Towards robust and trustworthy AI

Mathieu Sinn, PhD

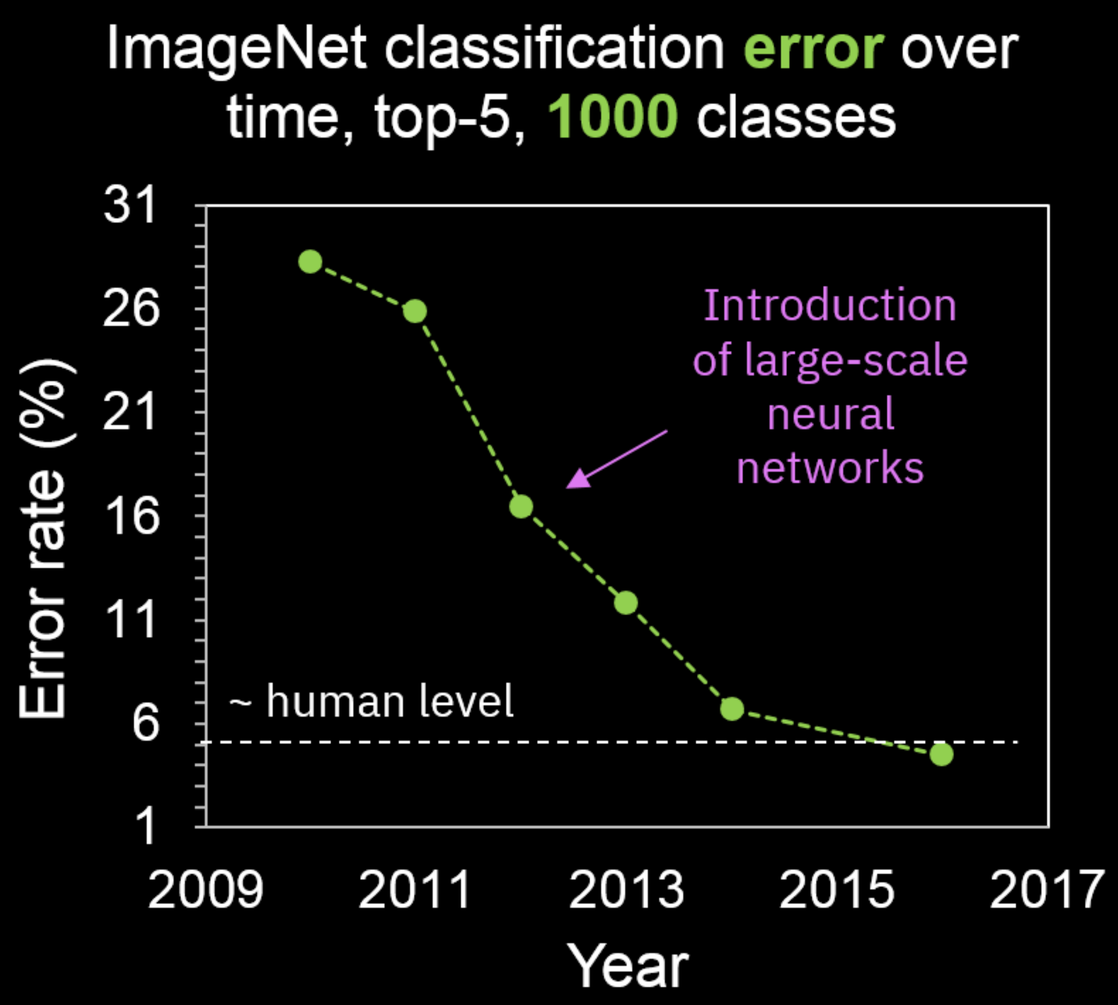
Manager AI, Security & Privacy
IBM Research Europe
Dublin, Ireland

