AI/ML Security

CyberWarriors Summer Camp 2024

Amanda Fernandez, PhD
Assistant Professor - UTSA

UTSA VAIL

Vision & Artificial Intelligence Lab

North Paseo Building 2.210 & 2.214

CONNECT
UTSA • STMU • UNLV
CONsortium on Nuclear sECurity Technologies
In Colaboration w/
Los Alamos National Laboratory
Argonne National Laboratory

matrix
The UTSA AI Consortium for Human Well-Being

Amanda Fernandez, Ph.D.
Assistant Professor
Department of Computer Science
https://bit.ly/UTSA-VAIL

Students:

- Michael Nootbaar (PhD)
- Logan Robinson (PhD)
- Daniel Mohanadhas (PhD)
- John Weaver (PhD)
- Niklesh Akula (MD/MS)
- John Wilburn (BS)
- Daniel Castillo (BS)
- Pedro Davila (BS)
- Diego Enriquez (BS)
- Diego Garcia (BS)
- Sergio Contreras (BS)

Research

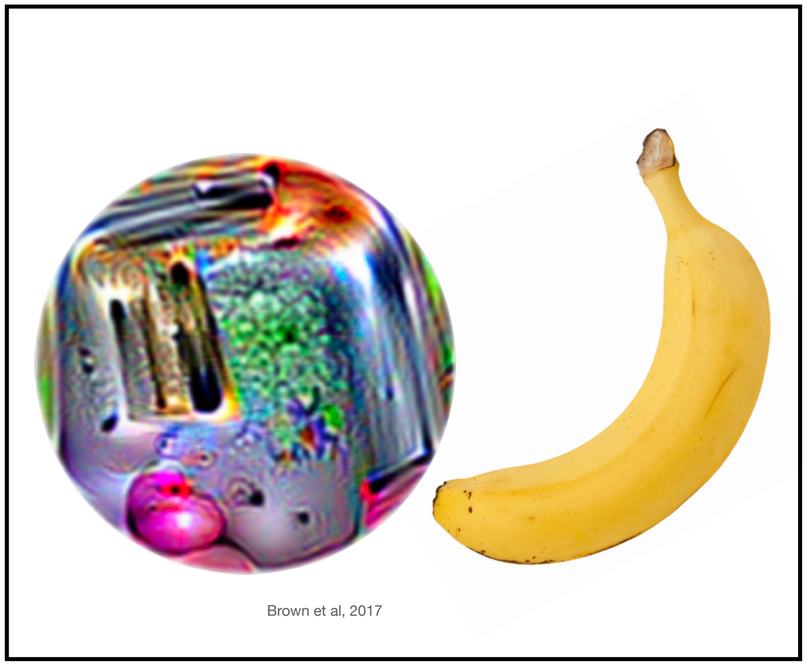
- **Explainable & efficient deep learning**
- **Computer vision**
- **Applications in the physical sciences**

Funding

- **NSF**
- **DoD**
- **DOE**

Robust Visual Understanding

How can we deepfake entire movies, but not recognize bananas?

