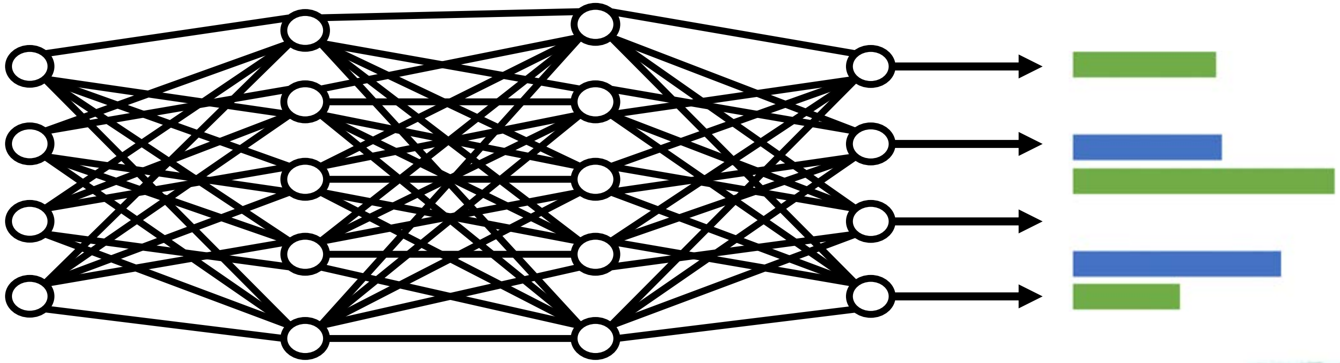
Cyber Security and AI

Britta Hale

Does the training work?

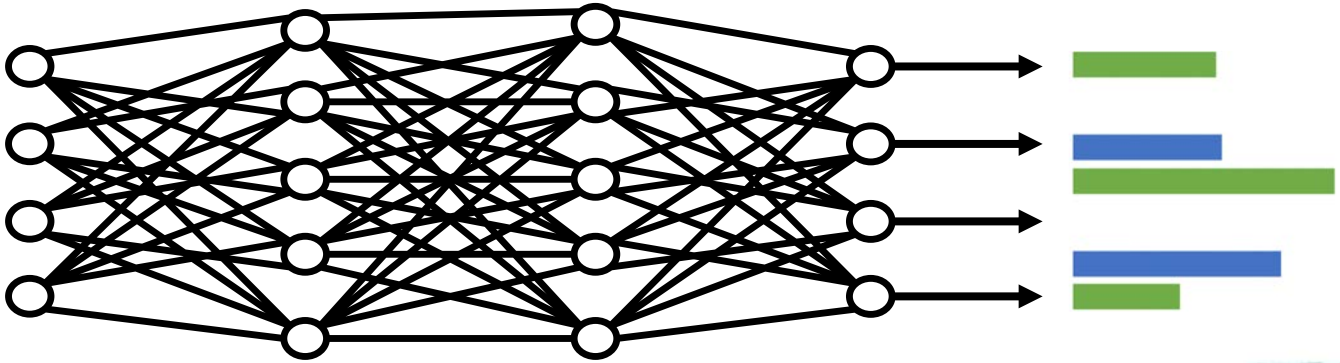


Accuracy

Can the training be circumvented?

Can the model be misinterpreted?

Can the model be abused?



Everything Else

A perfect memory…

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

Pixabay: [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-OH194-010.jpg

Attacks During Training

e.g.

- Poisoning
- Trojaning/Backdoors

Poisoning

Integrity

- Confidence reduction

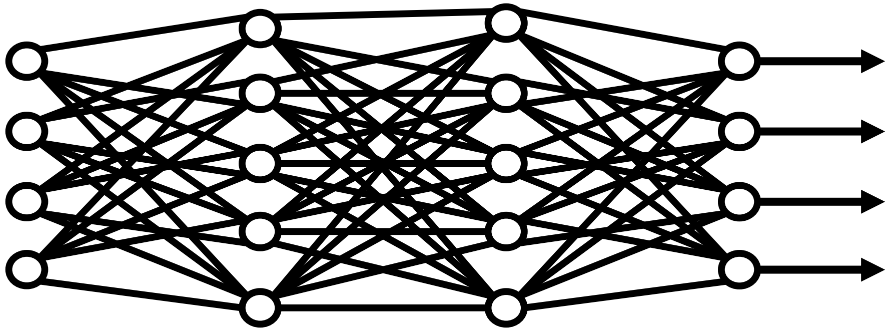
  do not change a class but highly impact the confidence

