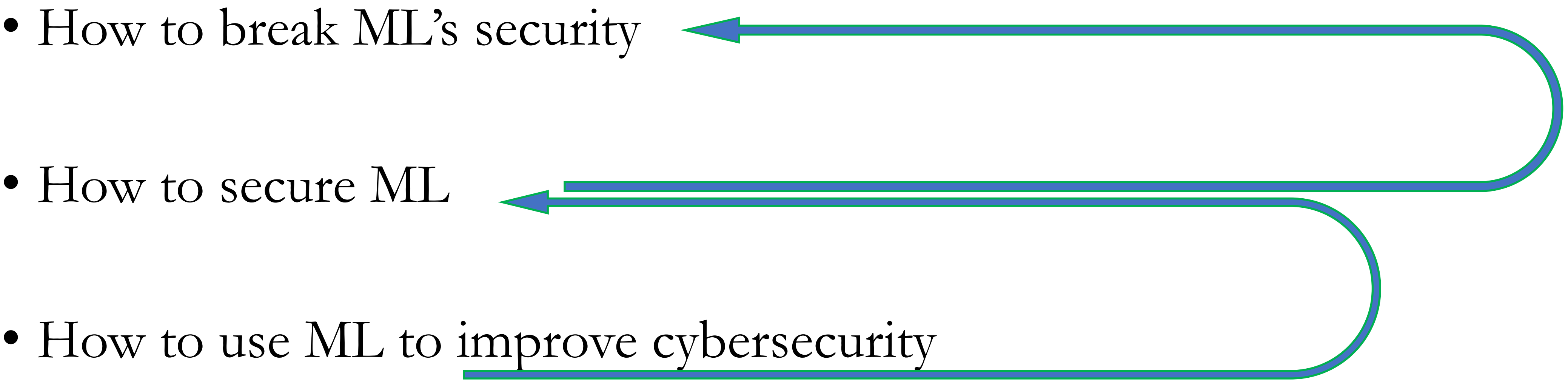
Cyber Security and AI

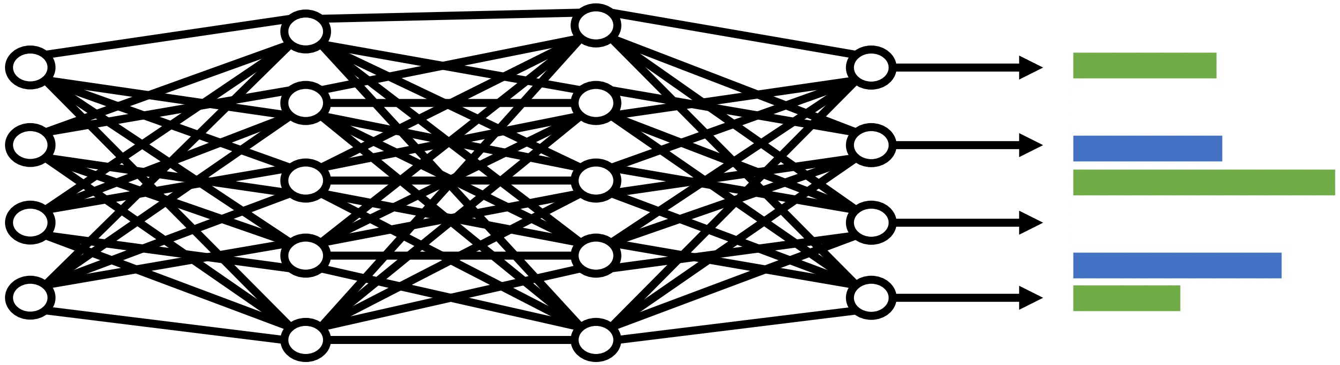
Dr. Britta Hale

Cybersecurity and Machine Learning



- How to break ML's security
- How to secure ML
- How to use ML to improve cybersecurity

Does the training work?

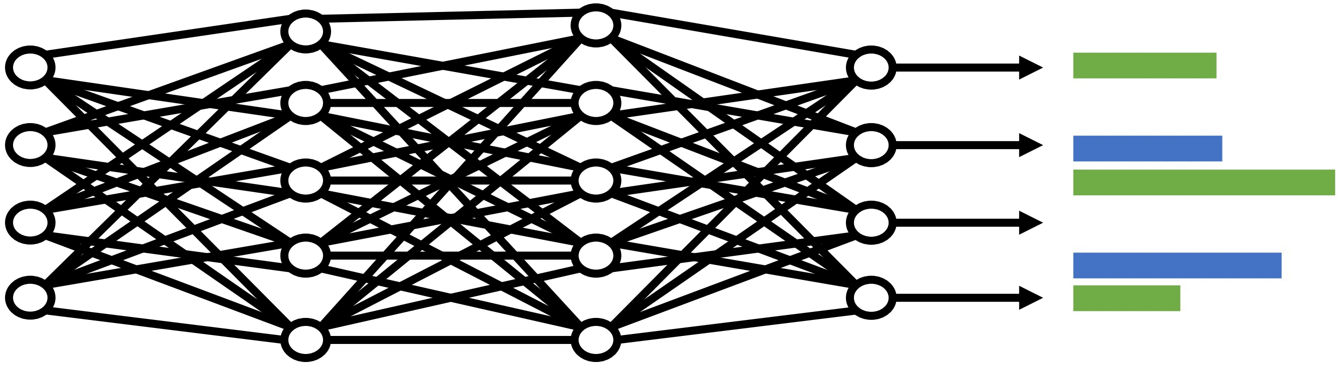


Accuracy

Can the training be circumvented?

Can the model be misinterpreted?

Can the model be abused?



Everything Else

Attacks During Training

e.g.

