKELT WERKZEUGE GEGEN HACKERANGRIFFE DURCH "BÖSE" KI

IBM、AIシステムを保護するオープンソースライブラリ「Adversarial Robustness Toolbox」

Выпущена Adversarial Robustness Toolbox, открытая библиотека от IBM для защиты ИИ

Adversarial Robustness Toolbox : IBM propose une boite à outils open source pour sécuriser l'intelligence artificielle

23-04-2018 | door: Witold Kepinski
IBM Adversarial Robustness Toolbox beschermt tegen kwaadaardige AI

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