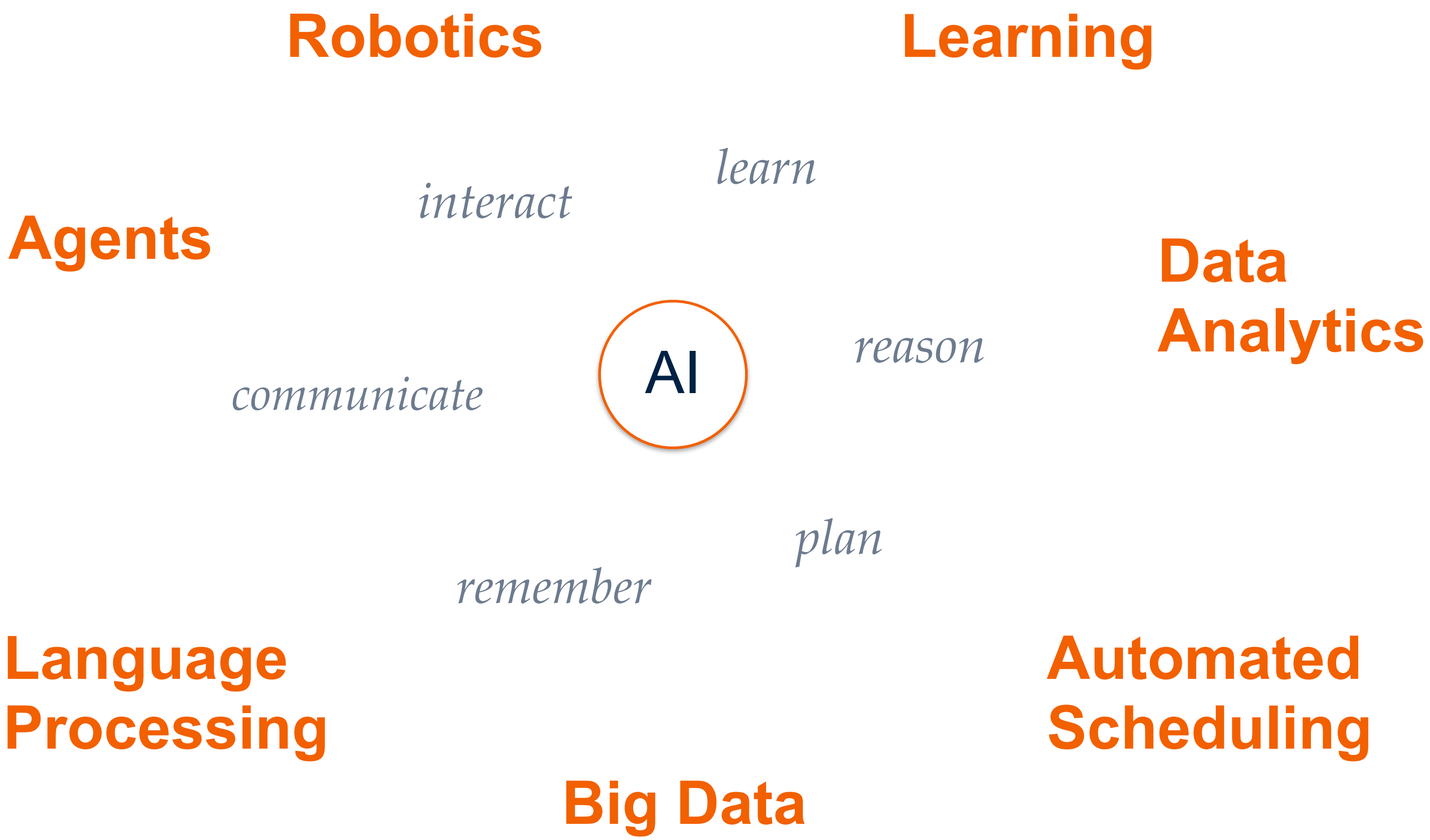


AI/ML Security

Outline

- Artificial Intelligence & Machine Learning
- Attacking ML Models
- Generative Networks
- Explainable A.I.

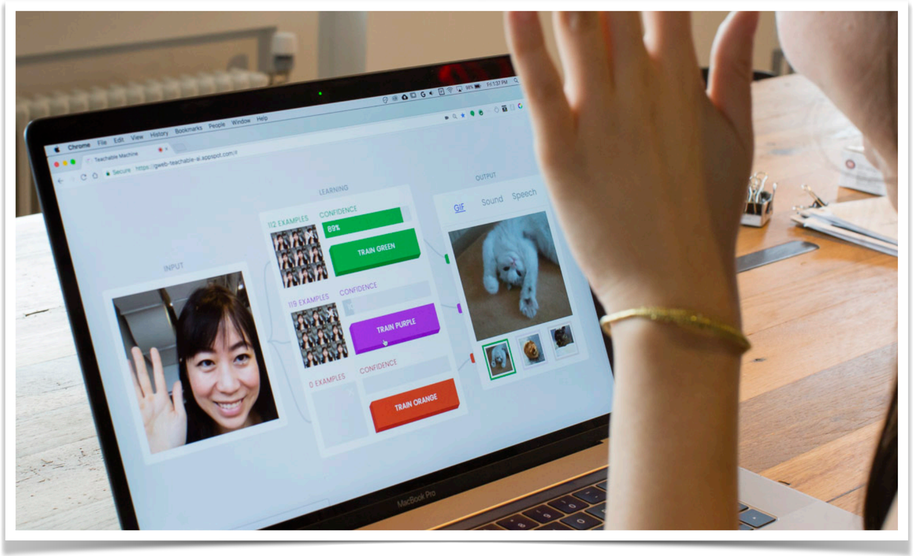
Artificial Intelligence



Interested in AI/ML?

Start here:

- Teachable Machine
- YouTube, podcasts
- Course websites, eDx, Coursera, Udemy, …
- Futurism, MIT News
- *Start slow! Learn Python, review linear algebra*