- Poisoning

- Trojans/Backdoors

A perfect memory…

https://www.nps.gov/safr/learn/news/free-volunteer-docent-training-tuesday-may-28-2019.htm;

[illegible], Free-Photos, DokaRyan
https://www.nps.gov/safr/learn/news/free-volunteer-docent-training-tuesday-may-28-2019.htm; https://www.flickr.com/photos/yellowstonenps/26648480439; https://commons.wikimedia.org/wiki/File:US_Navy_070921-N-0696M-084_Chief_of_Naval_Operations_(CNO)_Adm._Mike_Mullen_speaks_with_Vice_Adm._David_T._Oliver,_president_of_Naval_Post_Graduate_School_(NPS),_prior_to_speaking.jpg; https://navylive.dodlive.mil/2016/07/08/navy-innovation-silicon-valley-style/,
https://commons.wikimedia.org/wiki/File:NPS,_navy_leaders_continue_community_outreach_efforts_130110-N-[illegible]OH194-010.jpg

Poisoning

Integrity

- Confidence reduction

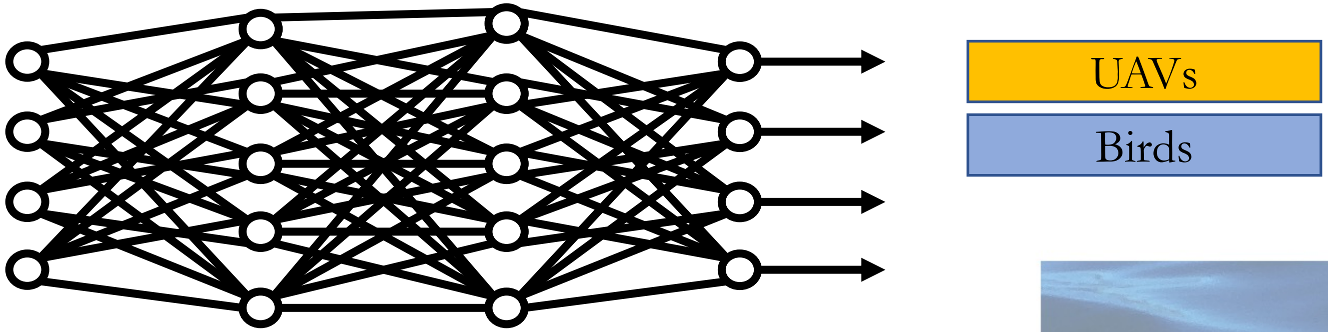
  do not change a class but highly impact the confidence