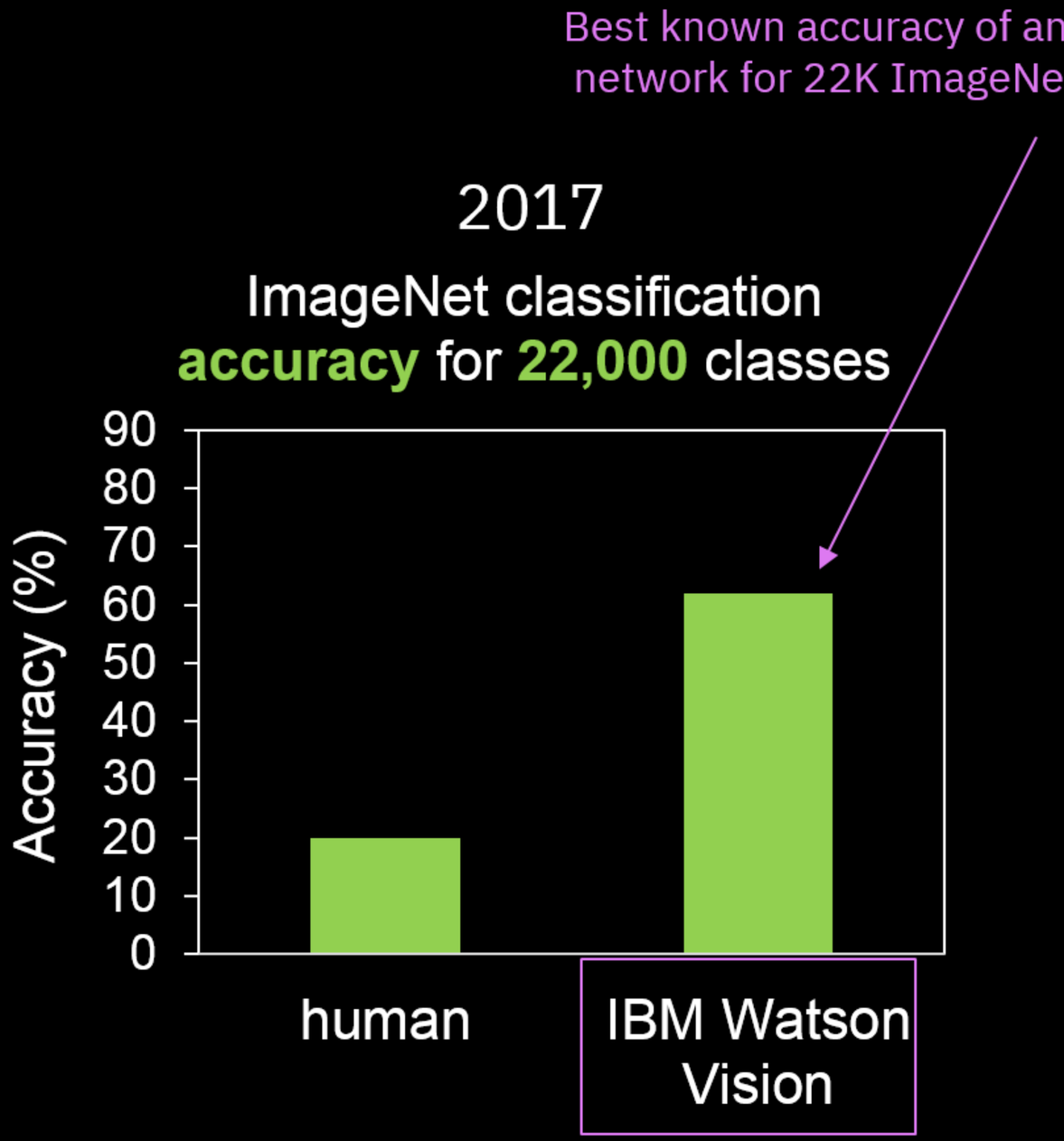
CeADAR
Online Tech Talk
September 3, 2020