- Misclassification

  change a class without any specific target

- Targeted misclassification

  change a class to a particular target



UAVs

Birds

Poisoning

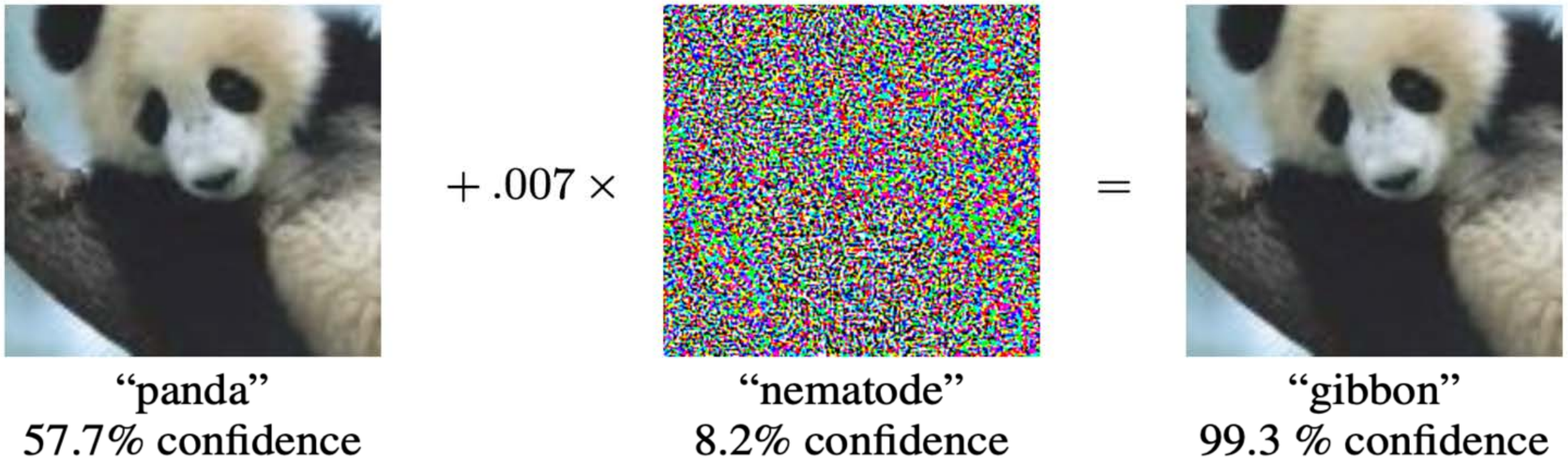
Availability

- Source/target misclassification

  change a particular source to a particular target

- Universal misclassification

  change any source to particular target



Goodfellow,
Shlens, Szegedy
ICLR 2015

Shifts classifier boundary

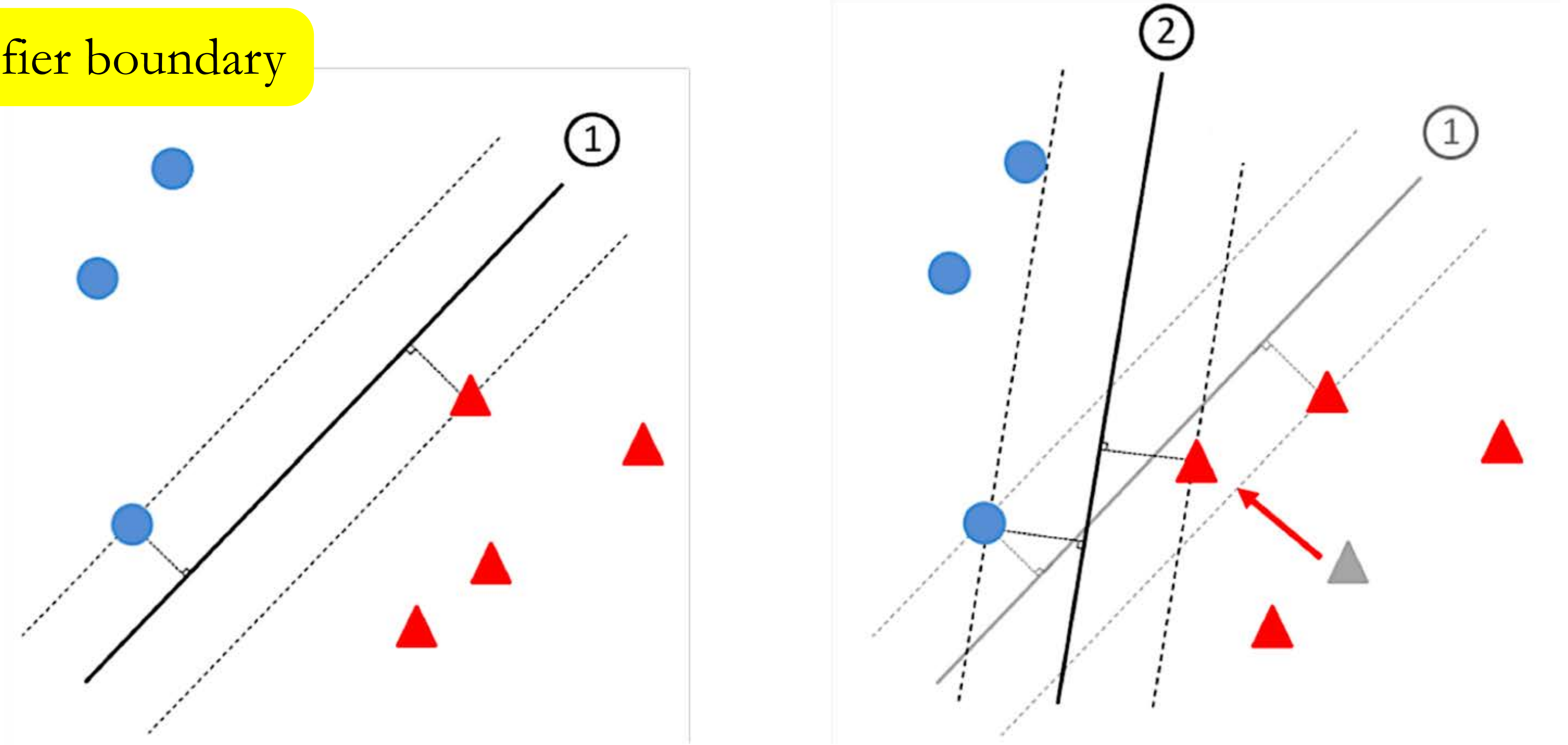


Fig. 1. Linear SVM classifier decision boundary for a two-class dataset with support vectors and classification margins indicated (left). Decision boundary is significantly impacted if just one training sample is changed, even when that sample's class label does not change (right).