- Misclassification

  change a class without any specific target

- Targeted misclassification

  change a class to a particular target



UAVs

Birds

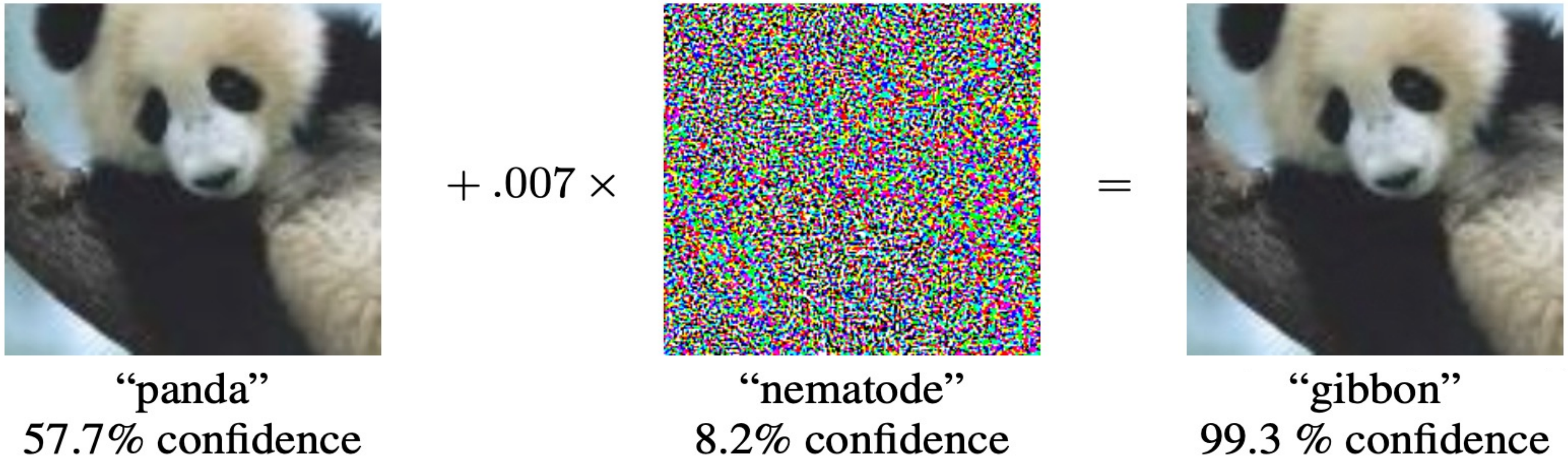
Poisoning

- Source/target misclassification

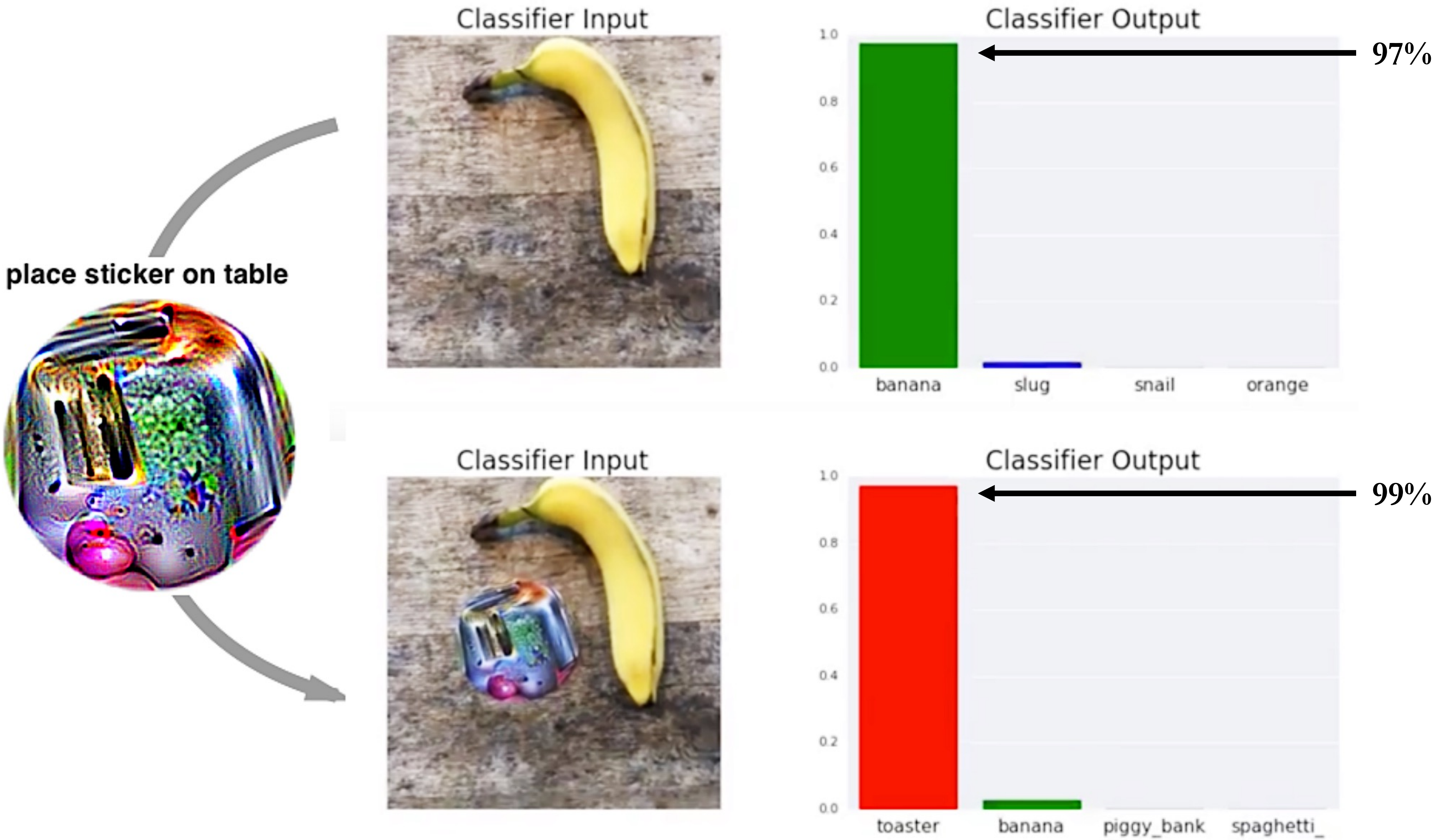
  change a particular source to a particular target

- Universal misclassification

  change any source to particular target



Goodfellow, Shlens, Szegedy ICLR 2015



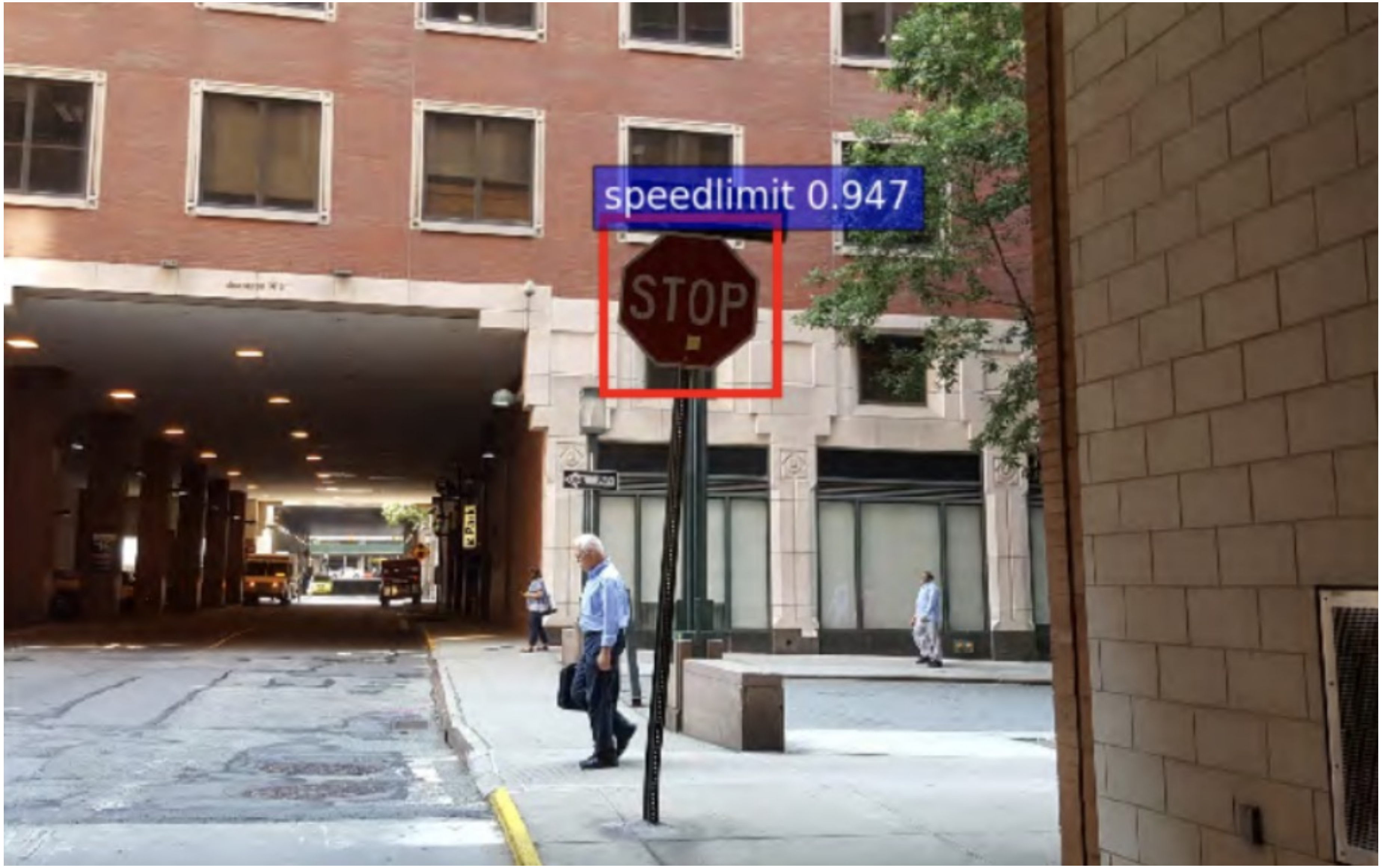
Tom B. Brown, Dandelion Mané , Aurko Roy, Martín Abadi, Justin Gilmer
https://arxiv.org/pdf/1712.09665.pdf