Outline

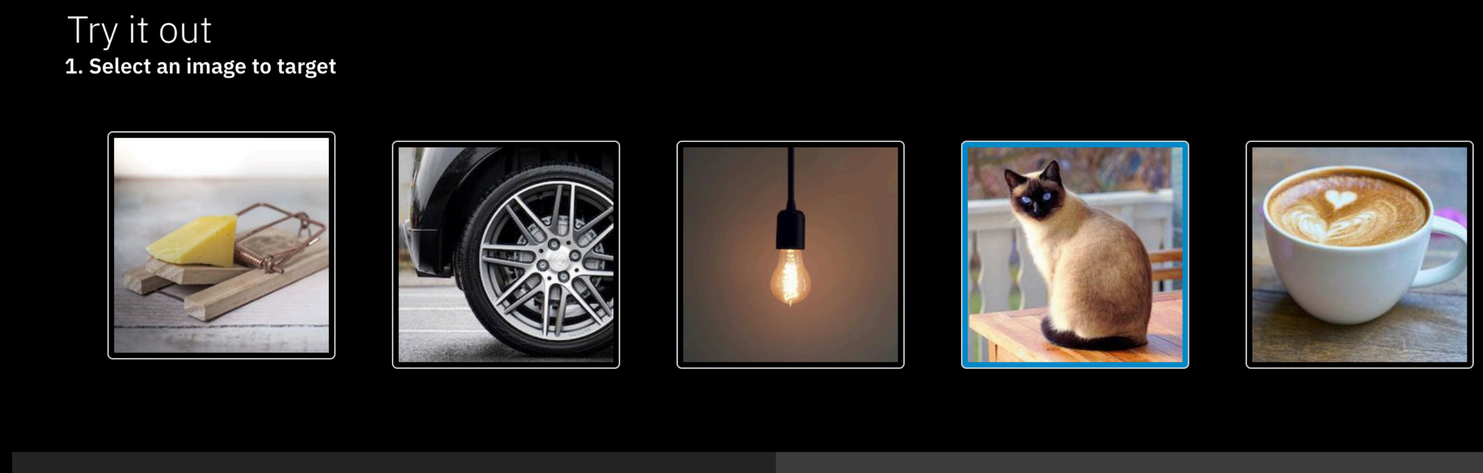
- Adversarial threats to AI
- The Adversarial Robustness Toolbox
- Resources & Conclusions

“AI is the new electricity”





But... AI is also surprisingly brittle!



https://art-demo.mybluemix.net/

This does not only apply to images…

Original

“Basketball throw” (72.5%)

Adversarial

“Tennis swing” (49.5%)

https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/adversarial_action_recognition.ipynb

What is happening??

Elephant (81.4%)

Cat (71.1%)

Elephant (63.9%)

[Geirhos et al., 2019]

ML models LOVE short cuts!

What is happening??

1.	**frying pan**	**76.5%**
2.	wok	15.8%
3.	stove	5.4%

1.	**wok**	**63.2%**
2.	frying pan	35.1%
3.	spatula	0.6%

[Carter et al., 2019]

ML models LOVE short cuts!

