Computer Vision Applications

COMP 388-002/488-002 Computer Science Topics

Daniel Moreira
Fall 2022

LOYOLA UNIVERSITY CHICAGO

RECAP

Face Recognition

COMP 388-002/488-002 Computer Science Topics
Computer Vision Applications

Daniel Moreira
Fall 2022

LOYOLA UNIVERSITY CHICAGO

Practical Activity 1

RECAP

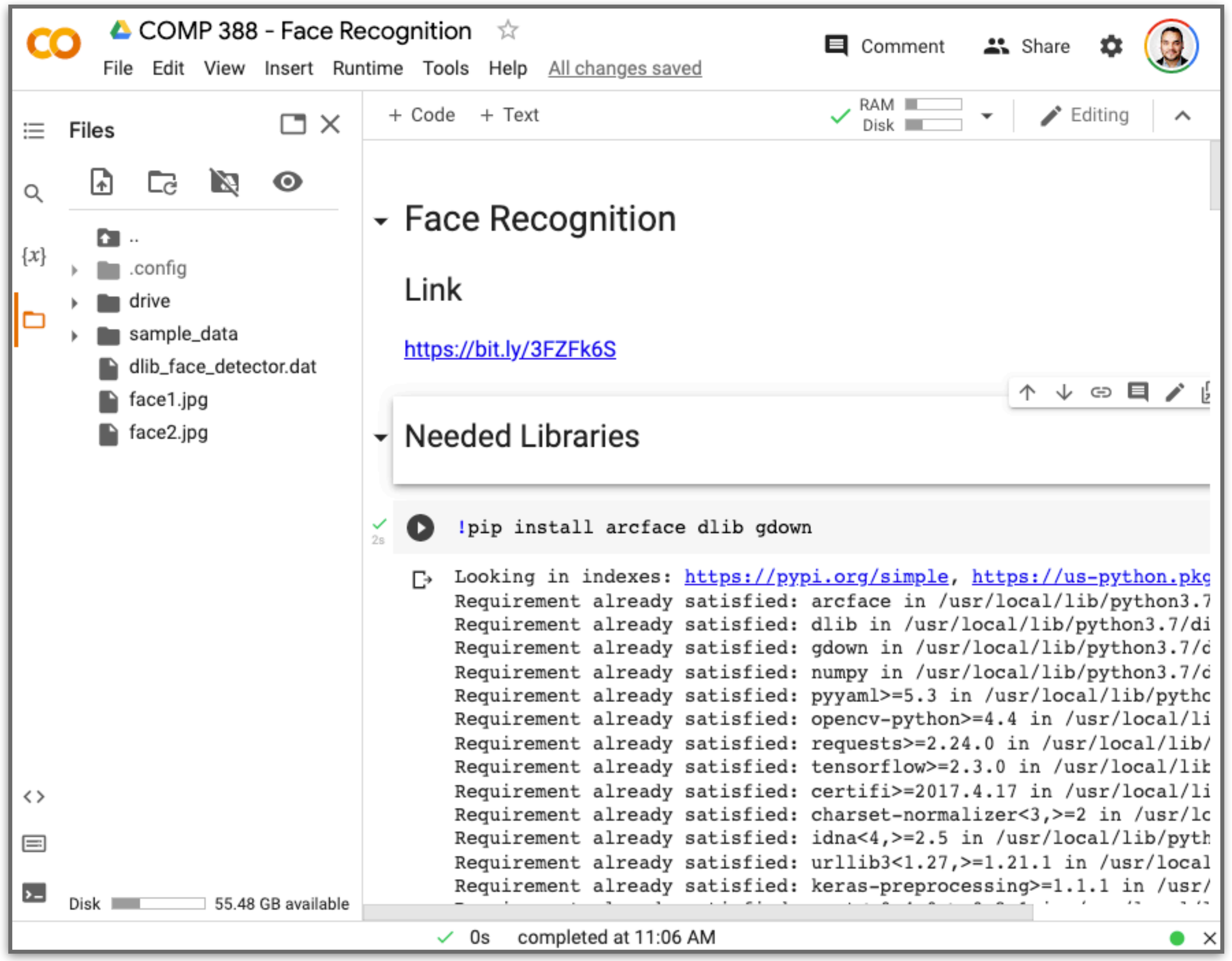
Work in Pairs

Use Google Colab at https://bit.ly/3FZFk6S

Observe ArcFace's[1] distance

- Different faces
- Different poses
- Different "accessories" (e.g., glasses, mask, etc.)

Contributions or Question?



1. https://arxiv.org/abs/1801.07698

LOYOLA UNIVERSITY CHICAGO

How about Attacks?

RECAP

https://www.wired.com/story/10-year-old-face-id-unlocks-mothers-iphone-x/


Support The Guardian
Available for everyone, funded by readers
Contribute → Subscribe →
Sign in
The Guardian
News Opinion Sport Culture Lifestyle
World ▸ Europe US Americas Asia Australia Middle East Africa Inequality More
Surveillance
This article is more than 6 months old
The fashion line designed to trick surveillance cameras

https://www.theguardian.com/world/2019/aug/13/the-fashion-line-designed-to-trick-surveillance-cameras

LOYOLA UNIVERSITY CHICAGO

Attacks

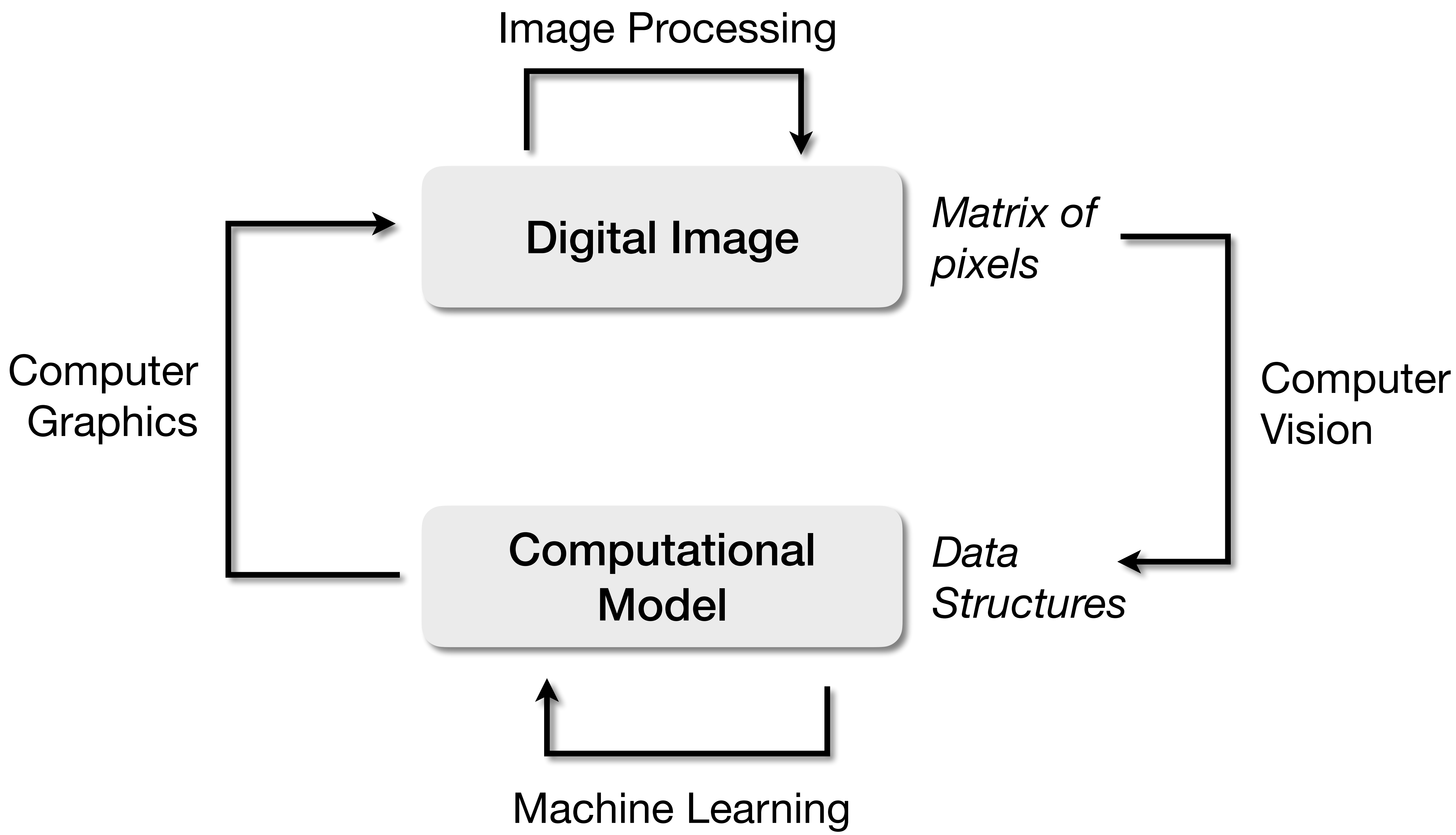
COMP 388-002/488-002 Computer Science Topics
Computer Vision Applications