Trojans/Backdoor

1. Inverse network to create a trojan trigger
2. Retrain model with malicious data
3. Real inputs which activate the trojan trigger generate malicious behavior

Access to original dataset not necessarily required

Retraining can take minutes/hours
(vs. weeks/months for original model)



Gu, Dolan-Gavitt, Garg 2019

Defense

1. Outlier detection

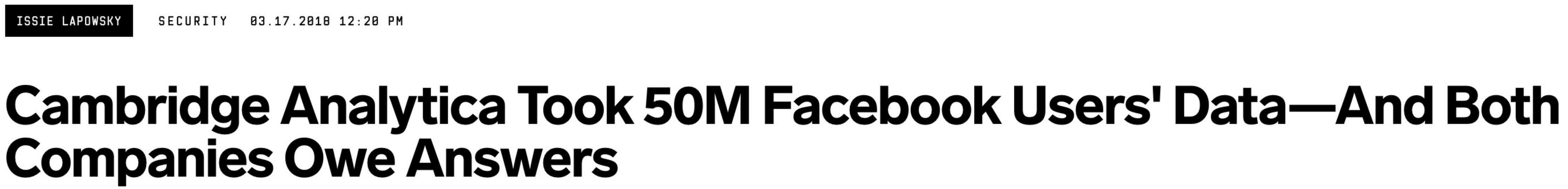
   How to define an outlier?

   What about data that was injected before filtering rules?

2. Test newly added training samples against current model for accuracy

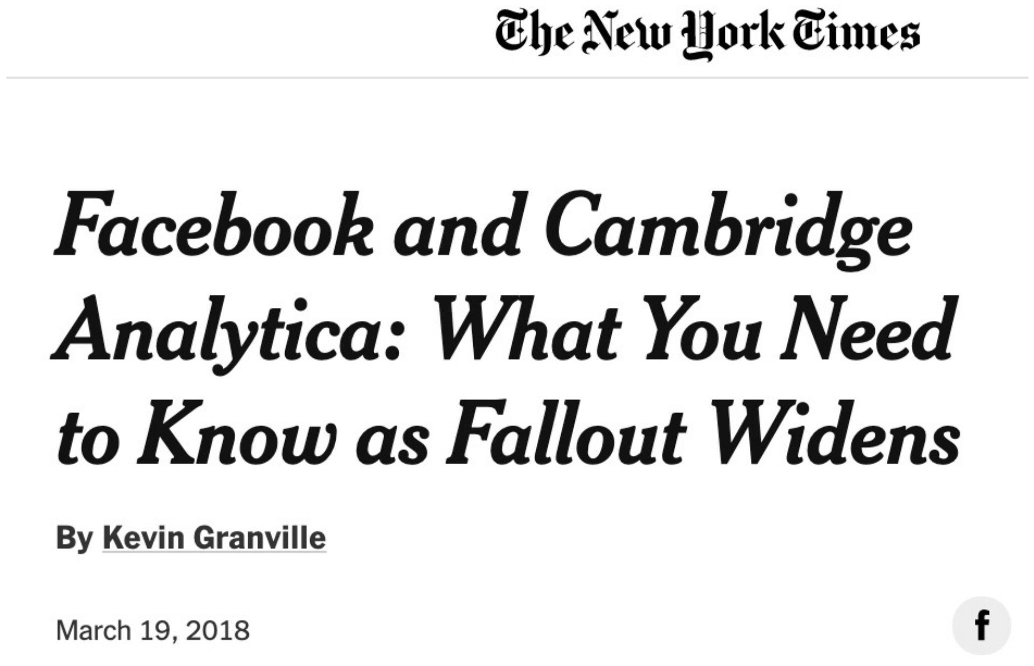
   What about trojans?