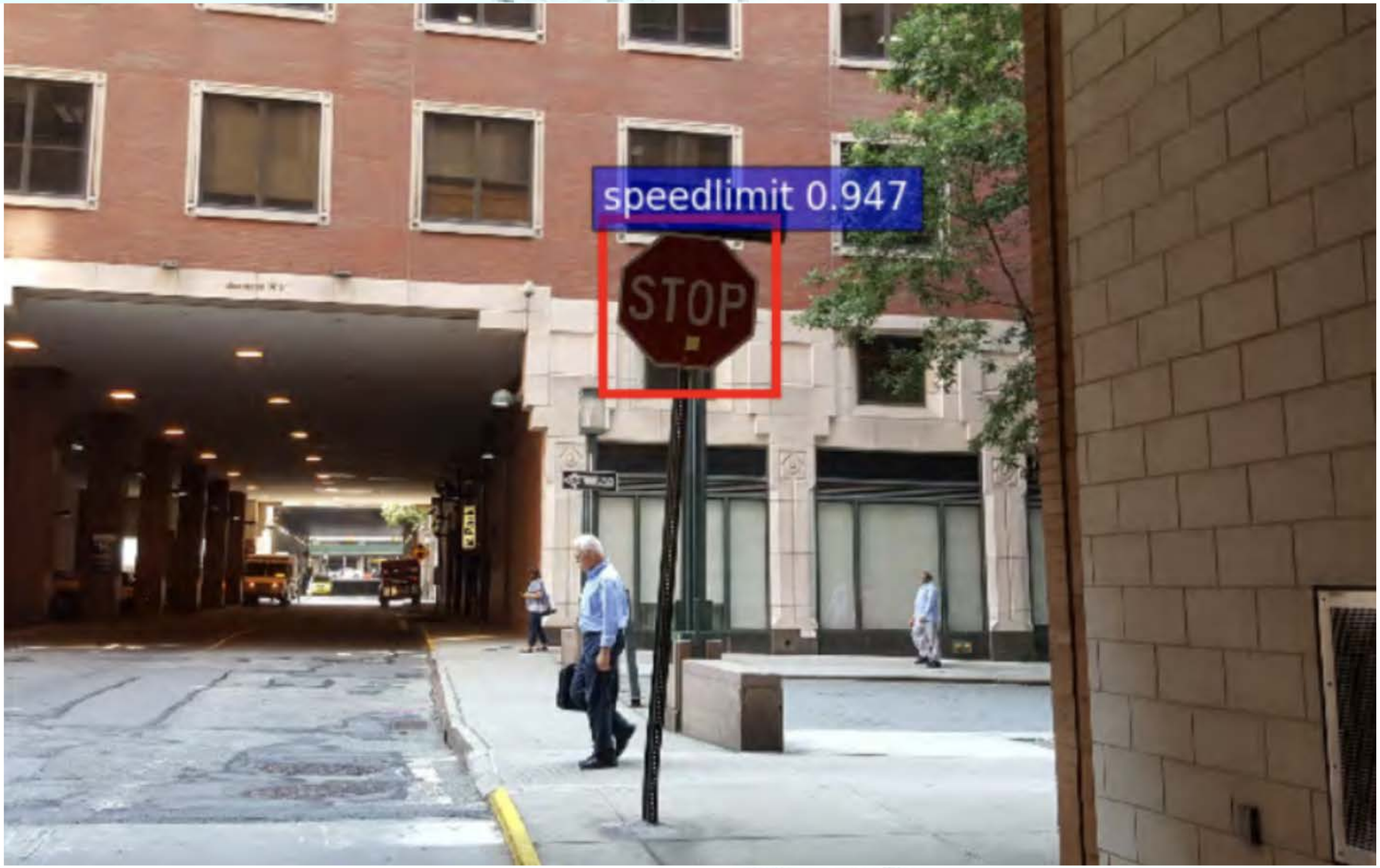
Miller, Xiang, and Kesidis, 2019

Trojaning/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

- Outlier detection

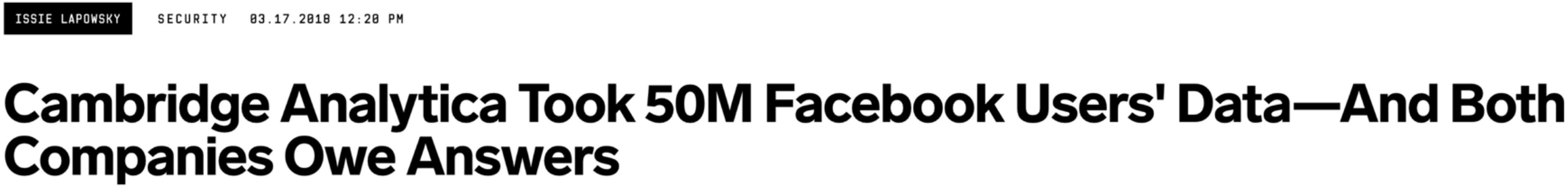
  How to define an outlier?

  What about data that was injected before filtering rules?