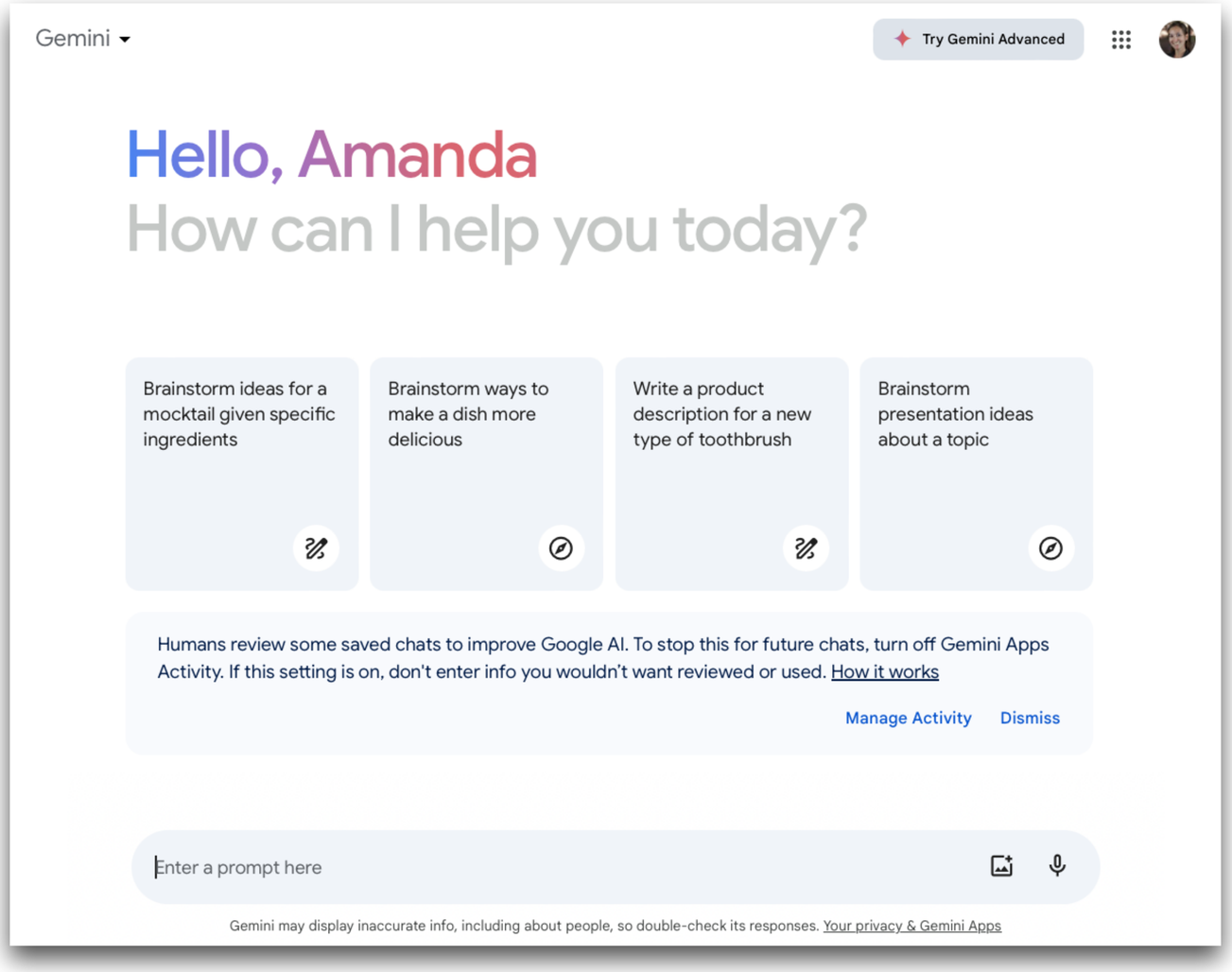
https://teachablemachine.withgoogle.com

Expert Systems

Large Language Models (LLMs)

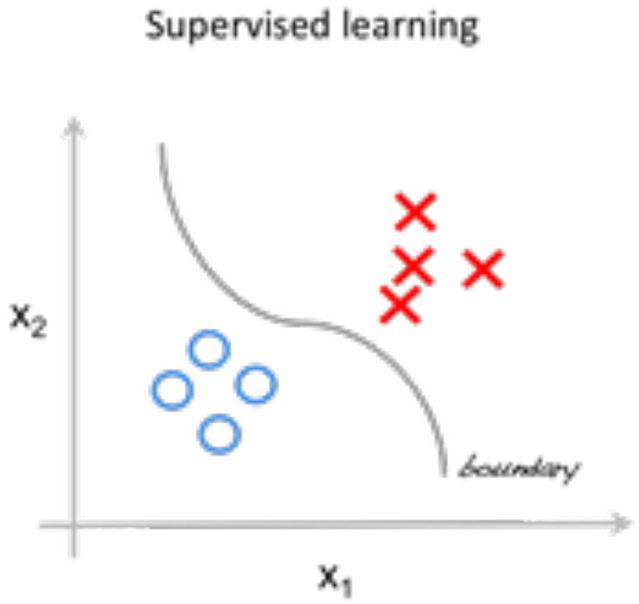
- *Generative* AI models

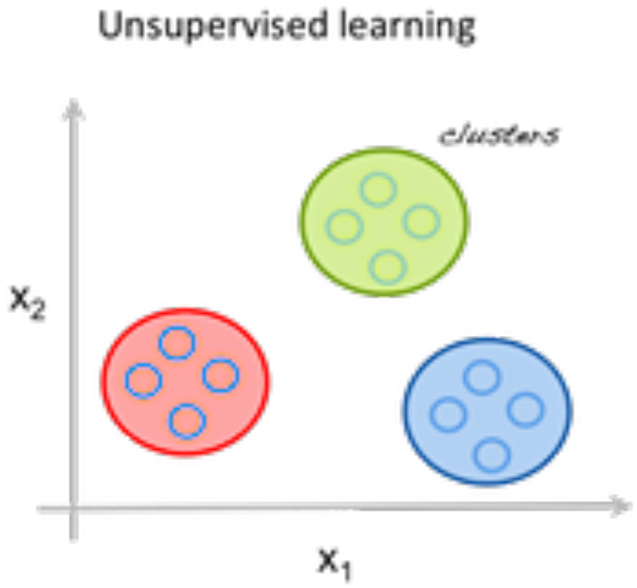
- Examples:
  - Gemini —>
  - GPT
  - Llama
  - DALL-E