Daniel Moreira

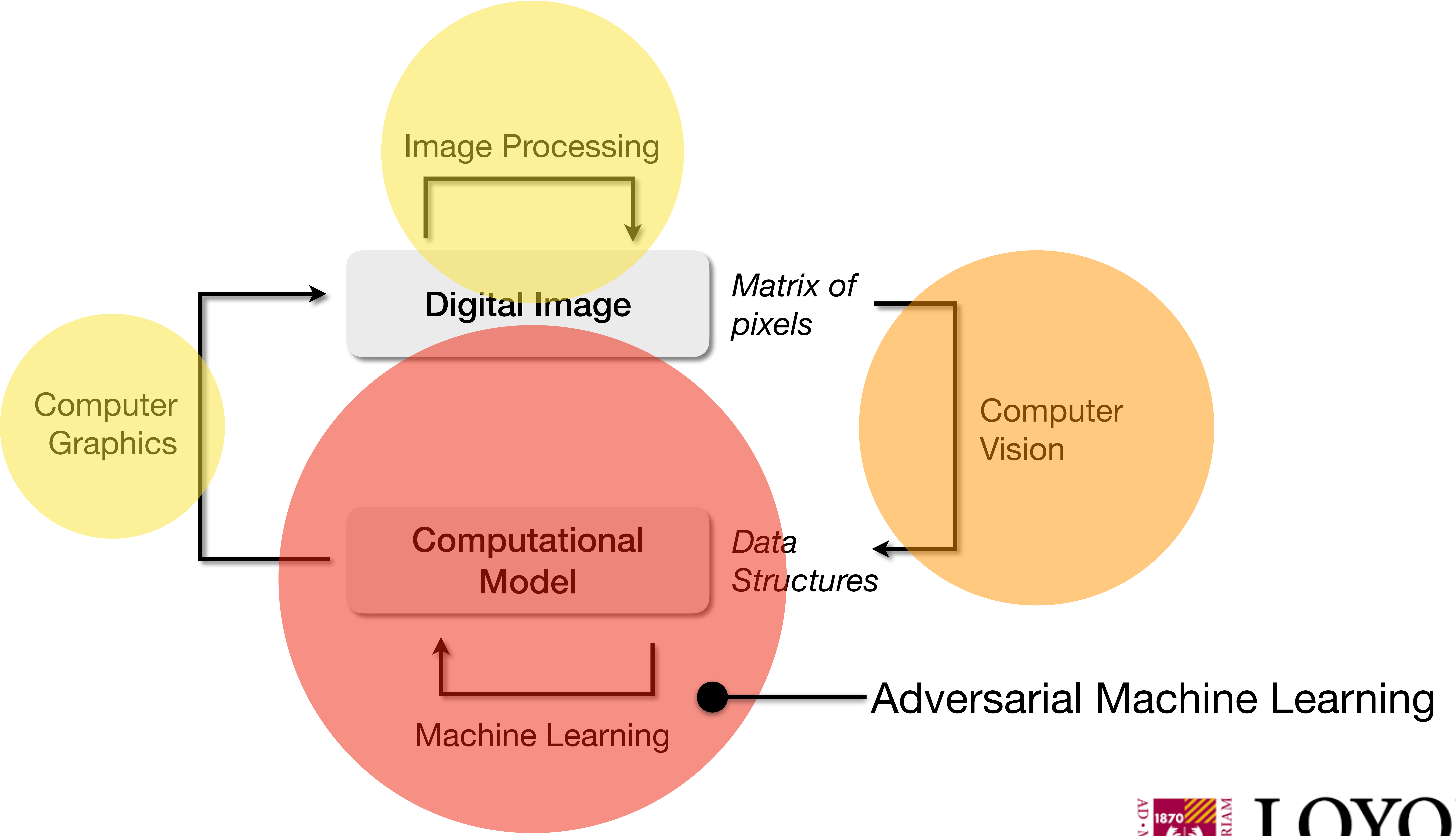
Fall 2022

LOYOLA
UNIVERSITY CHICAGO

Computer Vision Attacks



Computer Vision Attacks

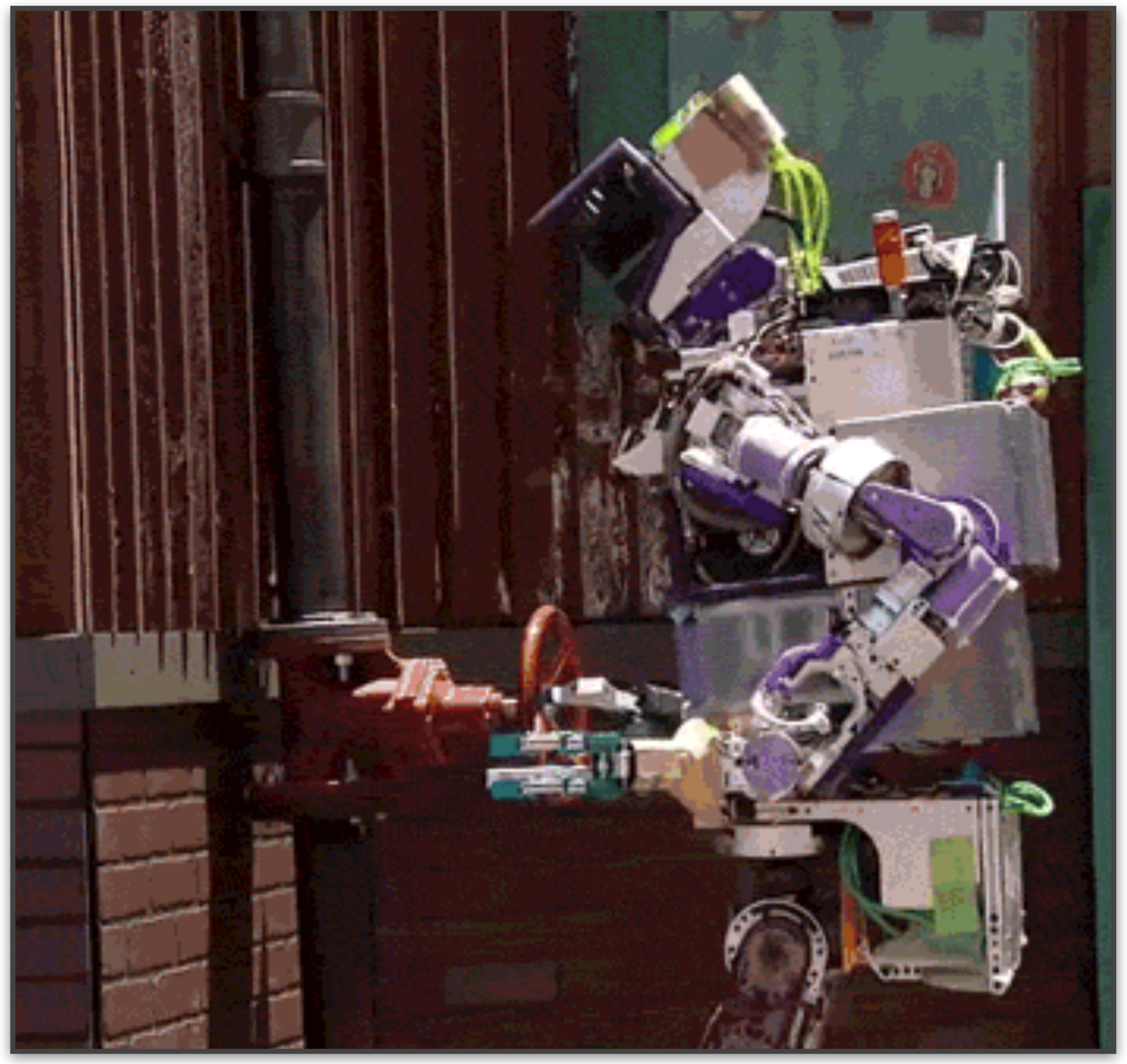


Adversarial Machine Learning

Attacks and Defenses

Attacks
Procedures performed by an **attacker** to explore the vulnerability of a computer system to make it behave out of specification.

Defenses
Procedures to avoid, detect, or mitigate attacks.

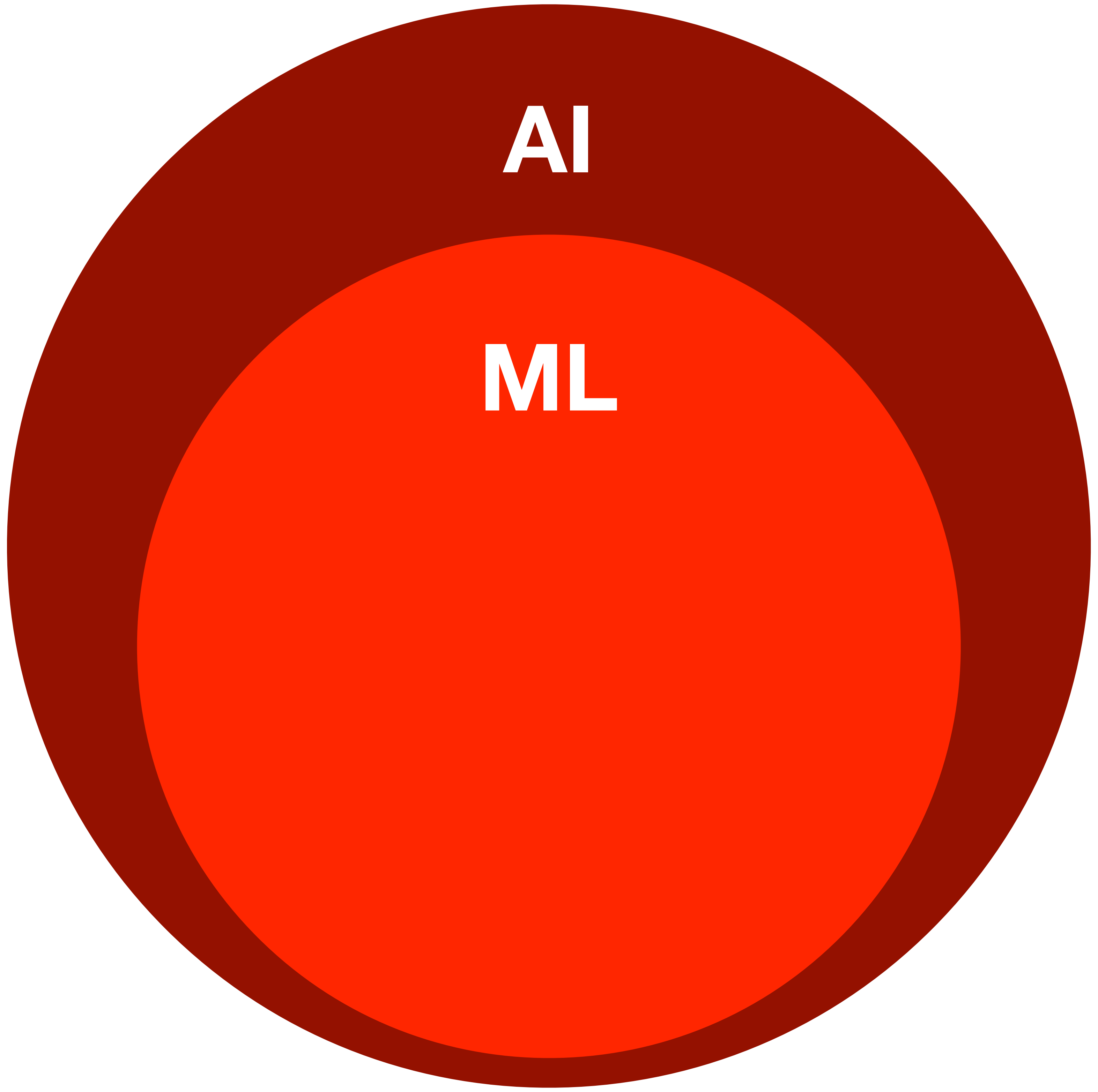


Adversarial Machine Learning

Within Machine Learning
Data driven.

Adversarial Data
Synthetic samples pretending to be real.
Manipulated samples pretending to be pristine.
Data copies pretending to be original (spoofing).

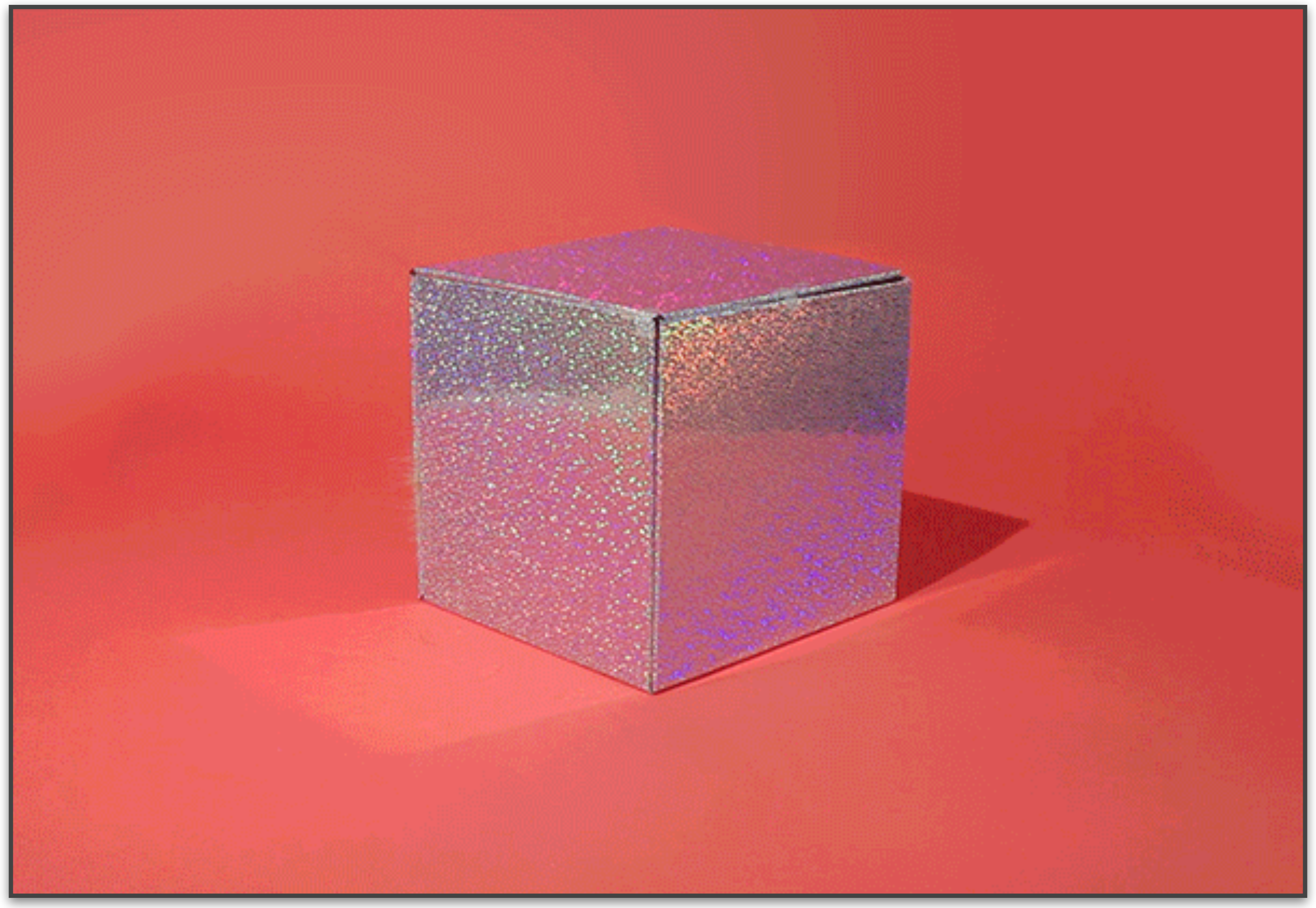
Two Attack Opportunities
At training time and at inference time.