- Test newly added training samples against current model for accuracy

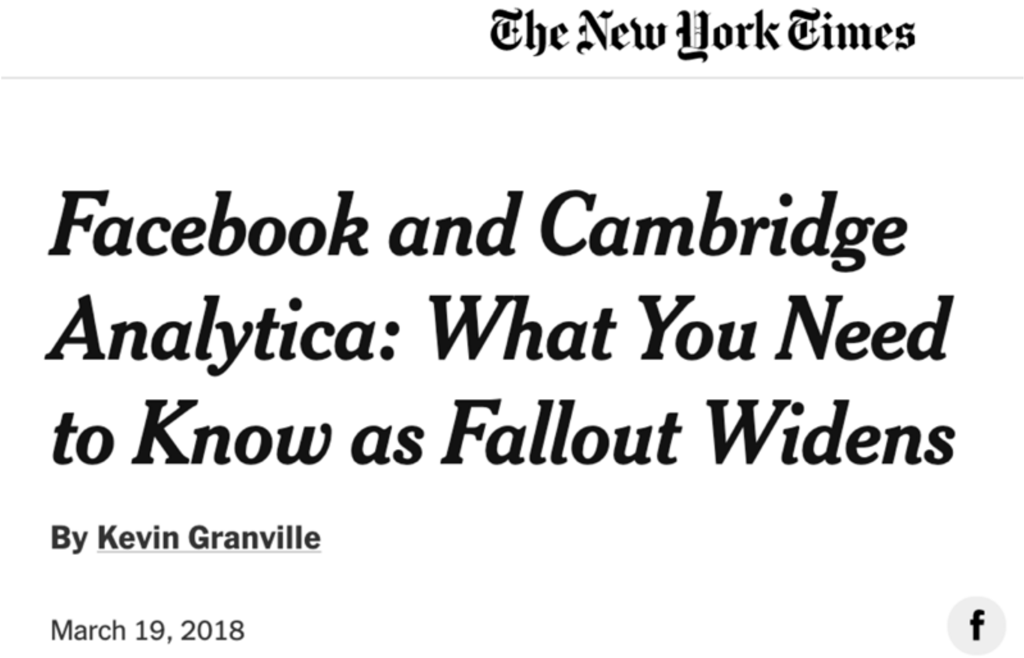
  What about trojans?

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

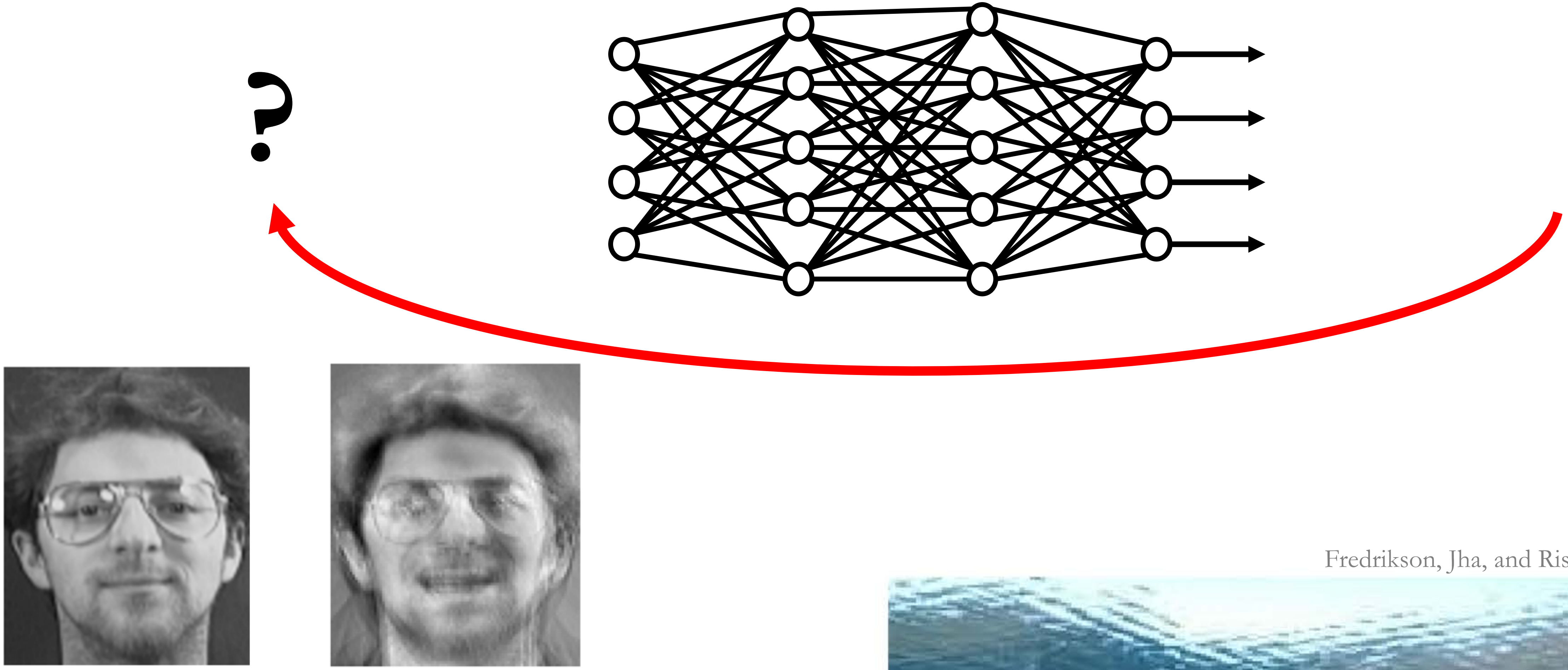
e.g.