Machine Learning

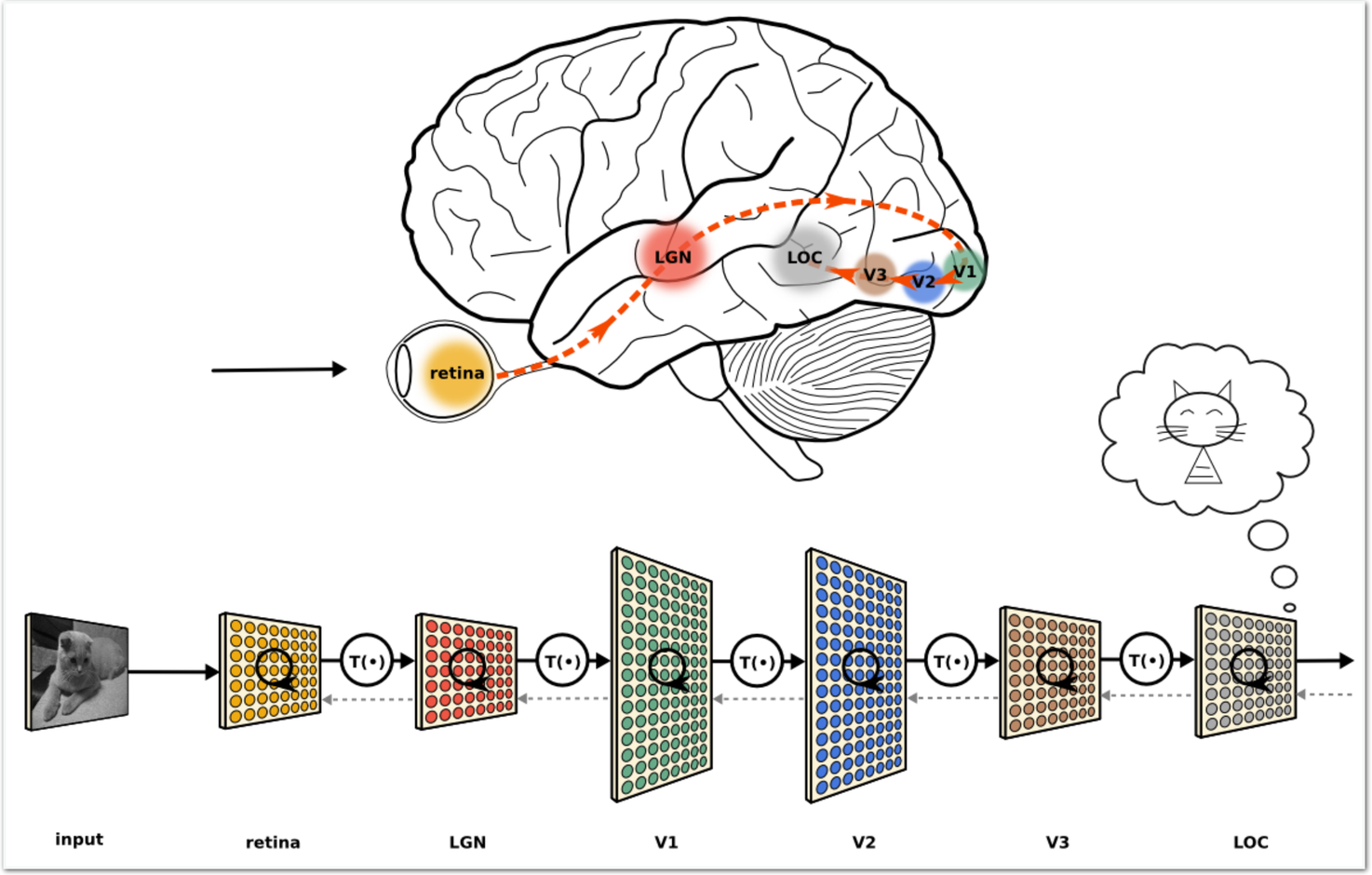
Types: Supervised & Unsupervised





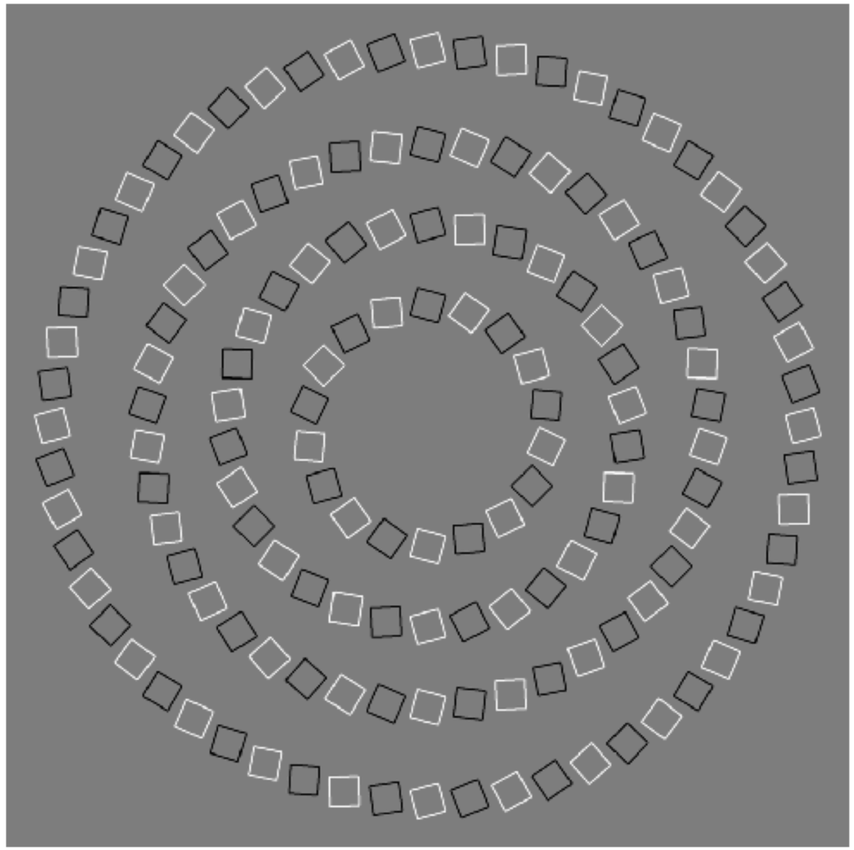
Machine Learning

Building models



Machine Learning

..easy, right?