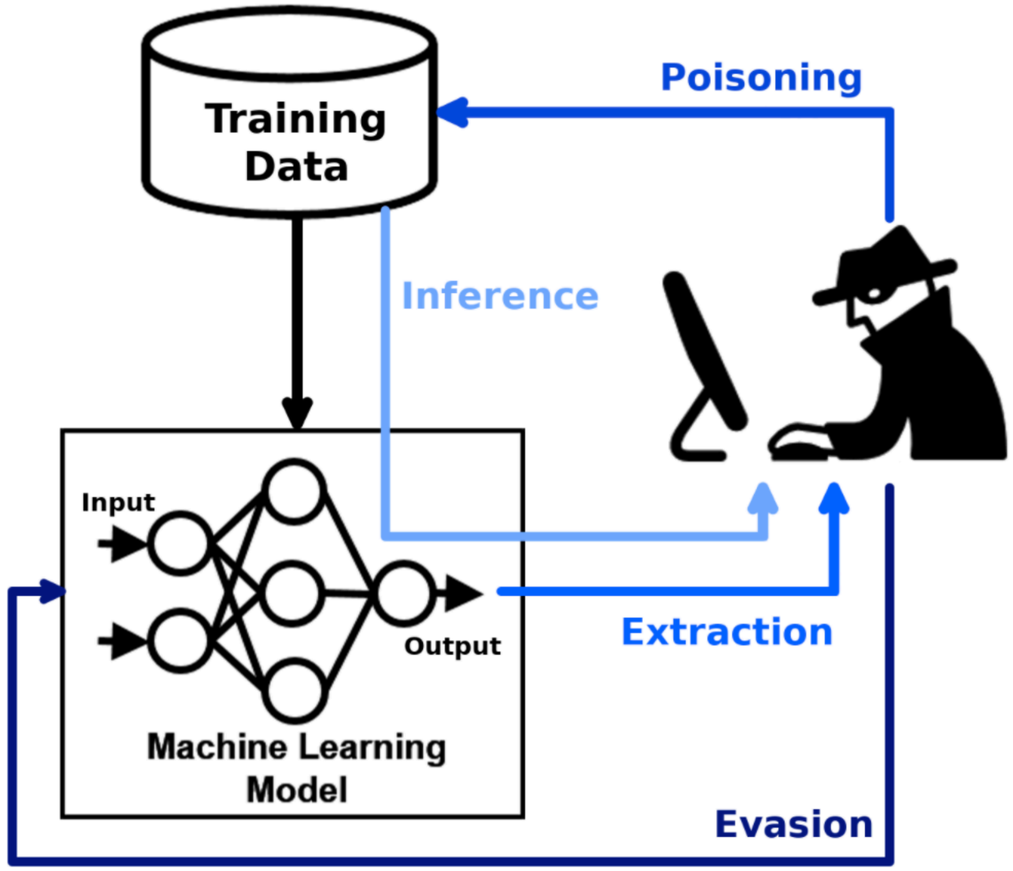
What is happening??

Schoolbus

[Nguyen et al., 2014]

ML models LOVE short cuts!

Adversarial Threats to AI



Adversarial Threats to AI
Scenarios

Security:

- **Disappearance** attacks against **CCTV surveillance**

Autonomous vehicles:

- Targeted/untargeted attacks on **object recognition** and **image segmentation** models

Cybersecurity:

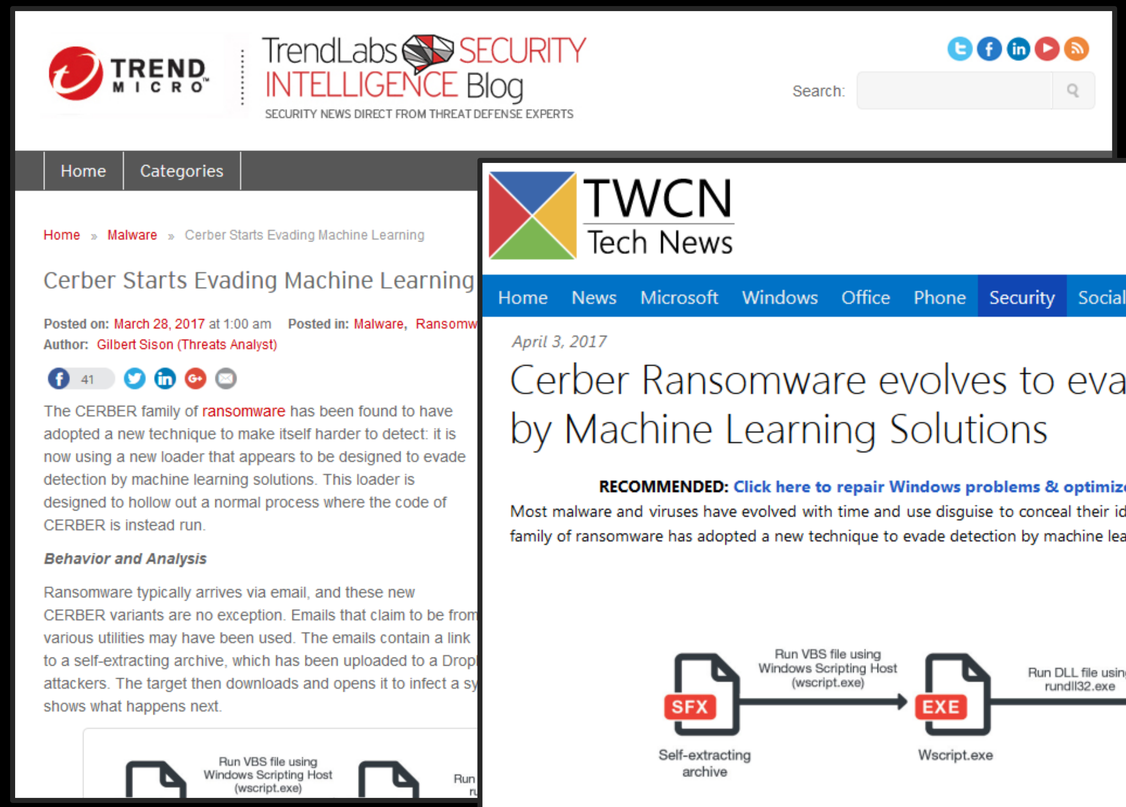
- Evade **spam filters**, **malware detectors**, **network intrusion detection** etc.