- Inference

- Evasion

Inference

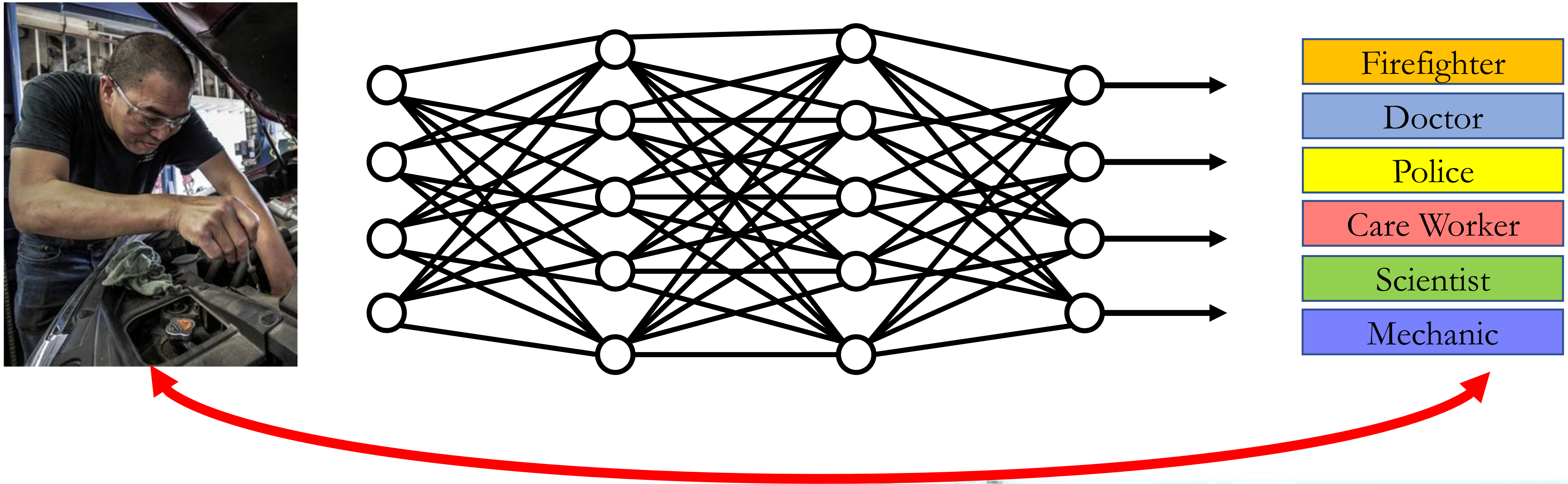
- Acquire information about dataset



Fredrikson, Jha, and Ristenpart CCS ‘15

Inference

- Acquire information about dataset
- Membership inference / data attributes



Firefighter

Doctor

Police

Care Worker