Adversarial Machine Learning

Within Machine Learning
Data driven.

Types of Attack
White Box
The attacker knows how the solution works and has access to its trained model.



Black Box
The attacker only sees the input and output of the solution; sometimes, even these are not clear.

Gray Box
The attacker has limited knowledge about the solution.

Adversarial Machine Learning

Within Machine Learning
Data driven.

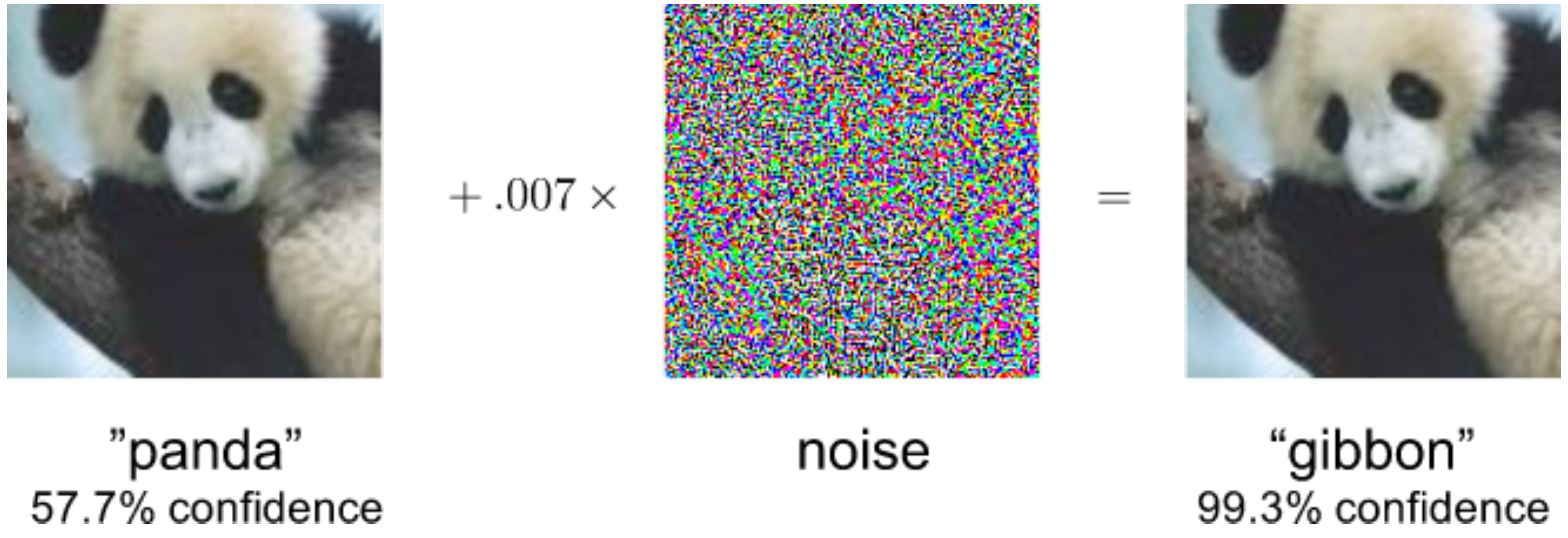
Attack Examples
Evasion
The attacker avoids data detection
(or causes misclassification) by the system,
usually by presenting adversarial data
(at inference time).

Example: evasion of text-based spam detection with images.

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Evasion
The attacker avoids data detection
(or causes misclassification) by the system,
usually by presenting adversarial data
(at inference time).

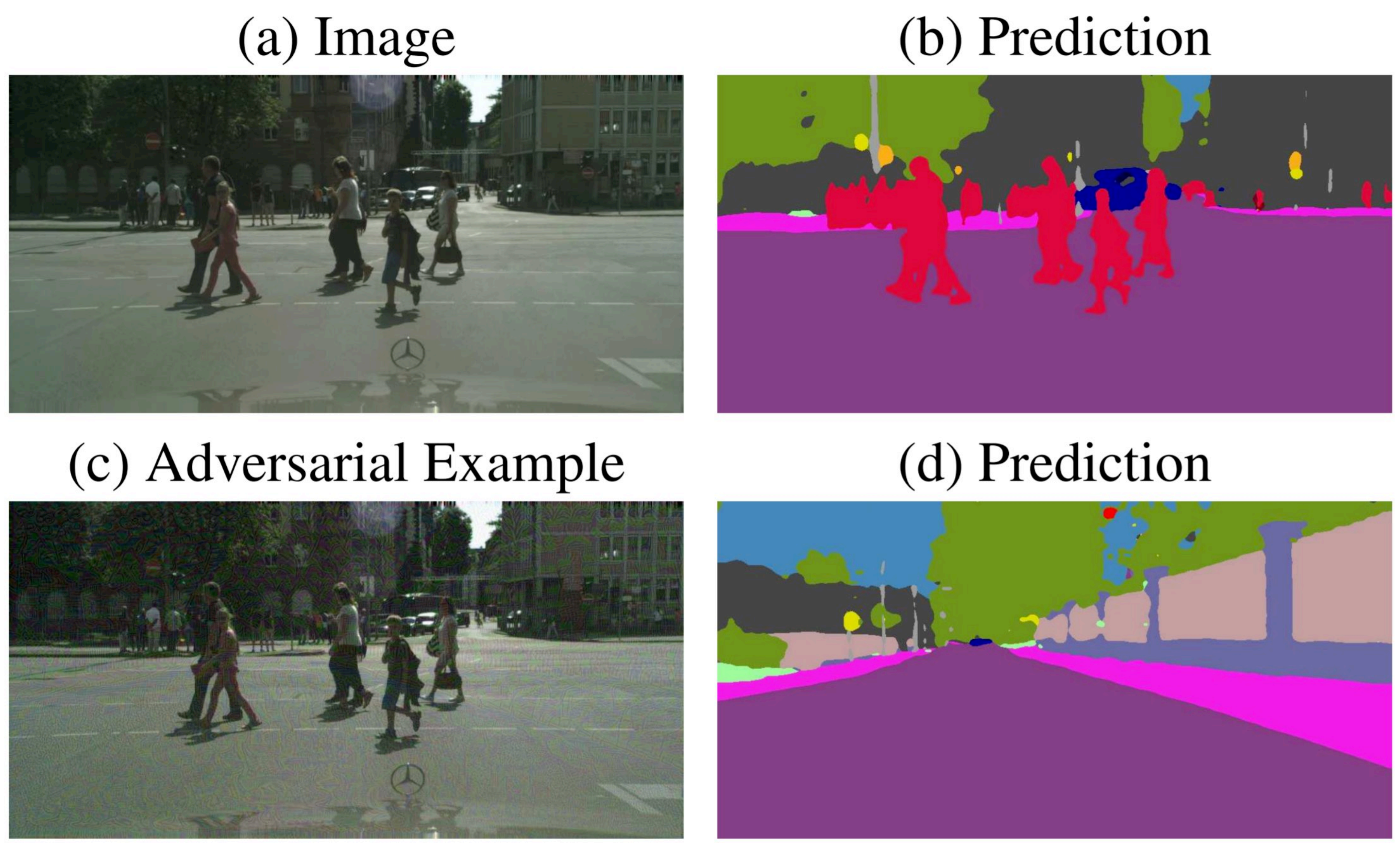


https://arxiv.org/pdf/1412.6572.pdf

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Evasion
The attacker avoids data detection
(or causes misclassification) by the system,
usually by presenting adversarial data
(at inference time).



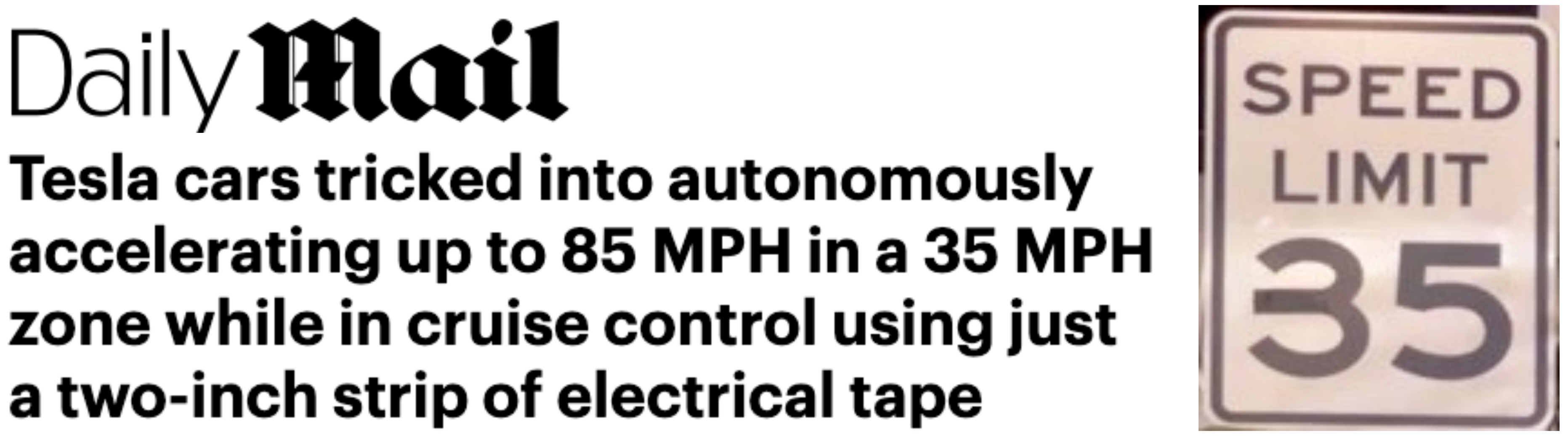
https://openaccess.thecvf.com/content_ICCV_2017/papers/Metzen_Universal_Adversarial_Perturbations_ICCV_2017_paper.pdf

LOYOLA UNIVERSITY CHICAGO

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Evasion
The attacker avoids data detection
(or causes misclassification) by the system,
usually by presenting adversarial data
(at inference time).