(Pinna and Gregory, 2002)

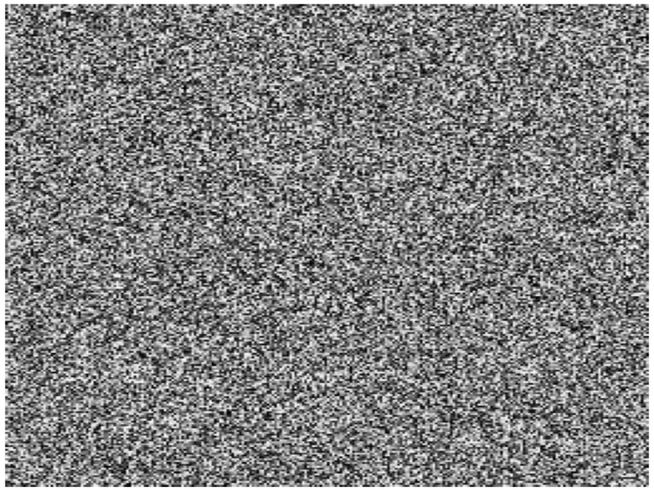
Adversarial Attacks

Examples & patches

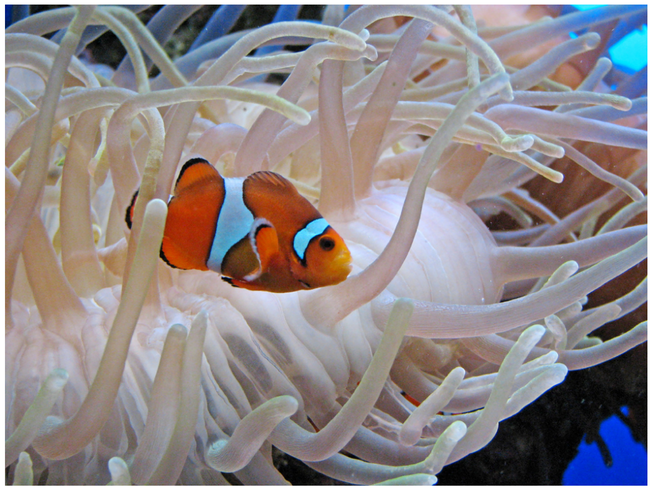
- Attacking a *neural network* involves providing **data** that affects its performance.
- *Goal*: data should look innocuous!

fish (92%)