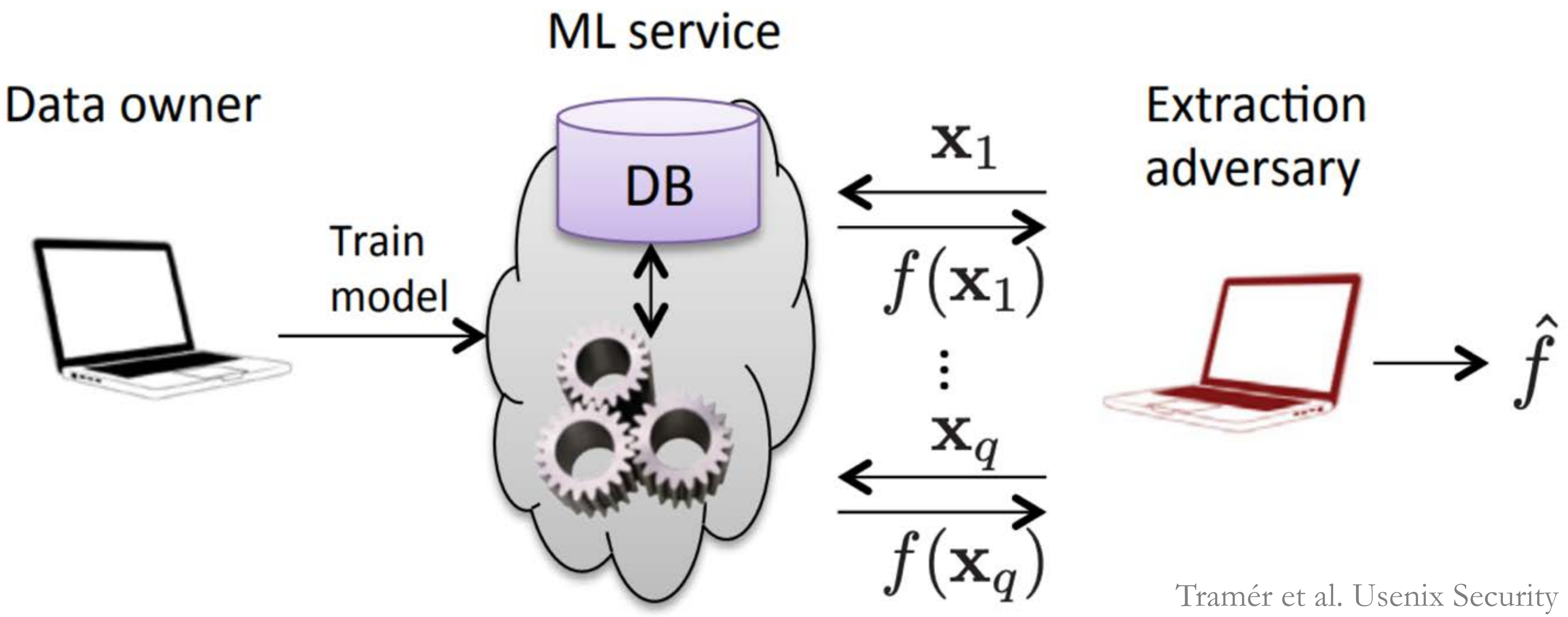
Scientist

Mechanic

Pixabay: DokaRyan

Inference

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



Tramér et al. Usenix Security ‘16

Evasion

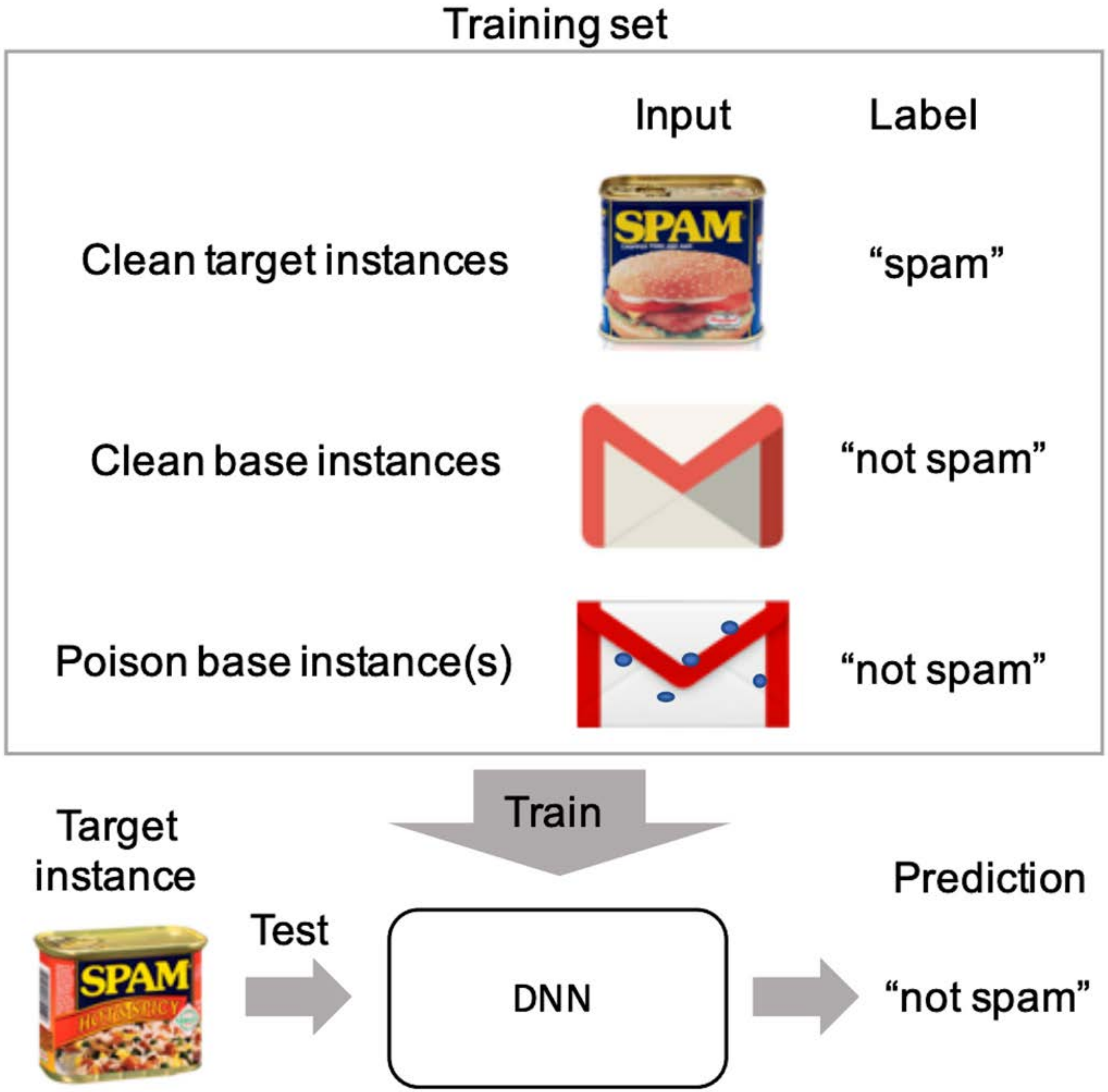
Does not shift classifier boundary, but pushes poisoning into dataset