Daily Mail
Tesla cars tricked into autonomously accelerating up to 85 MPH in a 35 MPH zone while in cruise control using just a two-inch strip of electrical tape

https://www.youtube.com/watch?v=4uGV_fRj0UA

LOYOLA
UNIVERSITY CHICAGO

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Evasion
The attacker avoids data detection
(or causes misclassification) by the system,
usually by presenting adversarial data
(at inference time).

Repudiation

https://www.youtube.com/watch?v=_PoudPCevN0

LOYOLA UNIVERSITY CHICAGO

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Evasion
The attacker avoids data detection
(or causes misclassification) by the system,
usually by presenting adversarial data
(at inference time).

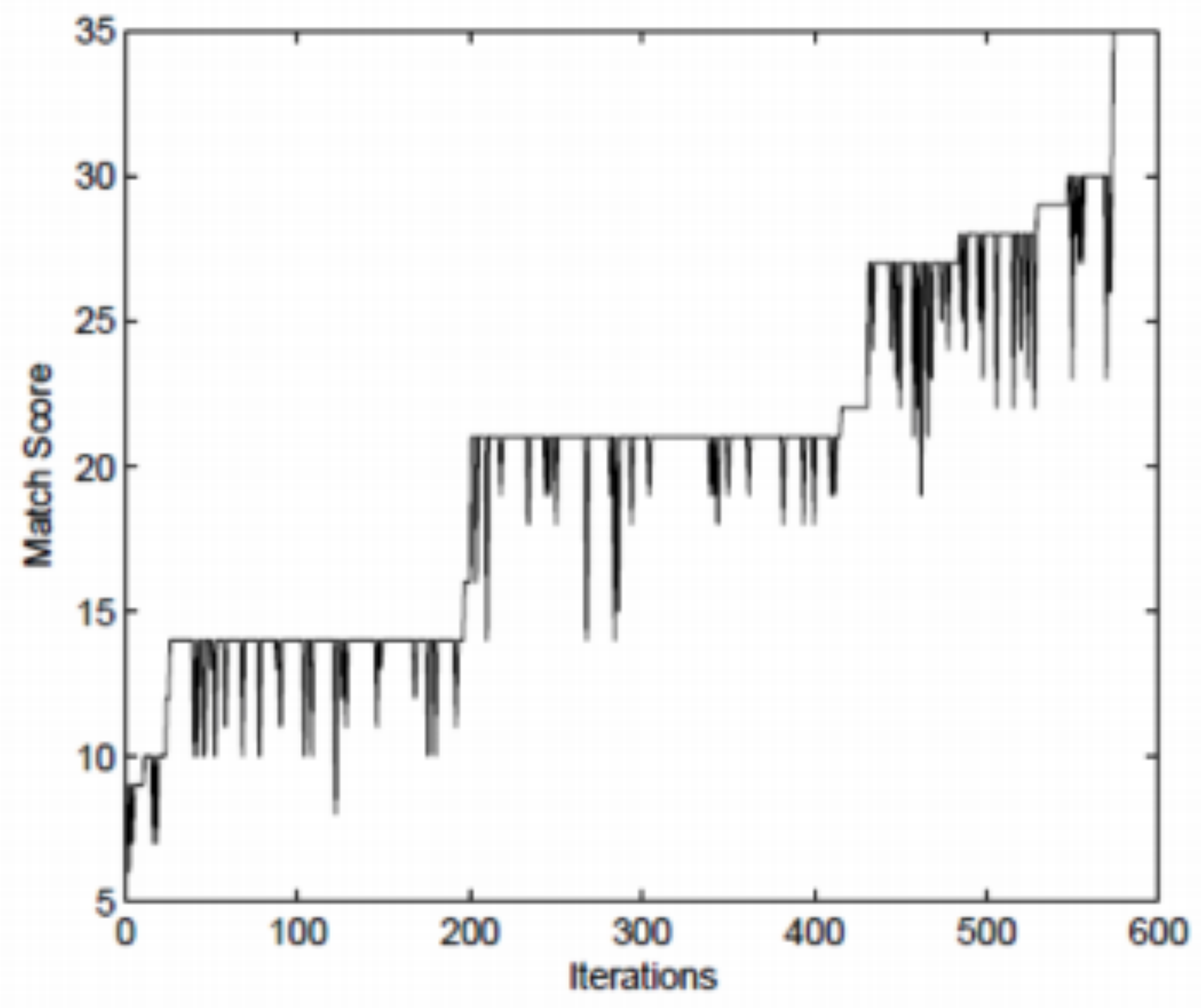
Spoofing

LOYOLA UNIVERSITY CHICAGO

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Hill Climbing
The attacker iteratively provides
synthetic adversarial samples to the system
(at inference time).
At each iteration, the attacker observes how the output
scores are progressing (gray-box attack).



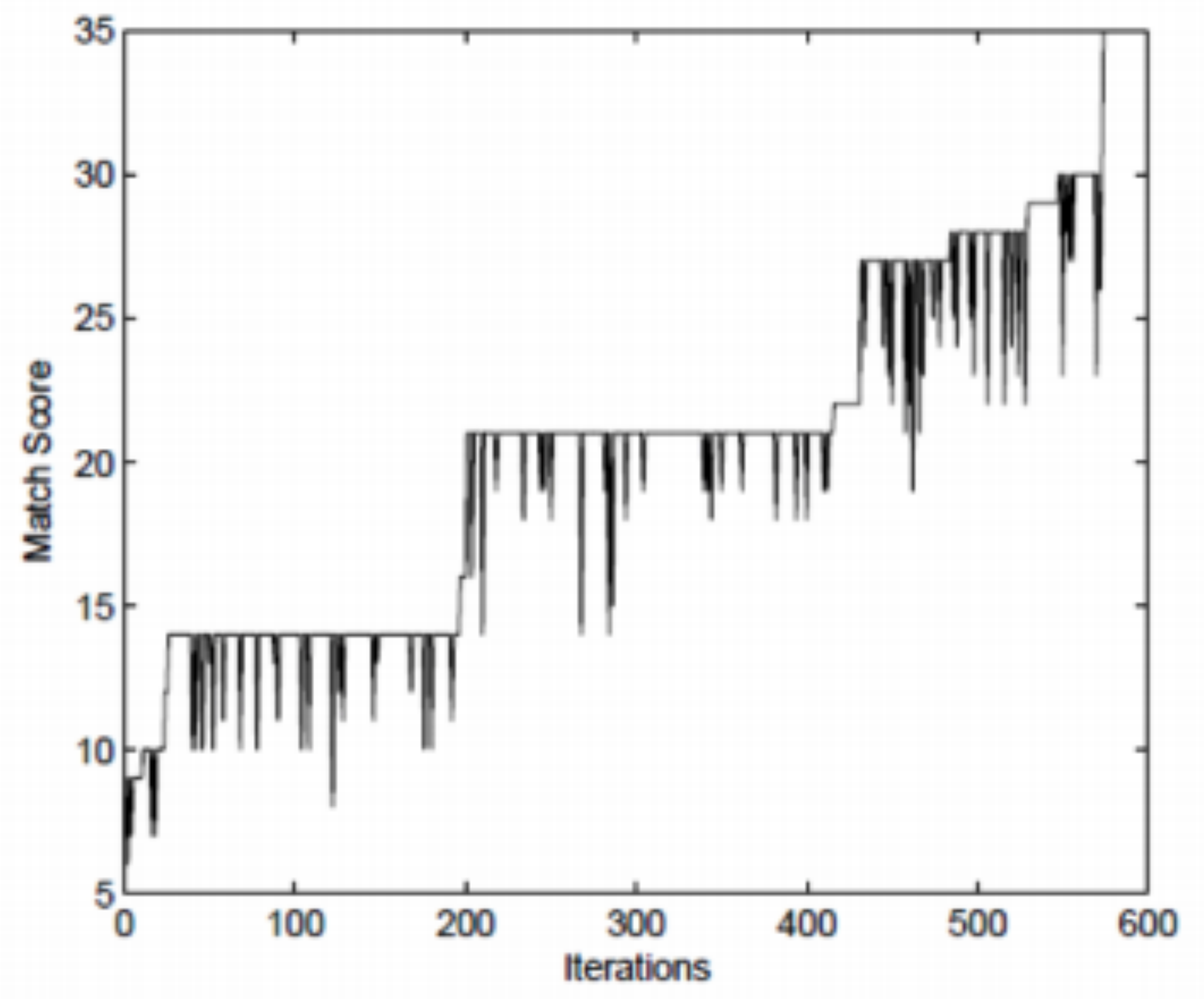
Martinez-Diaz et al.
Hill-Climbing and Brute-Force Attacks on Biometric Systems: A Case Study in Match-on-Card Fingerprint Verification
IEEE ICCST, 2006

LOYOLA UNIVERSITY CHICAGO

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Hill Climbing
With such progress feedback,
the attacker can guide the generation of better and better synthetic data samples, up the point of trespassing the decision threshold.



Martinez-Diaz et al.
Hill-Climbing and Brute-Force Attacks on Biometric Systems: A Case Study in Match-on-Card Fingerprint Verification
IEEE ICCST, 2006

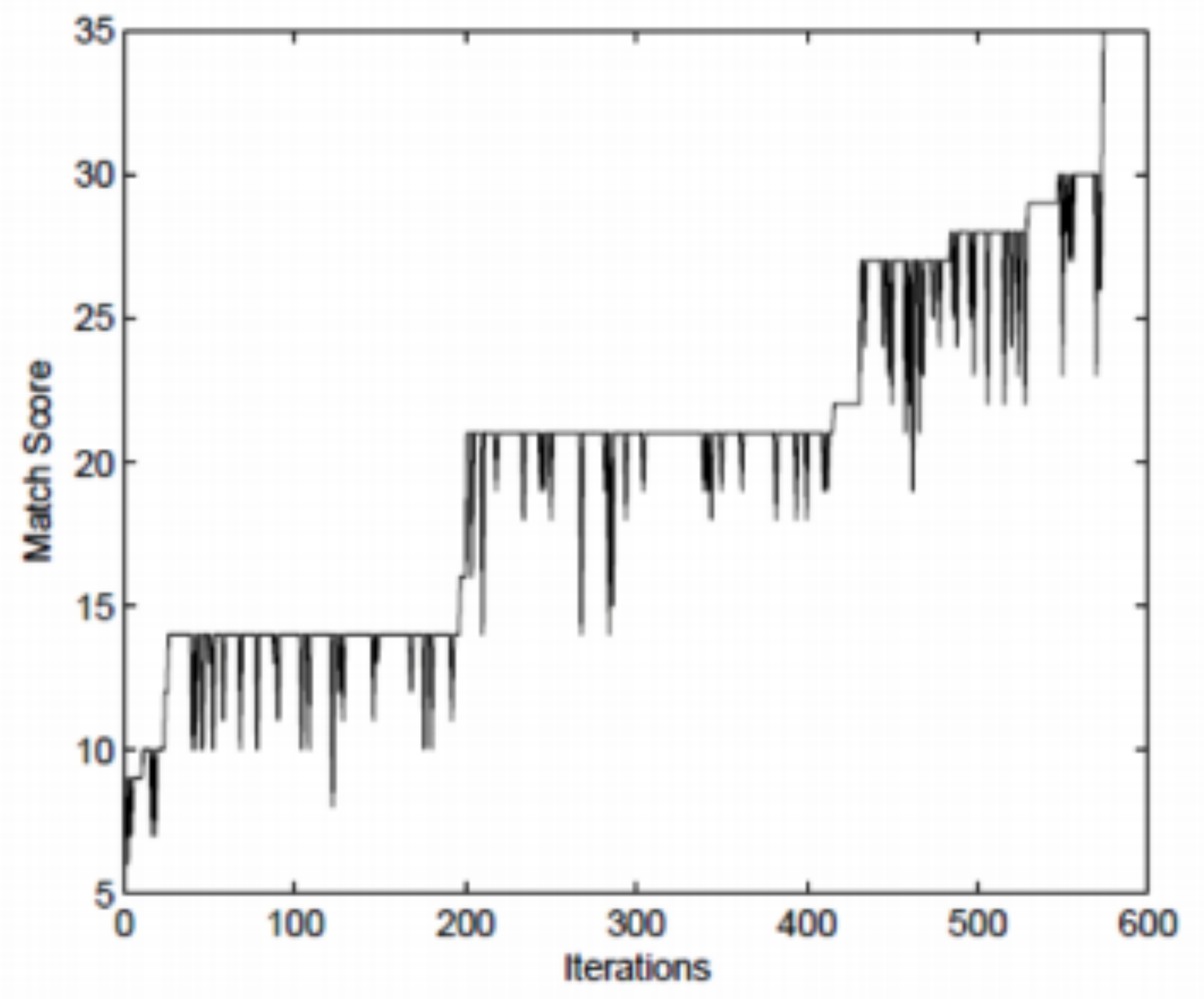
LOYOLA UNIVERSITY CHICAGO

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Hill Climbing
With such progress feedback,
the attacker can guide the generation of better and better synthetic data samples, up the point of trespassing the decision threshold.