+ 0.07 *



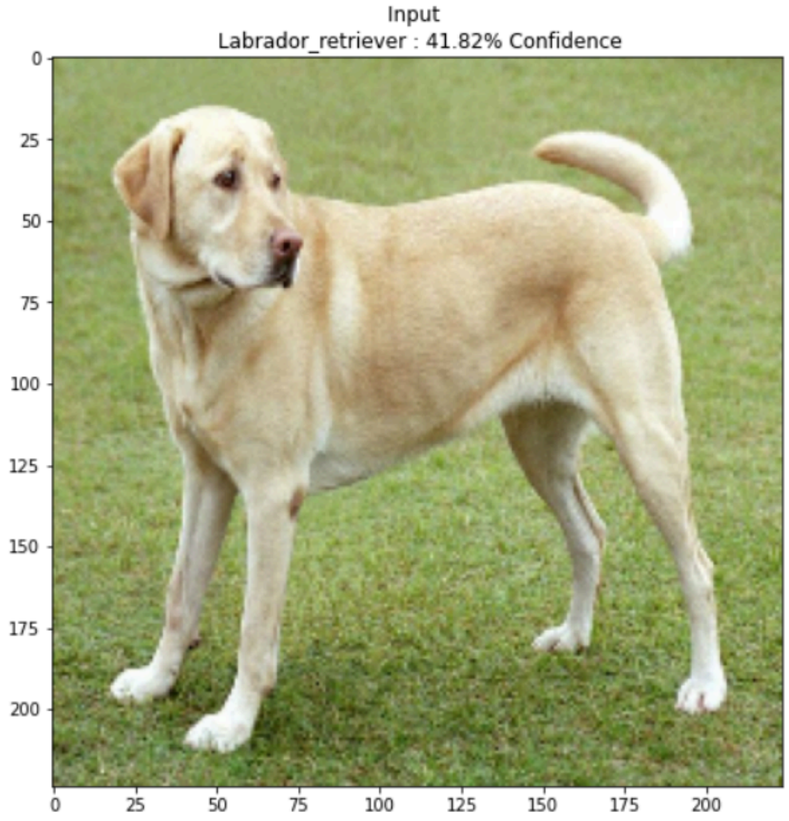
=

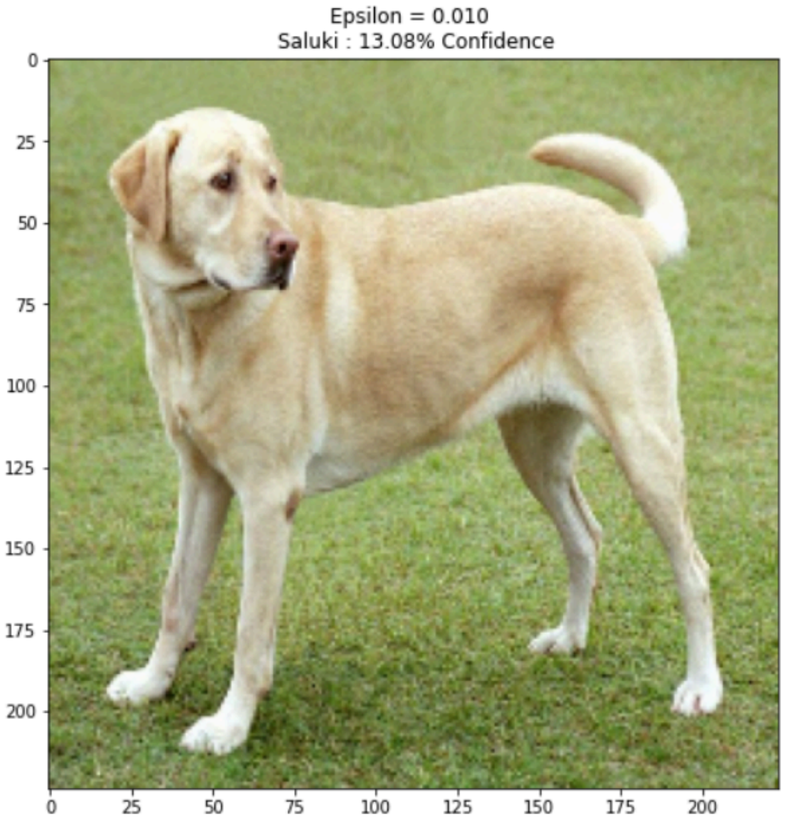


boat (79%)

Adversarial Attacks

Examples





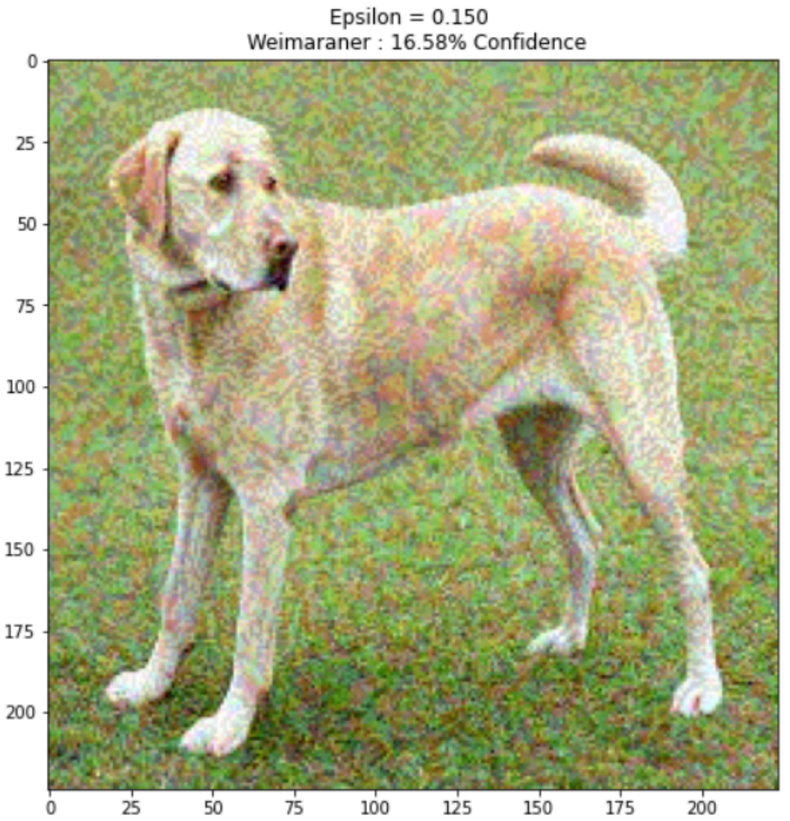


Image source:TensorFlow.org

Adversarial Attacks

Patches in the real world

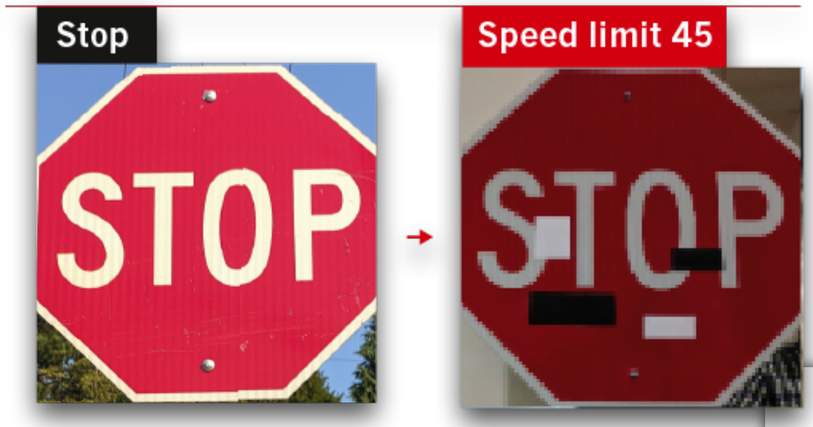


Image source: nature.com

Fig. 1: A silhouette of the eyeglasses we use.

Fig. 2: Examples of raw images of eyeglasses that we collected (left) and their synthesis results (right).