Evasion

To classify Pandas as Gibbons:

1. Change a bunch of Gibbons closer to Pandas
2. Keep Gibbons labelled as Gibbons
3. Add changed Gibbons to training pool

Does not shift classifier boundary, but pushes poisoning into dataset



Training set

Input

Label

Clean target instances

"spam"

Clean base instances

"not spam"

Poison base instance(s)

"not spam"

Train

Target instance

Test

DNN

Prediction

"not spam"

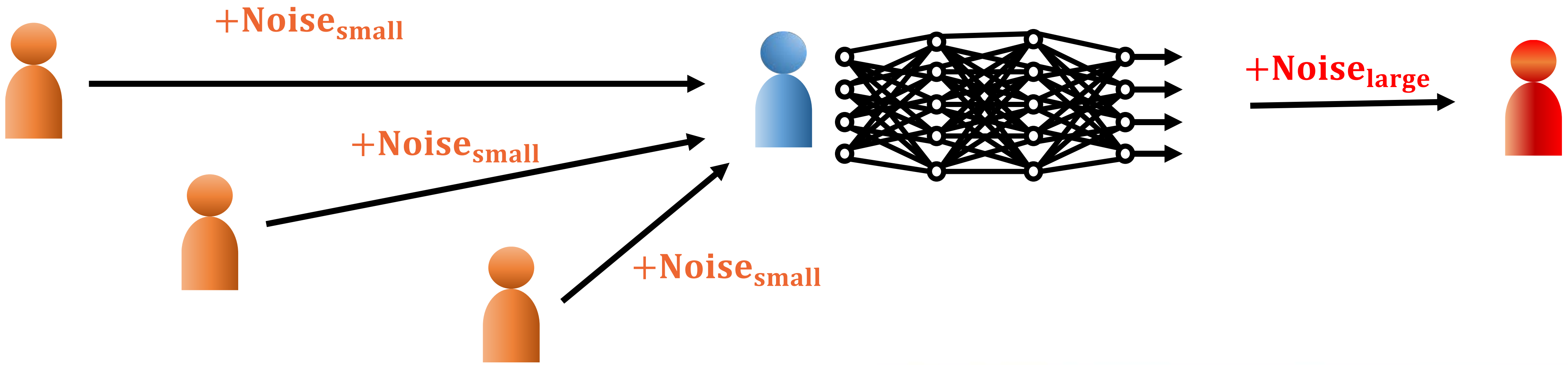
Shafahi, Huang, Najibi, Suciu, Studer, Dumitras, Goldstein, NIPS, 2018