One can make it fully data-driven with GANs.

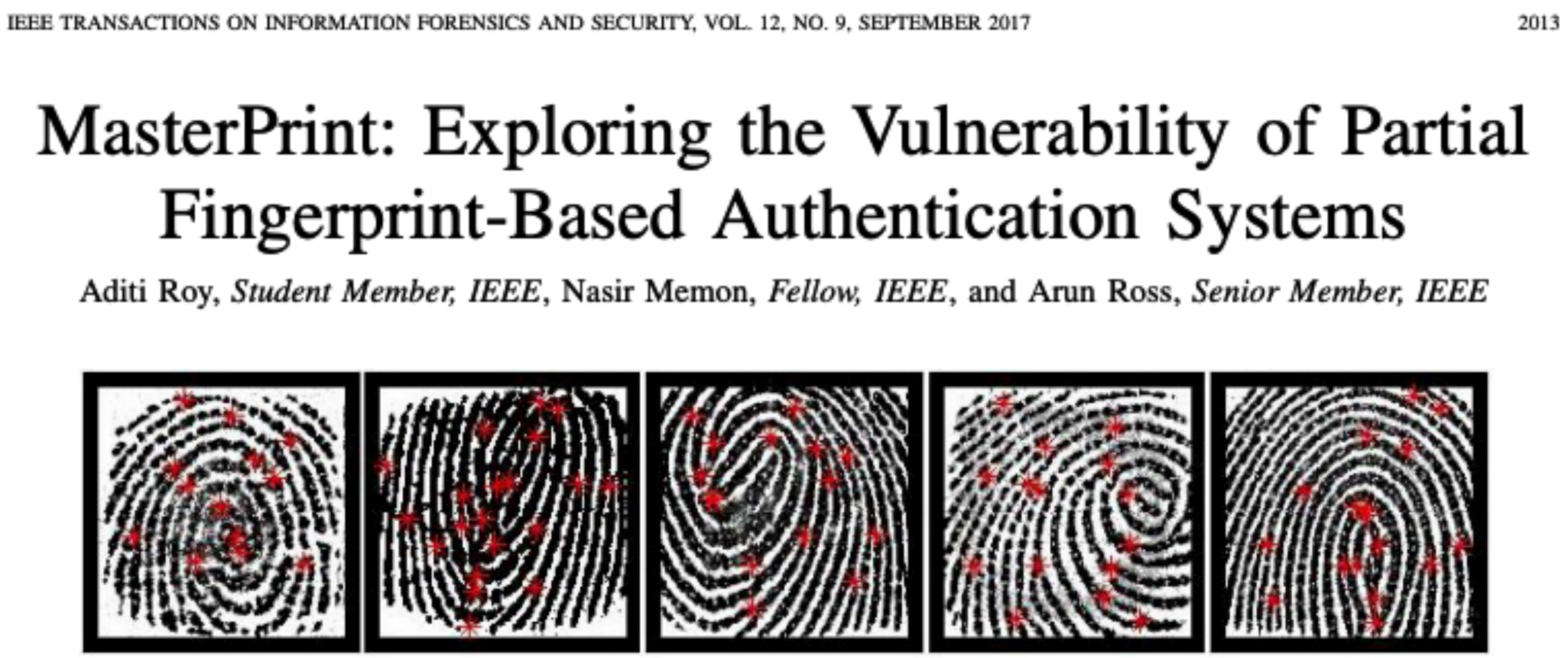


Martinez-Diaz et al.
Hill-Climbing and Brute-Force Attacks on Biometric Systems: A Case Study in Match-on-Card Fingerprint Verification
IEEE ICCST, 2006

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Master Key
Similar to hill climbing, but the target is to generate an adversarial sample that matches multiple classes.


IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 12, NO. 9, SEPTEMBER 2017 2013

MasterPrint: Exploring the Vulnerability of Partial Fingerprint-Based Authentication Systems

Aditi Roy, *Student Member, IEEE*, Nasir Memon, *Fellow, IEEE*, and Arun Ross, *Senior Member, IEEE*

https://www.cse.msu.edu/~rossarun/pubsRoyMemonRossMasterPrint_TIFS2017.pdf

LOYOLA UNIVERSITY CHICAGO

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Denial of Service
The system is either fooled to mostly take its more intense processing path, or flooded with multiple input data to process.

https://www.theguardian.com/world/2019/aug/13/the-fashion-line-designed-to-trick-surveillance-cameras

LOYOLA UNIVERSITY CHICAGO

Adversarial Machine Learning

Within Machine Learning
Data driven.

Attack Examples
Data Poisoning (or Backdoor, or Trojan)
Mislabeled adversarial data are covertly included among the training samples.

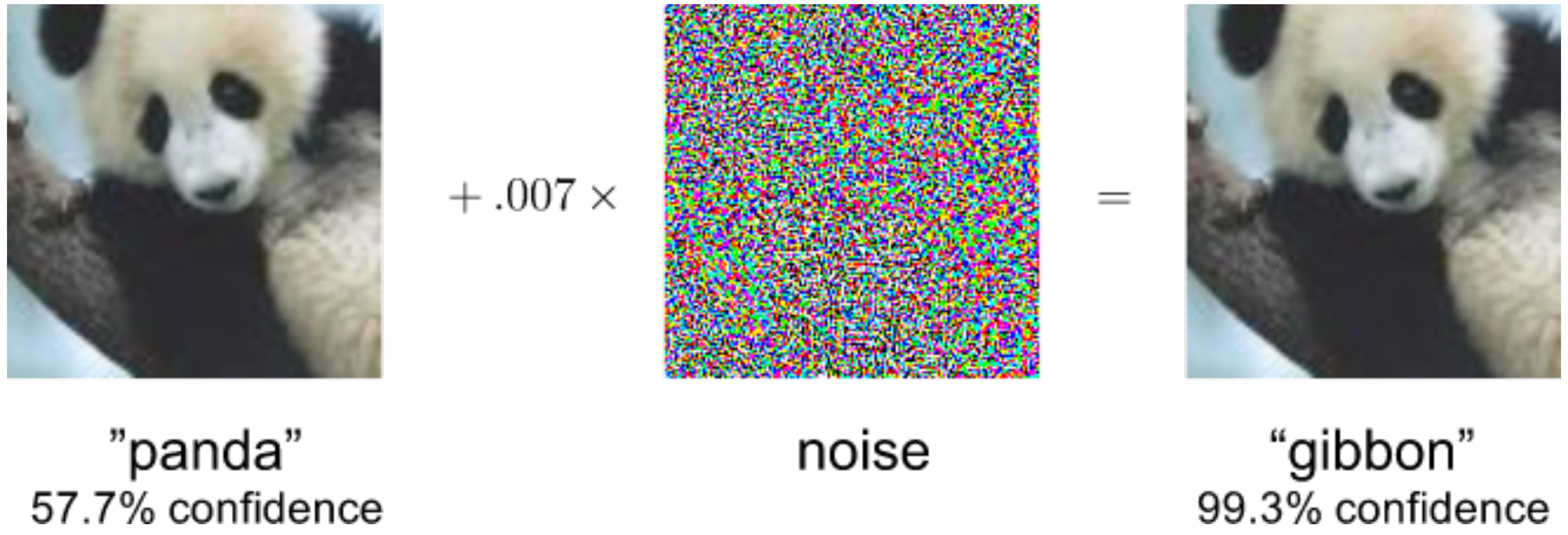
Gu et al.
BadNets: Evaluating Backdooring Attacks on Deep Neural Networks
IEEE Access, 2019
https://ieeexplore.ieee.org/document/8685687

LOYOLA UNIVERSITY CHICAGO

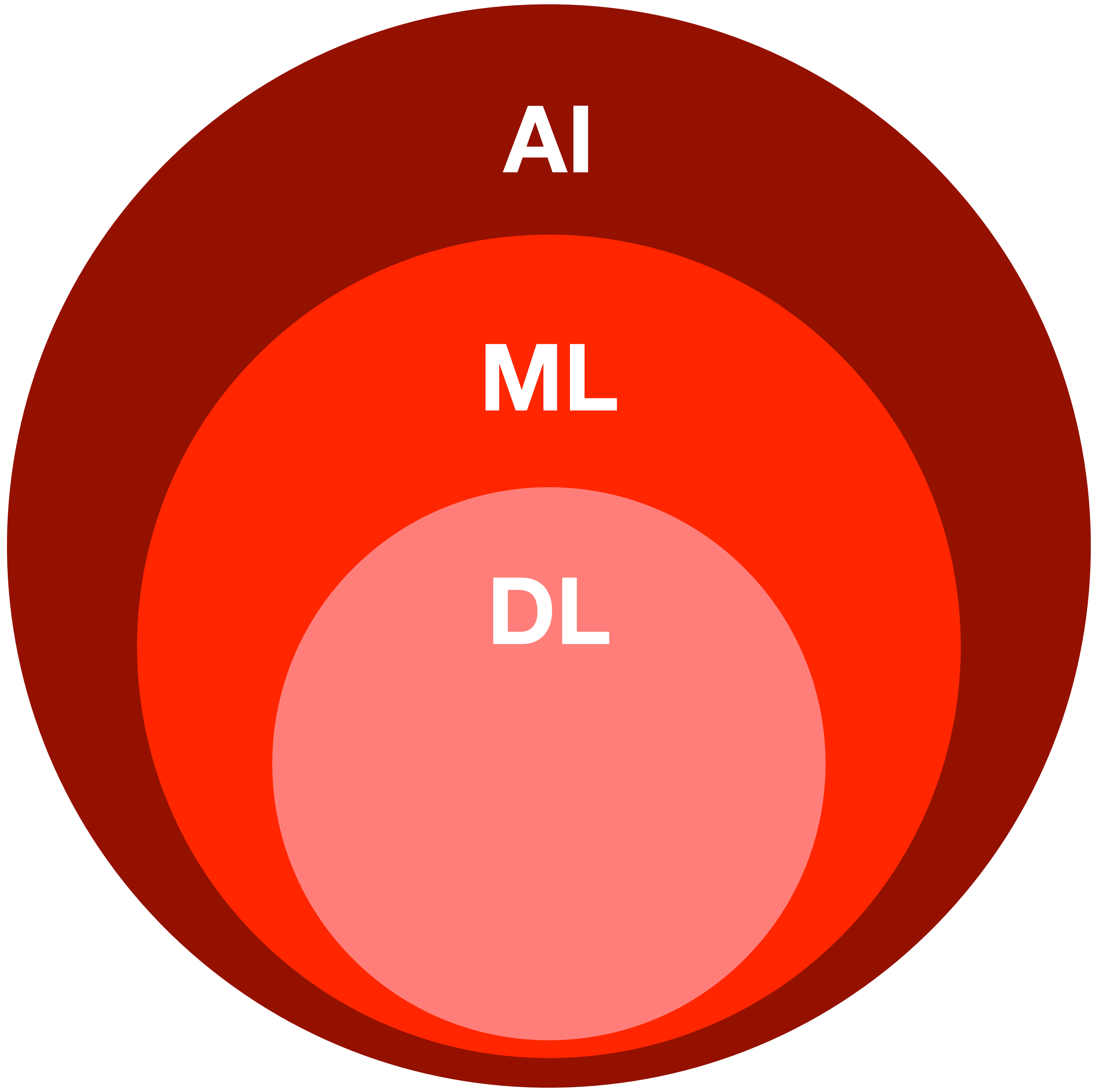
Adversarial Attacks on DL

Harnessing Adversarial Examples

Objective
Learn how to add imperceptive noise to input data in a way that changes the model's class prediction, causing misclassification.