Sharif et al, 2017

Tesla's autopilot tricked into driving on the wrong side of the road

TECHNOLOGY 1 April 2019

Tesla's autopilot can takeover some driving tasks

David Paul Morris/Bloomberg via Getty Images

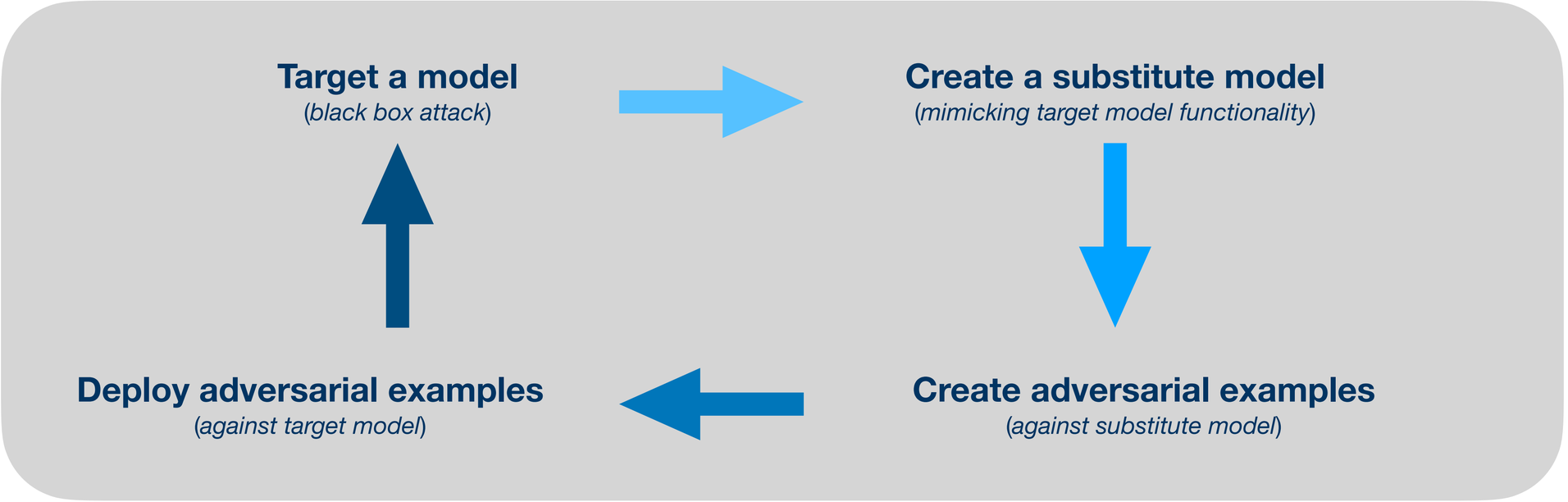
NewScientist

By **Chris Stokel-Walker**

Keen Security Labs, of Chinese tech company Tencent, confused a Tesla Model S by placing 3 stickers on the road.

How does a model know what's real?

To attack a model, typically you start with a target..



Machine Learning

Generative Models

Generating candy hearts

…or “motivational” posters!