DATA



ISSIE LAPOWSKY SECURITY 03.17.2018 12:20 PM

Cambridge Analytica Took 50M Facebook Users' Data—And Both Companies Owe Answers



The New York Times

Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens

By Kevin Granville

March 19, 2018

Attacks During Production

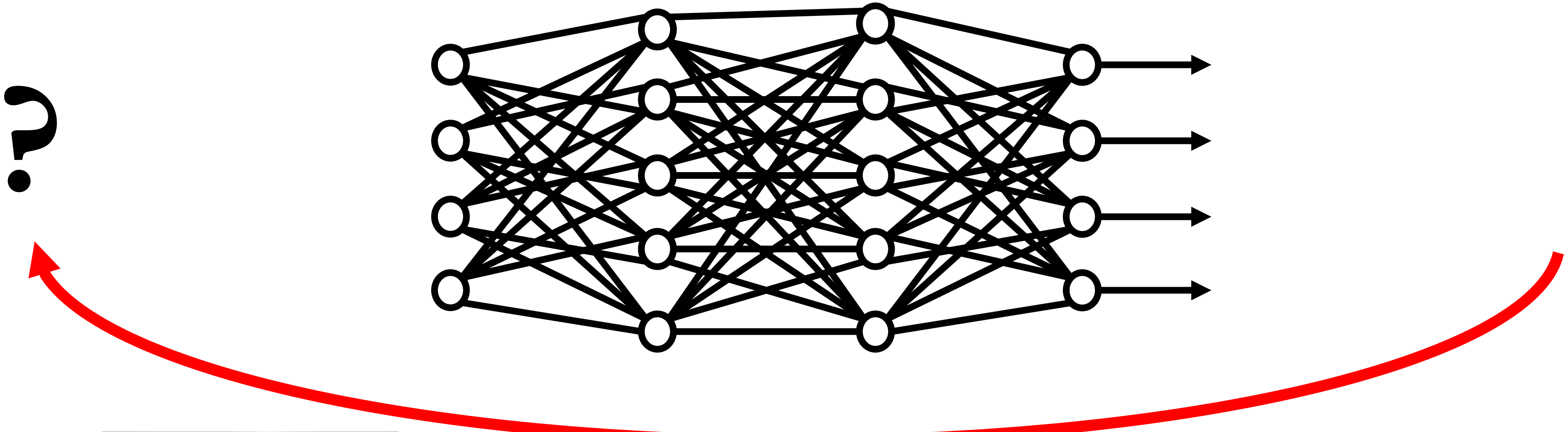
e.g.