https://arxiv.org/pdf/1412.6572.pdf

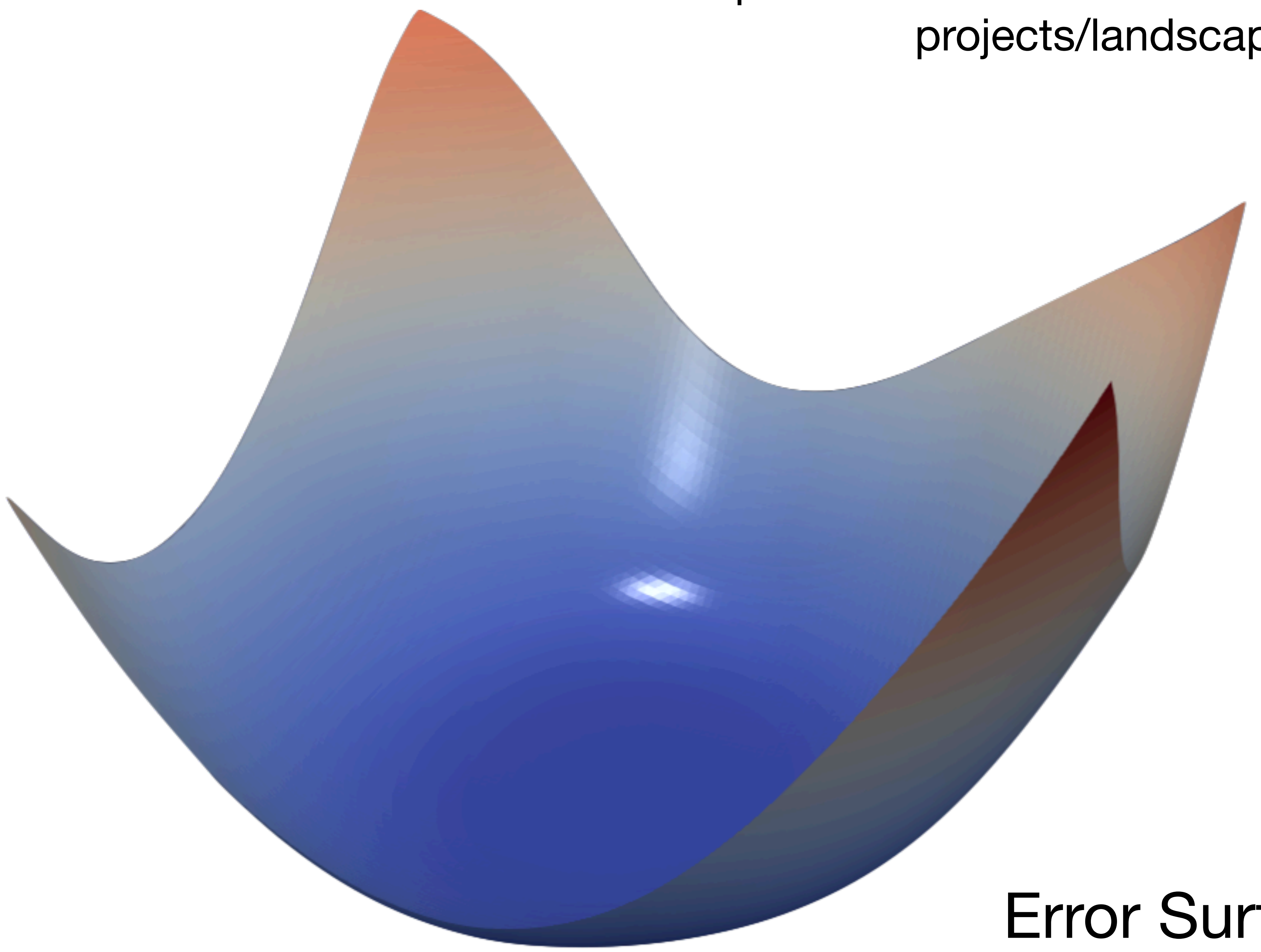


Adversarial Attacks on DL

Harnessing Adversarial Examples

Possible Solution
Fast Gradient Sign Method (FGSM)

https://www.cs.umd.edu/~tomg/projects/landscapes/



Error Surface

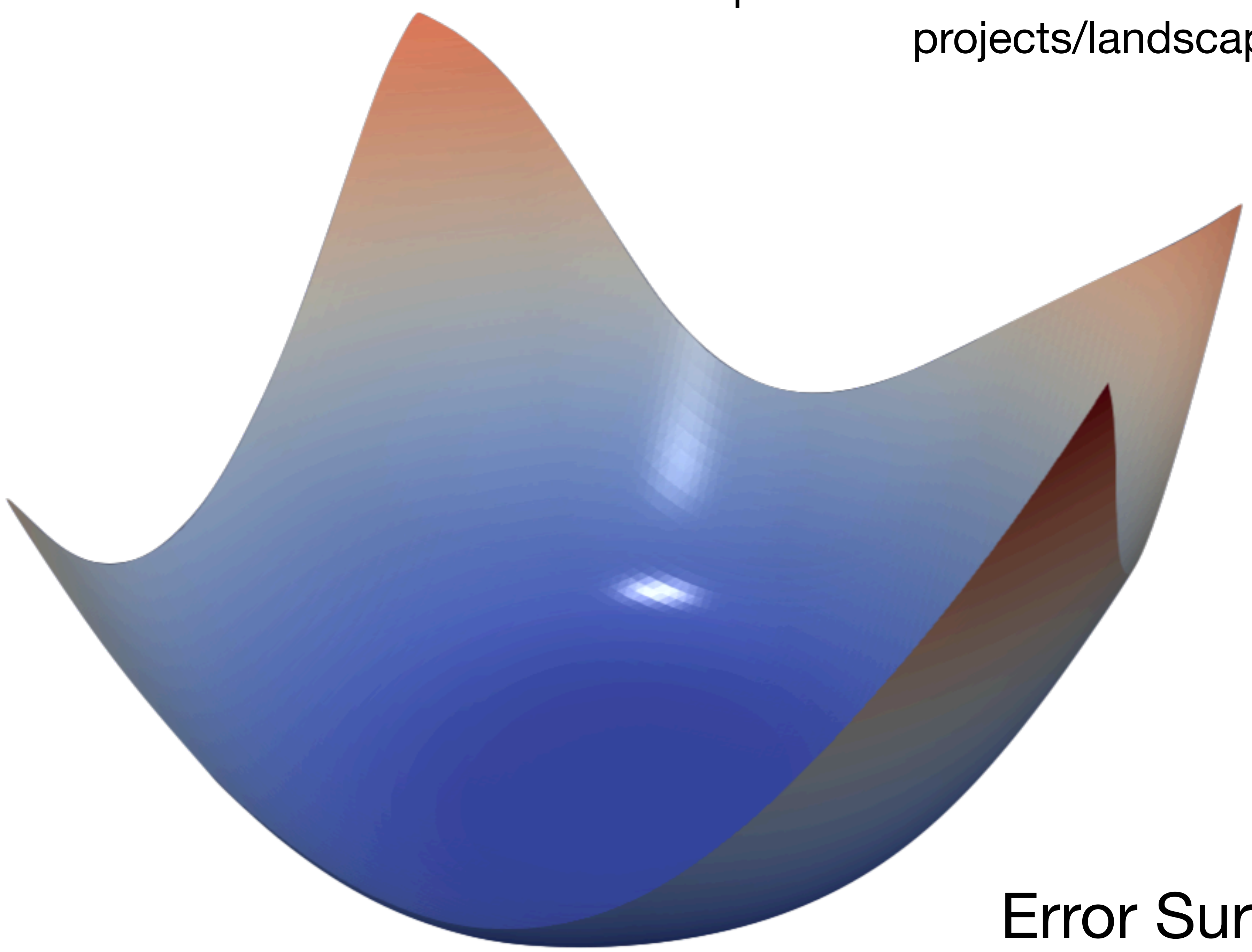
Adversarial Attacks on DL

Harnessing Adversarial Examples

Possible Solution
Fast Gradient Sign Method (FGSM)

Typical Training Solution
Minimize the error (actual label, predicted label) according to the training data batch by "walking" on the error surface to the opposite direction of the error gradient.