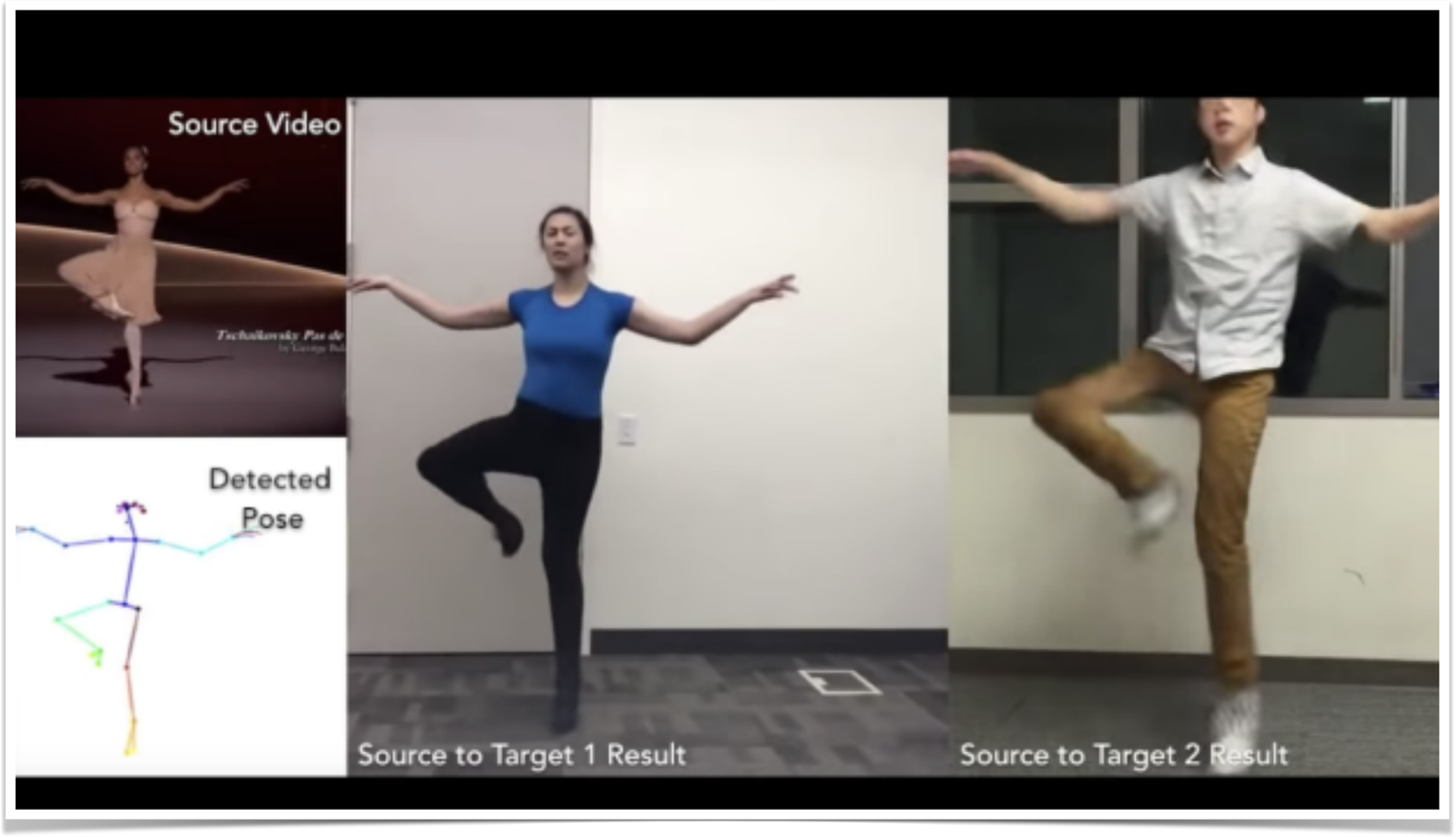
Deep fakes

Generating realistic digital media.

Brundage et al., 2018

https://www.youtube.com/playlist?list=PLpaGT3slbH0AGdScmPBAqKZ3SZEeOGnpd

Deep fakes

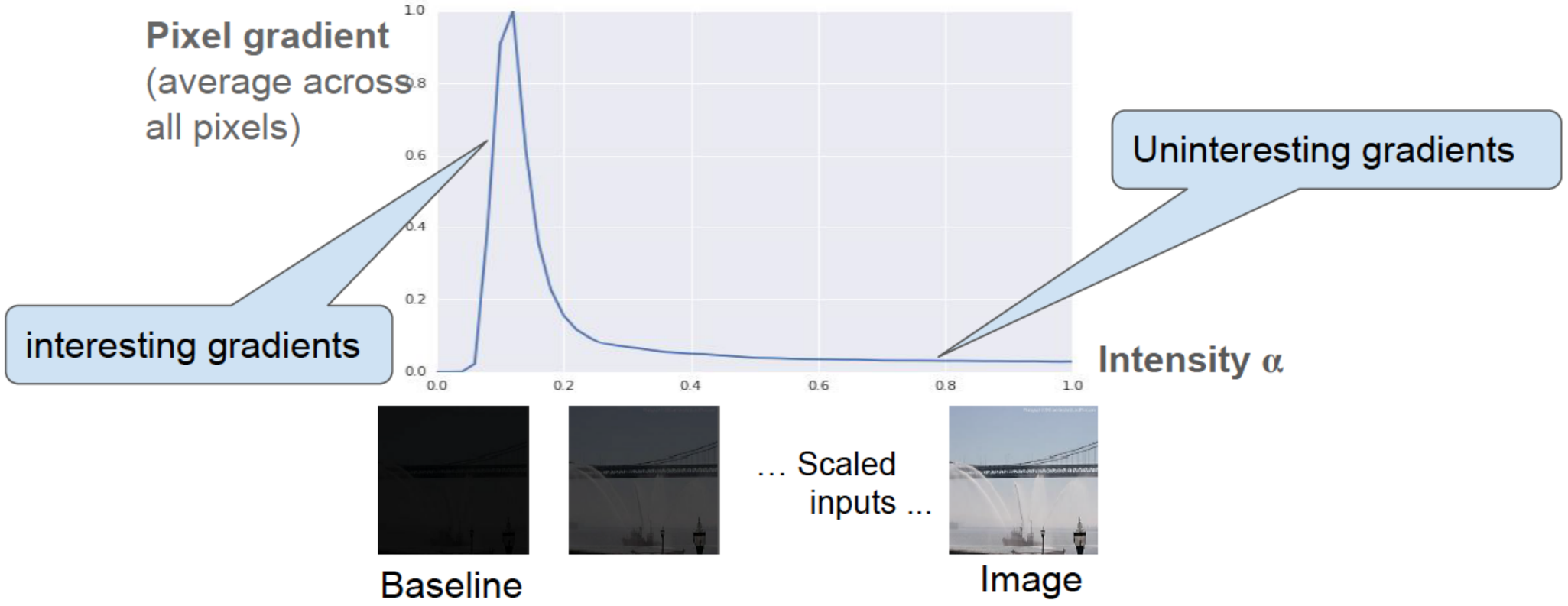


“Everybody Dance Now”

https://www.youtube.com/watch?v=PCBTZh41Ris

Explaining Deep Learning Models

Integrated Gradients for Attribution



Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "**Axiomatic attribution for deep networks**." *International conference on machine learning*. PMLR, 2017.

UTSA VAIL

Vision & Artificial Intelligence Lab

North Paseo Building 2.210 & 2.214

Deep Learning

- Explainability
- Optimization
- Foundation models

Computer Vision

- Segmentation
- Adversarial corruption
- Multimodality

Applications

- Nuclear materials
- Fabrication analysis
- Medical imaging

CS Education

- Broadening participation
- Undergraduate research
- AI tools

Q&A

https://bit.ly/UTSA-VAIL