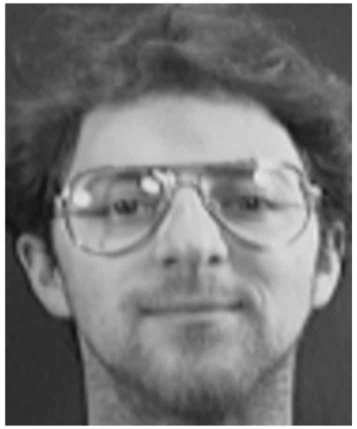
- Inference

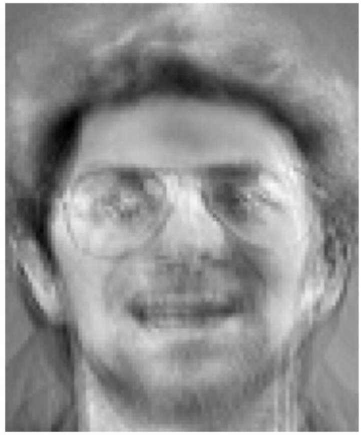
- Evasion

Inference

- Acquire information about dataset



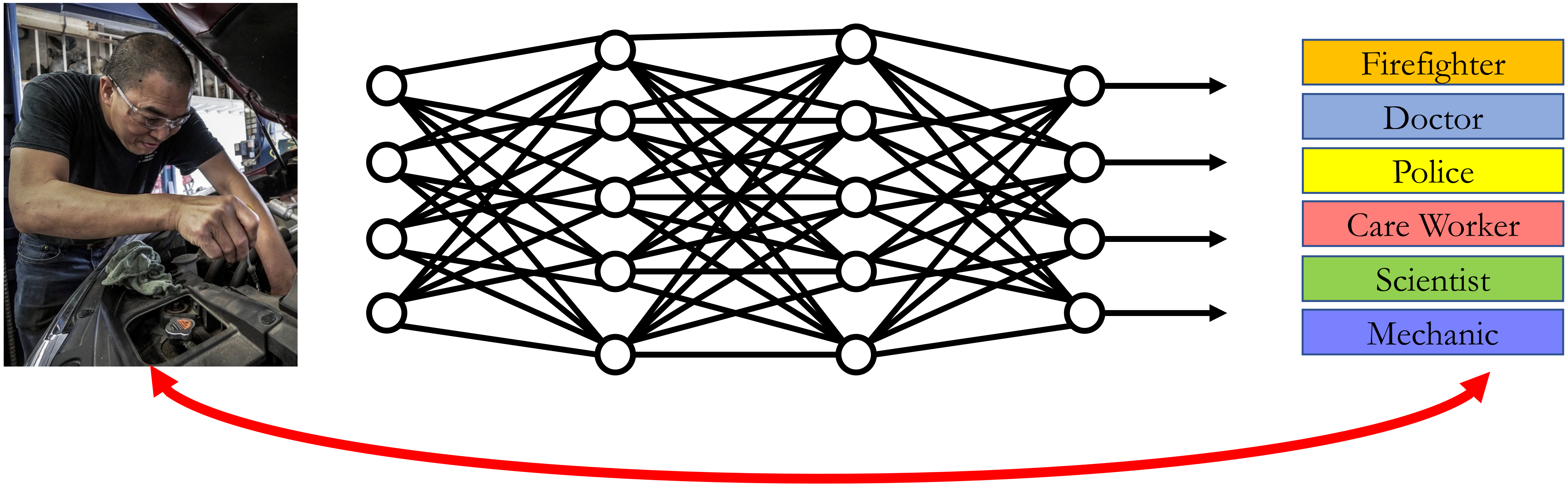




Fredrikson, Jha, and Ristenpart CCS ‘15

Inference

- Acquire information about dataset
- Membership inference / data attributes



Firefighter

Doctor

Police

Care Worker

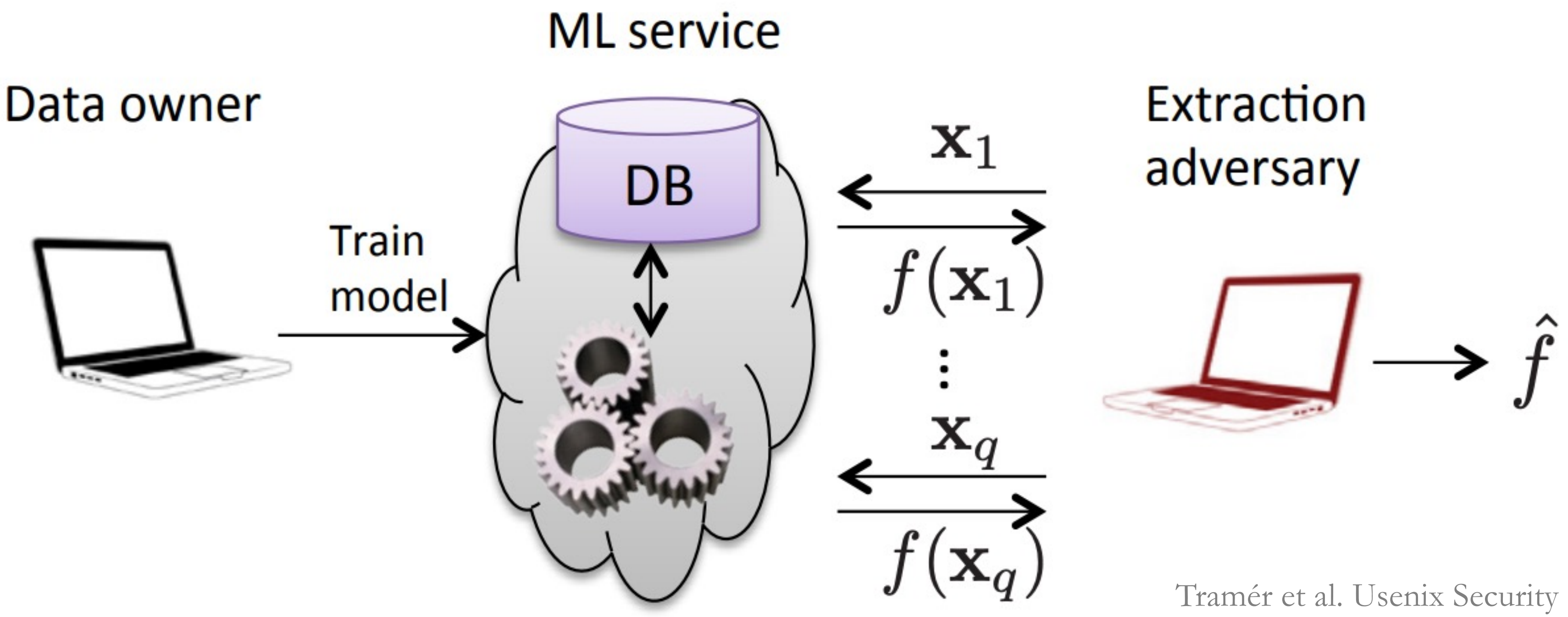
Scientist

Mechanic

Pixabay: DokaRyan

Inference

- Acquire information about dataset
- Membership inference / data attributes
- Model Extraction



Tramér et al. Usenix Security '16

Evasion

Does not shift classifier boundary, but pushes poisoning into dataset

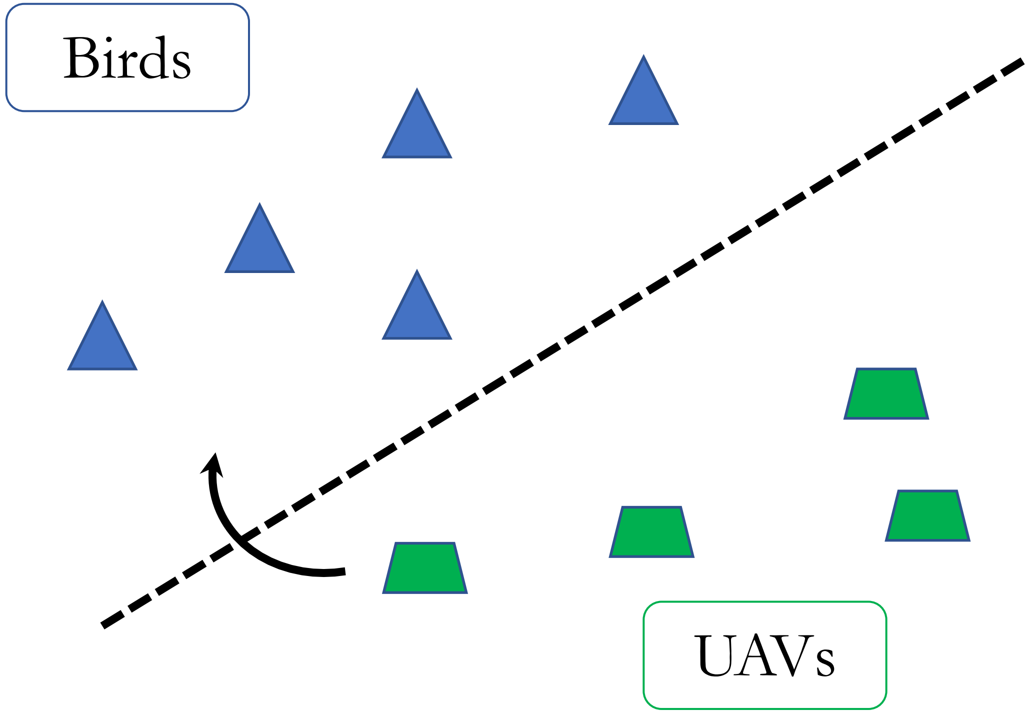
Evasion

To classify Birds as UAVs:

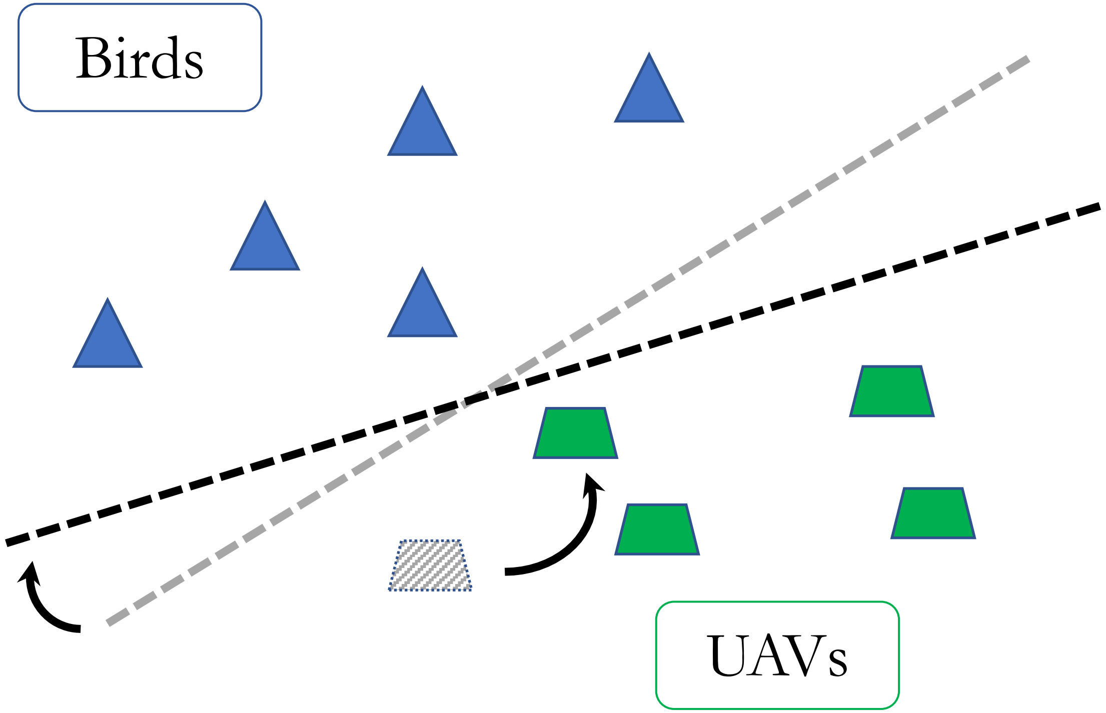
1. Change some UAVs to look closer to Birds
2. Keep UAVs labelled as UAVs
3. Add changed UAVs to training pool

Does not shift classifier boundary, but pushes poisoning into dataset

Evasion



Poisoning



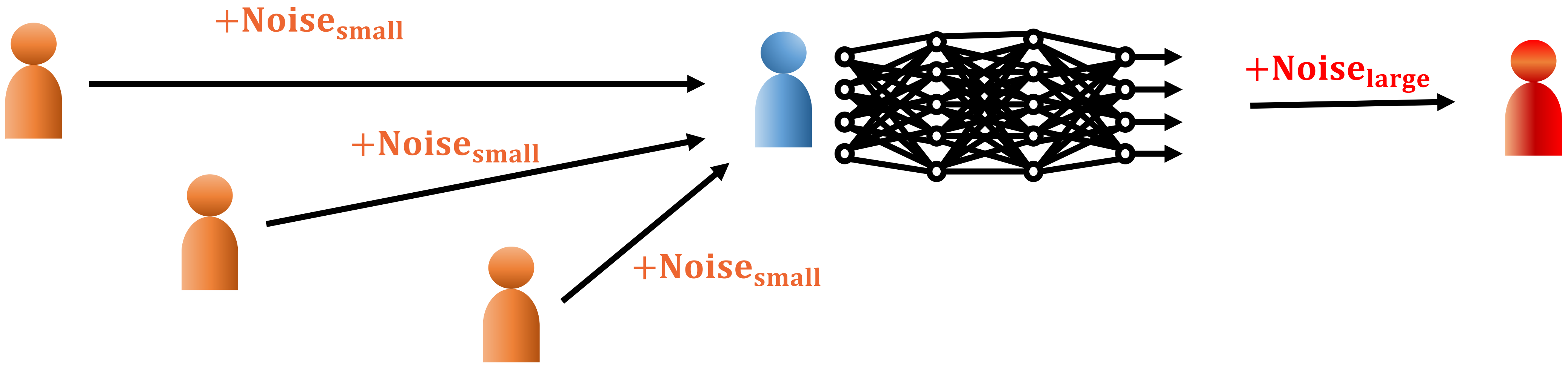
Defense

1. Differential Privacy

Defense

1. Differential Privacy

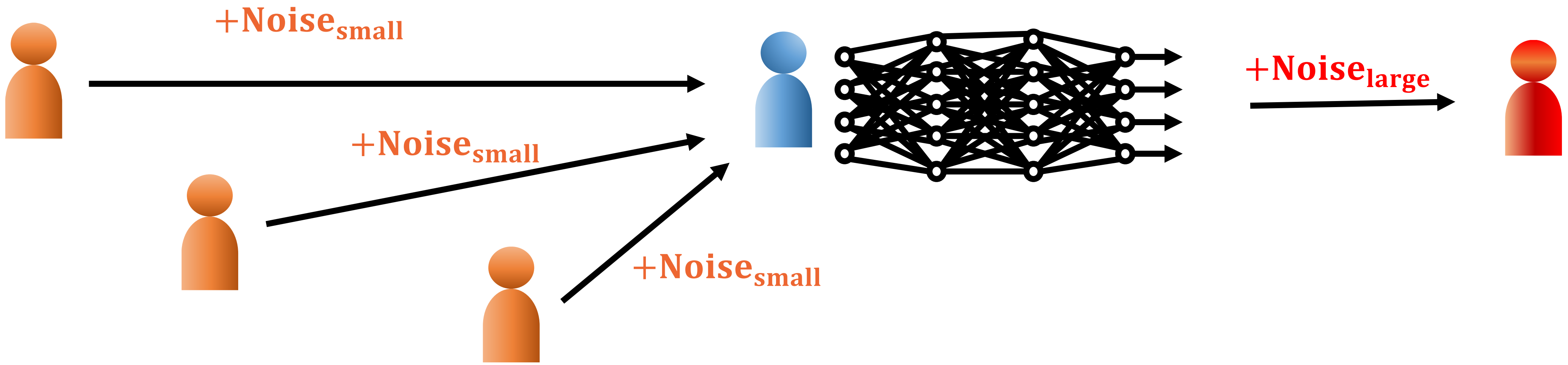
Goal: Try to hide individual data points