https://www.cs.umd.edu/~tomg/projects/landscapes/



Error Surface

LOYOLA UNIVERSITY CHICAGO

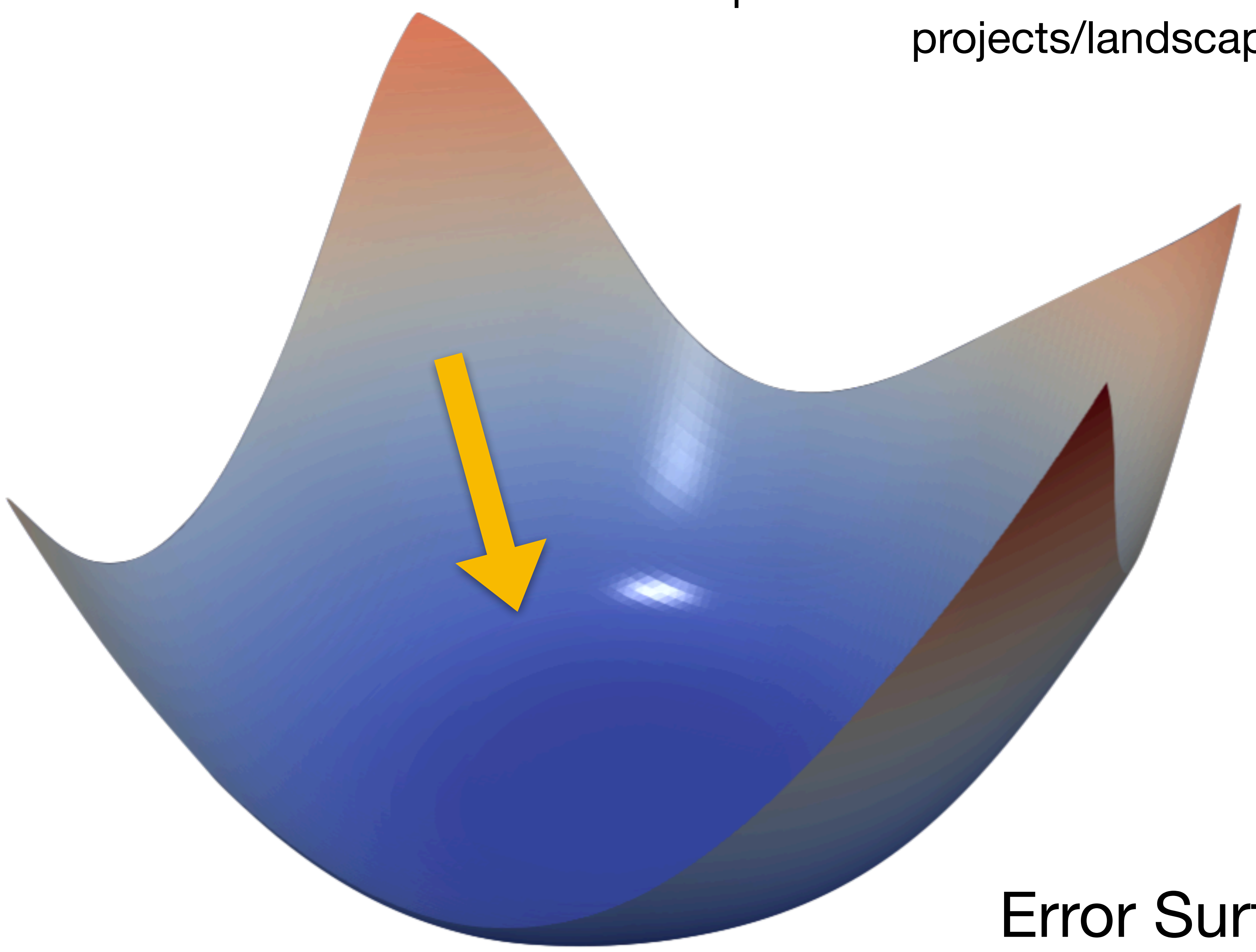
Adversarial Attacks on DL

Harnessing Adversarial Examples

Possible Solution
Fast Gradient Sign Method (FGSM)

Typical Training Solution
Minimize the error (actual label, predicted label) according to the training data batch by “walking” on the error surface to the opposite direction of the error gradient.

https://www.cs.umd.edu/~tomg/projects/landscapes/



Error Surface

$$new\_weight = current\_weight - learning\_rate \times sign(\nabla_{batch\_error})$$

LOYOLA UNIVERSITY CHICAGO

Adversarial Attacks on DL

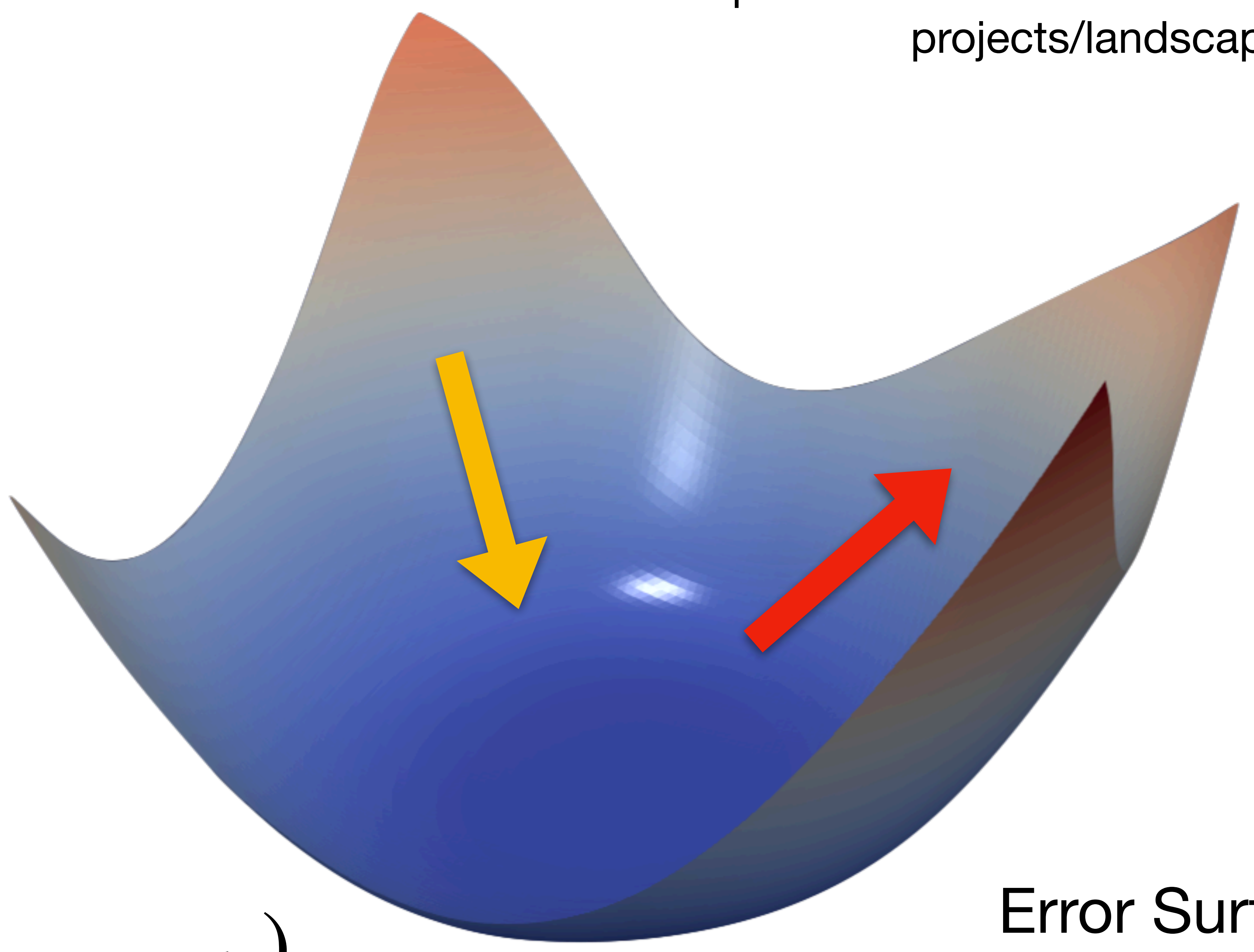
Harnessing Adversarial Examples

Possible Solution
Fast Gradient Sign Method (FGSM)

FGSM
Given one sample (that will be the adversarial data), maximize the error by "walking" on the error surface to the direction of the error gradient.

$$new\_sample = current\_sample + \epsilon \times sign(\nabla_{new\_sample})$$

https://www.cs.umd.edu/~tomg/projects/landscapes/



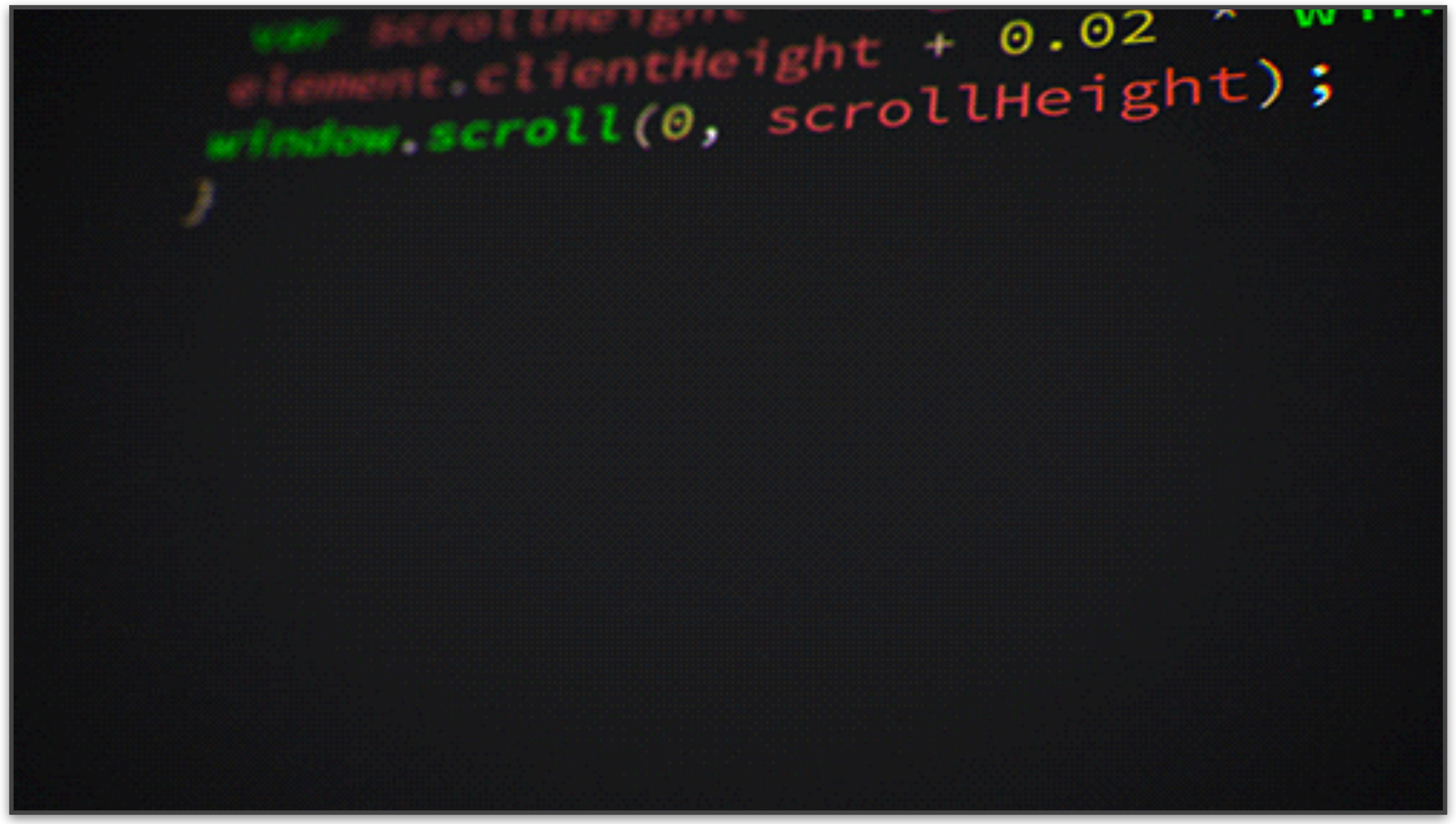
Error Surface

LOYOLA UNIVERSITY CHICAGO

Adversarial Attacks on DL

Harnessing Adversarial Examples

Possible Solution
Fast Gradient Sign Method (FGSM)