Financial services:

- Evade **fraud detection**

Undermine trust in AI

Such attacks are already happening...



Reports of cybersecurity vulnerabilities due to evasion attacks against AI in anti-malware / -virus products.

Adversarial Robustness Toolbox (ART)

Repo: https://github.com/Trusted-AI/adversarial-robustness-toolbox
Docs: https://adversarial-robustness-toolbox.readthedocs.io/
Demo: https://art-demo.mybluemix.net

- Python library, 12K 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 →

```python
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))
```

TensorFlow · mxnet · dmlc XGBoost · Keras · PYTORCH · scikit learn · LightGBM

Open-source release @ RSA 2018:

Current stats:

- **1.6K** GitHub stars
- **450+** forks
- **250+** clones/w
- **1K+** downloads/w

siliconANGLE

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

BY JAMES KOBIELUS
UPDATED 00:53 EST . 22 APRIL 2018

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.

By Charlie Osborne

IBM ENTWICKELT WERKZEUGE GEGEN HACKERANGRIFFE DURCH "BÖSE" KI

20. April 2018

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

18 апреля 2018 в 0:24, Новости · 2 минуты · 277

ZDNet Japan > セキュリティ

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」

Charlie Osborne （Special to ZDNet.com） 翻訳校正： 編集部 矢倉美登里 吉武稔夫 （ガリレオ） 2018年04月19日 13時42分