$+\text{Noise}_{\text{small}}$
$+\text{Noise}_{\text{small}}$
$+\text{Noise}_{\text{small}}$
$+\text{Noise}_{\text{large}}$

Defense

1. Differential Privacy

Problem: Model may become imbalanced



Defense

1. Differential Privacy

2. Don’t force guessing (“null” class)

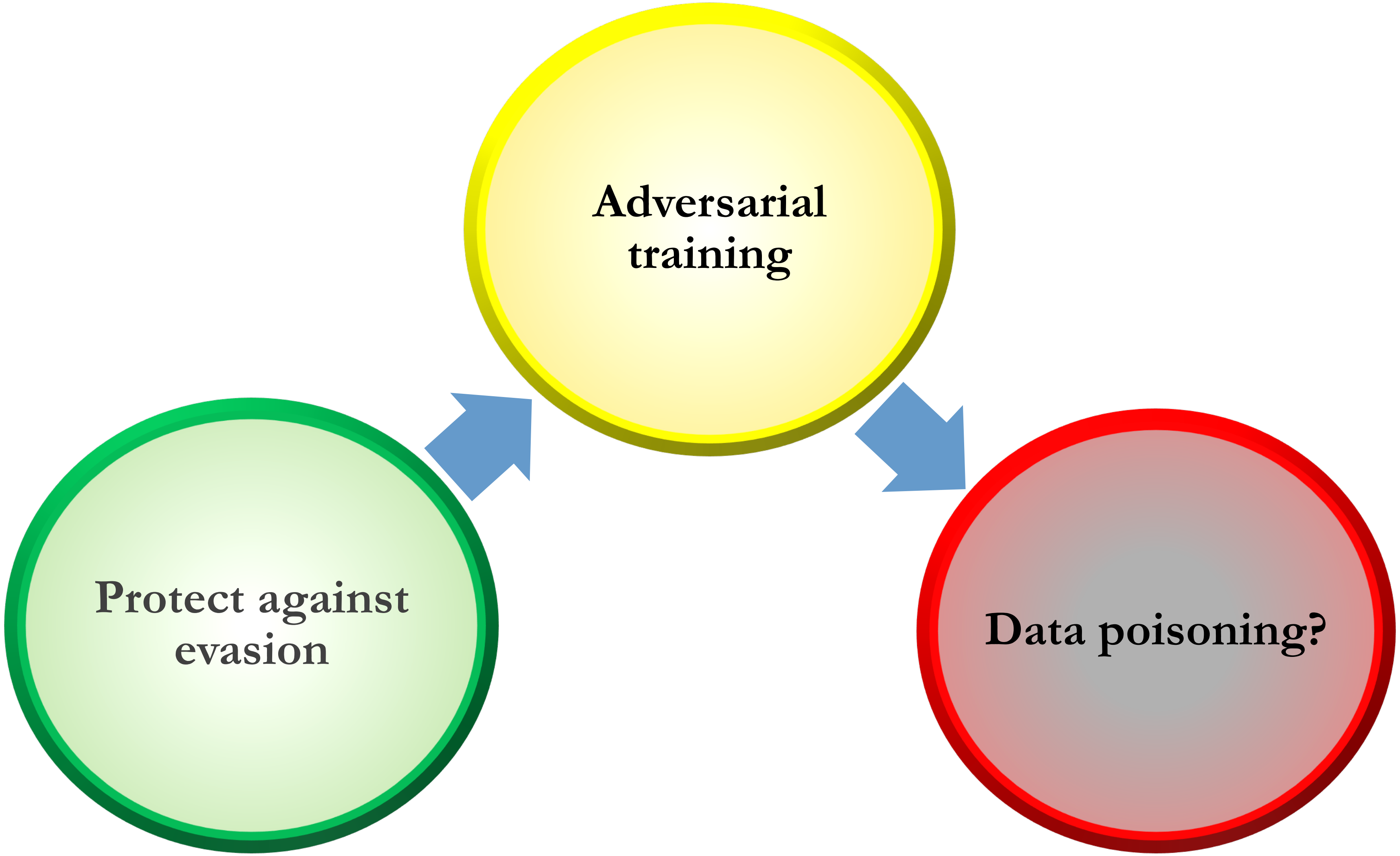
   Human overhead

3. Adversarial training

   What if the adversary uses different examples?

   What if you train on too many adversarial examples?

Poisoning vs. Evasion


Protect against evasion

Adversarial training

Data poisoning?

Human-in-the-loop experiment

190910-M-JX780-1337M, 200923-M-WH885-1259M, 190215-M-MJ974-0550Y, 190122-N-RI884-0063C, 200909-M-IP473-[illegible] 181019-M-ON629-1014C, 190909-M-VX661-1

210702-N-DO281-1665, 210105-M-OY155-1003C, https://pxhere.com/en/photo/741704, https://www.pxfuel.com/en/free-photo-eoimd, https://www.pxfuel.com/en/free-photo-itqni

DevSecOps-AI

Q&A

Cyber Security and AI