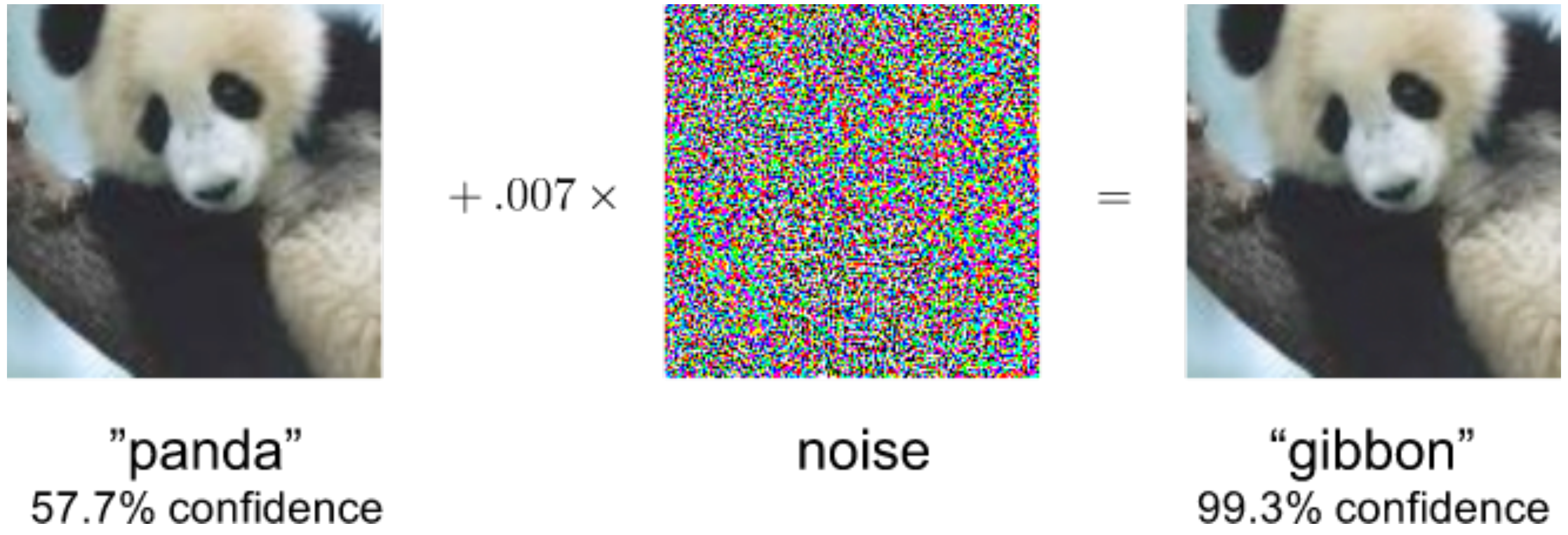


Algorithm
1. Forward-propagate sample through the network.
2. Compute the (actual label, predicted label) error.
3. Back-propagate the error gradient to the sample.
4. Tweak the sample features in the direction that maximizes the error by adding “noise”.

LOYOLA UNIVERSITY CHICAGO

Adversarial Attacks on DL

Harnessing Adversarial Examples



https://arxiv.org/pdf/1412.6572.pdf

Possible Solution
Fast Gradient Sign Method (FGSM)

Algorithm
1. Forward-propagate sample through the network.
2. Compute the (actual label, predicted label) error.
3. Back-propagate the error gradient to the sample.
4. Tweak the sample features in the direction that maximizes the error by adding "noise".

$\epsilon \times sign(\nabla_{new\_sample})$: noise.

$sign(\nabla_{new\_sample})$: direction that maximizes the error.

ϵ: control the perceptive-mislabel trade-off.

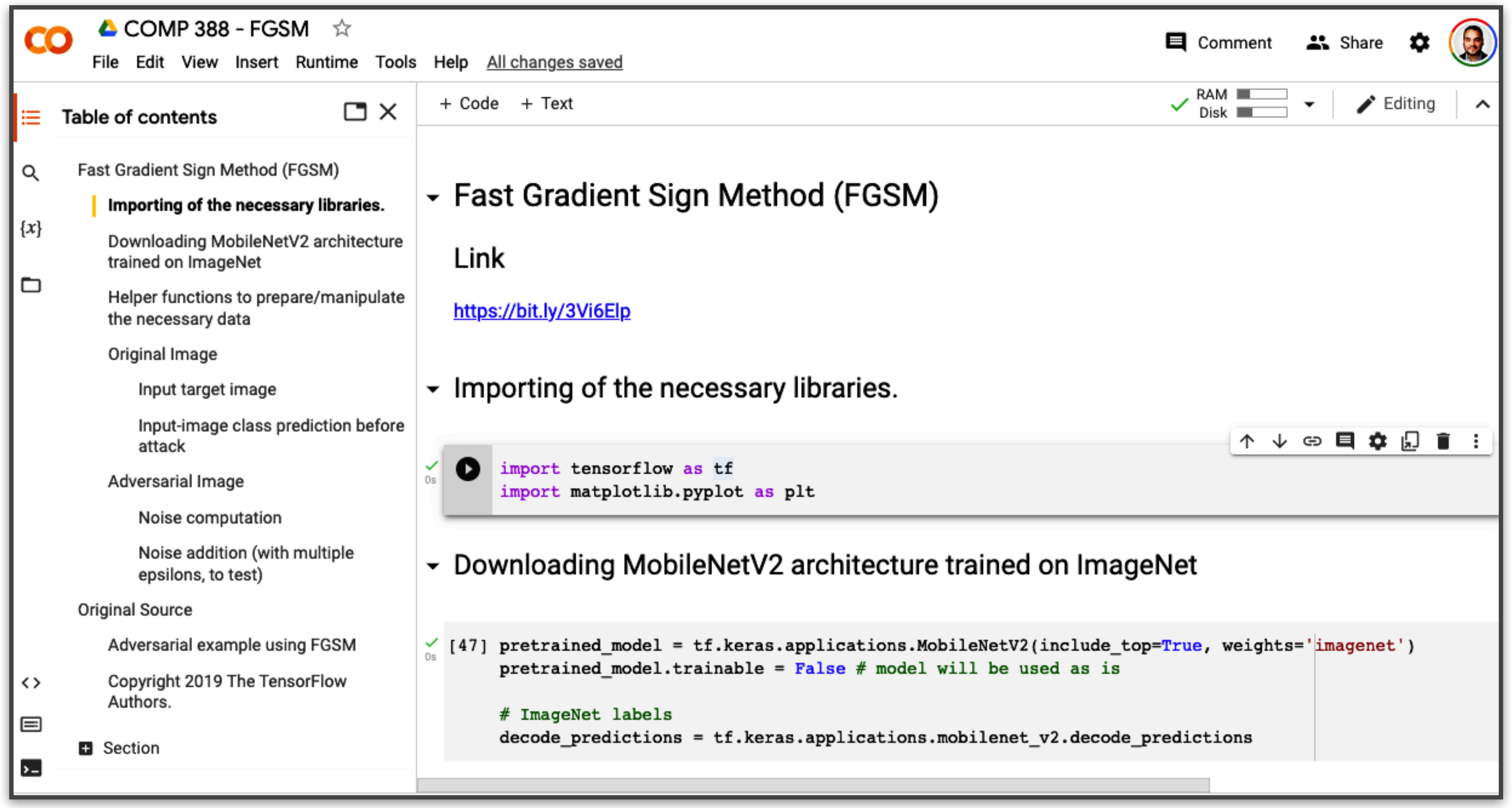
Practical Activity 2

Work in Pairs

Use Google Colab at https://bit.ly/3Vi6Elp

Observe the epsilon's perceptive-mislabel trade-off.
Try with different images (source: https://bit.ly/3OqEMsP).

Contributions or Question?



1. https://arxiv.org/abs/1801.07698

LOYOLA UNIVERSITY CHICAGO

Adversarial Machine Learning

Discussion Time

FGSM
What type of attack is FGSM
(e.g., evasion, white box)?

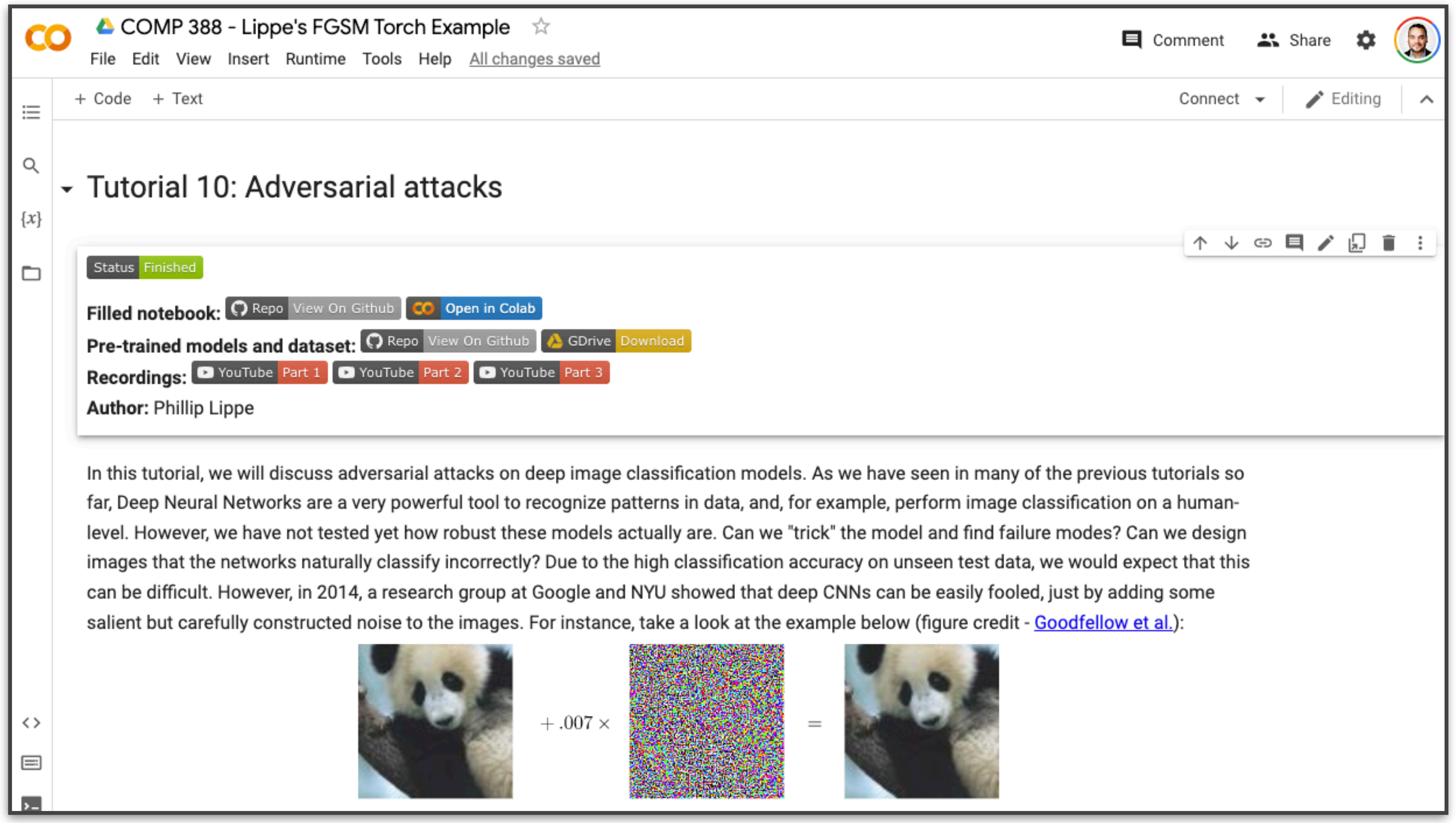
It worked on MobileNet;
would it work on other architectures?

LOYOLA UNIVERSITY CHICAGO

PyTorch Example

Watch it Later
Use Google Colab at
https://bit.ly/3V5VW1c

Phillip Lippe's PyTorch example
FGSM attack on ResNet-34
Other attack methods are available



1. https://arxiv.org/abs/1801.07698

Adversarial Machine Learning

Discussion Time

FGSM
What type of attack is FGSM
(e.g., evasion, white box)?

It worked on MobileNet;
would it work on other architectures?

Defenses
Procedures to avoid, detect,
or mitigate these attacks?

LOYOLA
UNIVERSITY CHICAGO

Deep Fake Detection

Discussion Time

Can a GAN-generated image be detected?
Yes, for now…
The discriminator always wins (the generator).

Open Problems
Cheap fake detection.
GAN architecture attribution.
GAN in the wild.

LOYOLA UNIVERSITY CHICAGO