いいね！ 8 · ツイート · G+ · B! 3 · Pocket 20

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

Par: fredericmazue | jeu, 19/04/2018 - 12:29

intelligence artificielle, attaque contradictoire

J'aime 1,7 K · Partager 4 · Tweeter · G+

23-04-2018 | door: **Witold Kepinski**

IBM Adversarial Robustness Toolbox beschermt tegen kwaadaardige AI

Resources & Conclusions

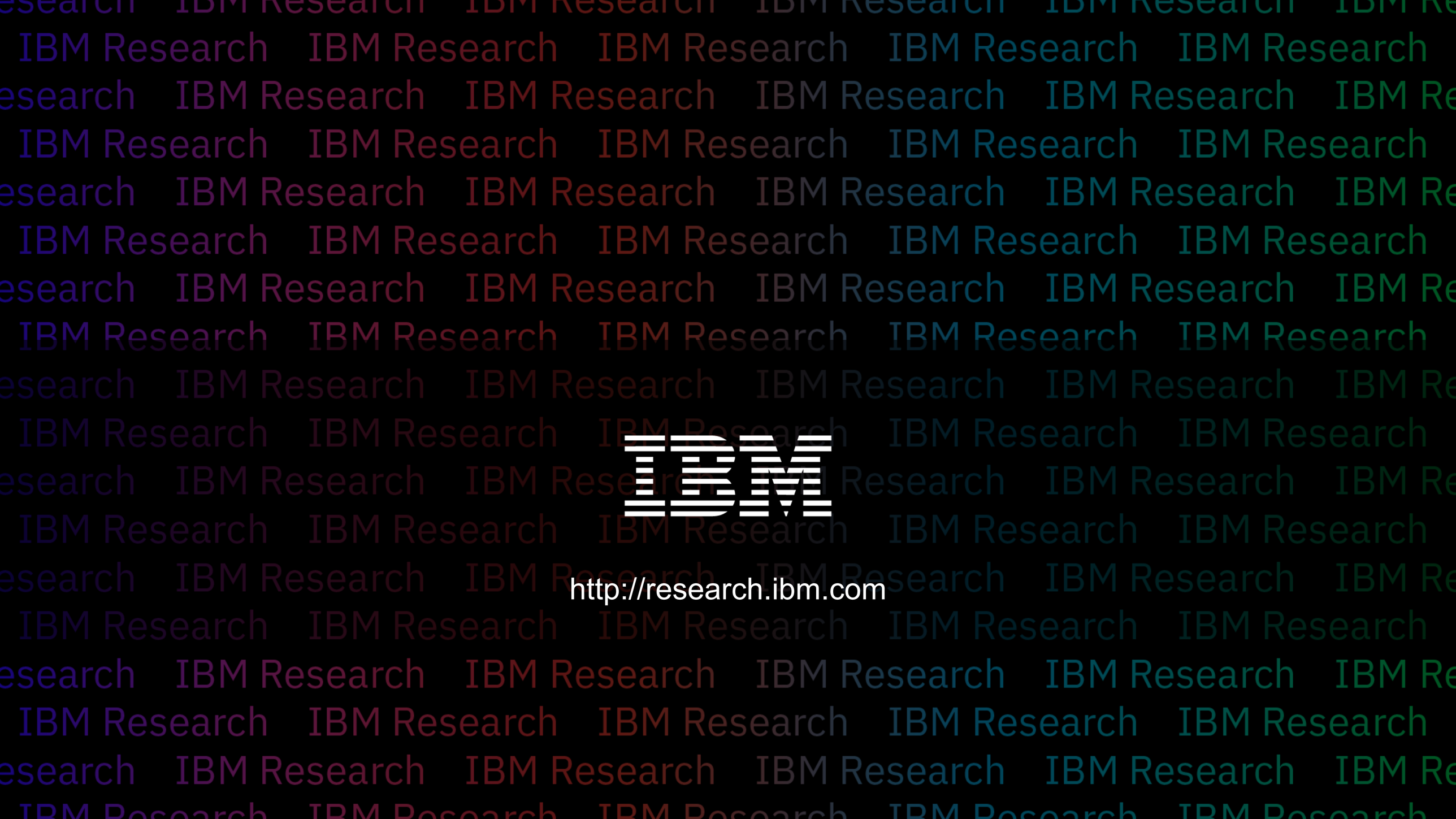
- GitHub
  - https://github.com/Trusted-AI/adversarial-robustness-toolbox
- Documentation
  - https://adversarial-robustness-toolbox.readthedocs.io
- Slack
  - https://ibm-art.slack.com
- Demo
  - https://art-demo.mybluemix.net/
- Blog
  - https://www.ibm.com/blogs/research/2019/09/adversarial-robustness-360-toolbox-v1-0
- White paper
  - https://arxiv.org/abs/1807.01069
- Tutorial
  - http://www.research.ibm.com/labs/ireland/nemesis2018/pdf/tutorial.pdf

https://www.research.ibm.com/artificial-intelligence/trusted-ai/

ART – How to contribute?

- Check Contributions page:

  https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/CONTRIBUTING.md
- Create github issues for suspected bugs, missing features, ideas for improvements etc.
- Contribute bug fixes, new features etc. via pull requests to dev branch
- Follow PEP 8 coding style, provide unit tests
- Sign DCO (via ‘-s’ flag) for every commit


IBM Research
http://research.ibm.com