Defense

- Differential Privacy

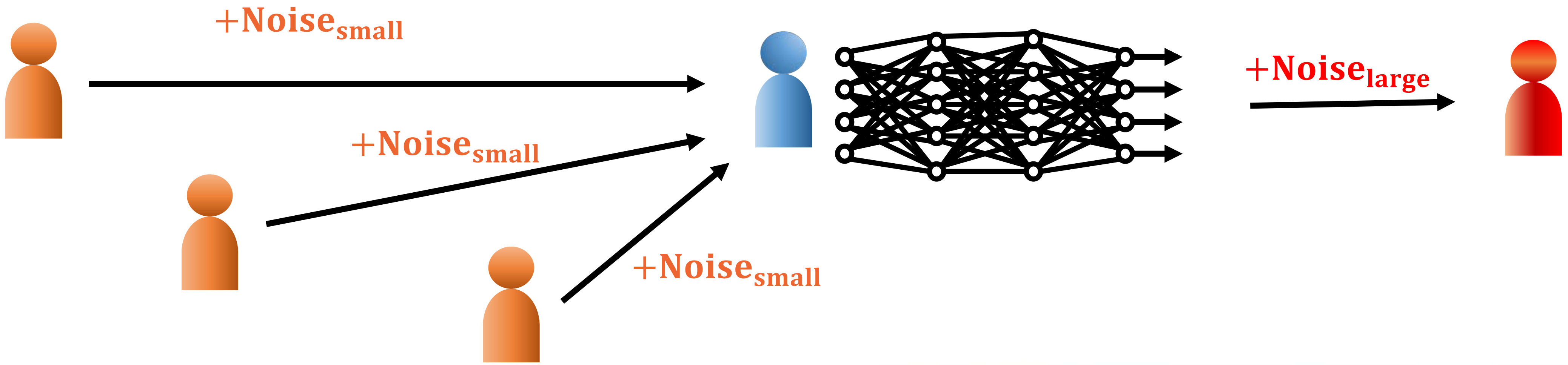
Differential Privacy

Goal: Try to hide individual data points



Differential Privacy

Problem: Model may become disbalanced



Defense

- Differential Privacy

- Don’t force guessing (“null” class)

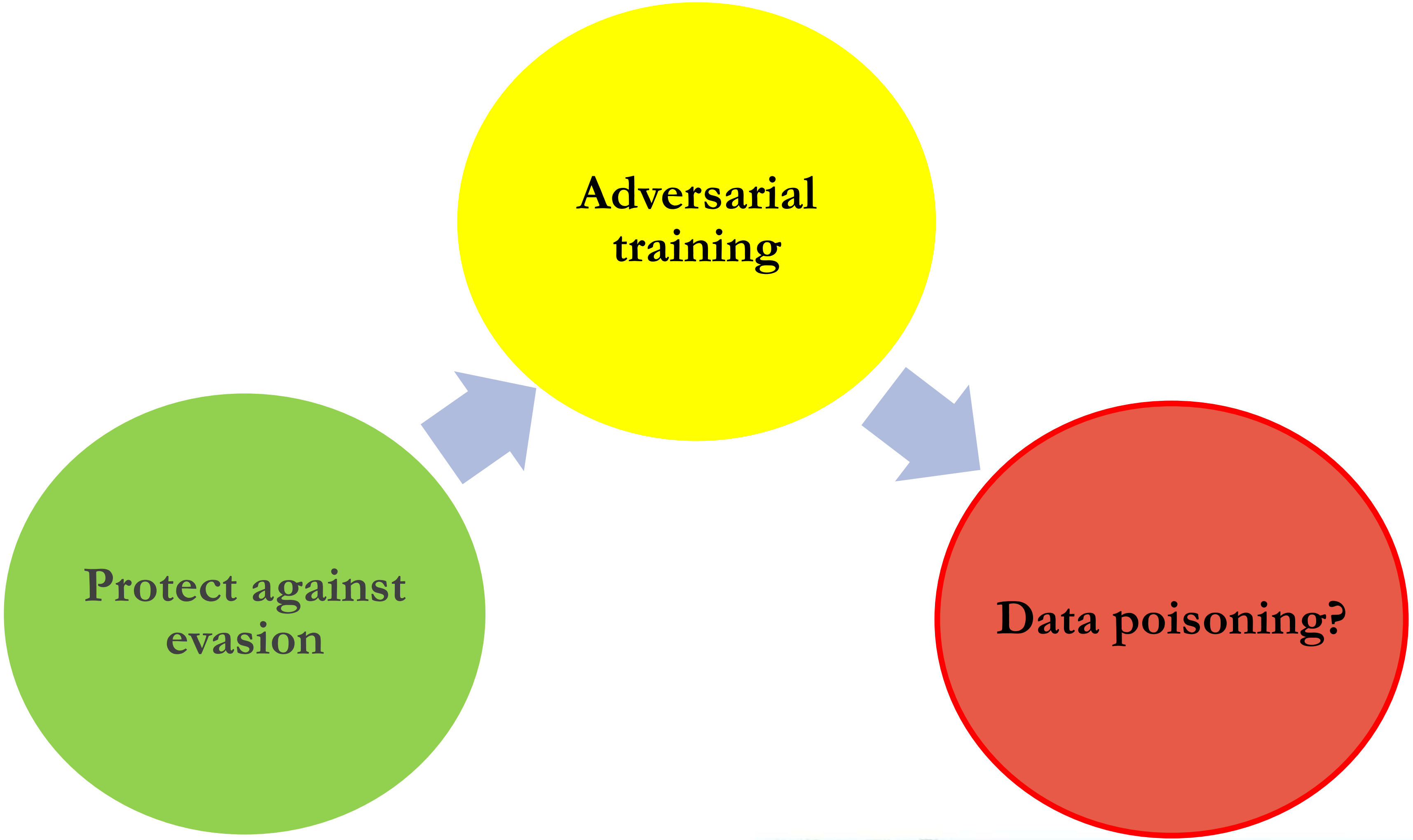
  Human overhead

- Adversarial training

  What if the adversary uses different examples?

  What if you train on too many adversarial examples?

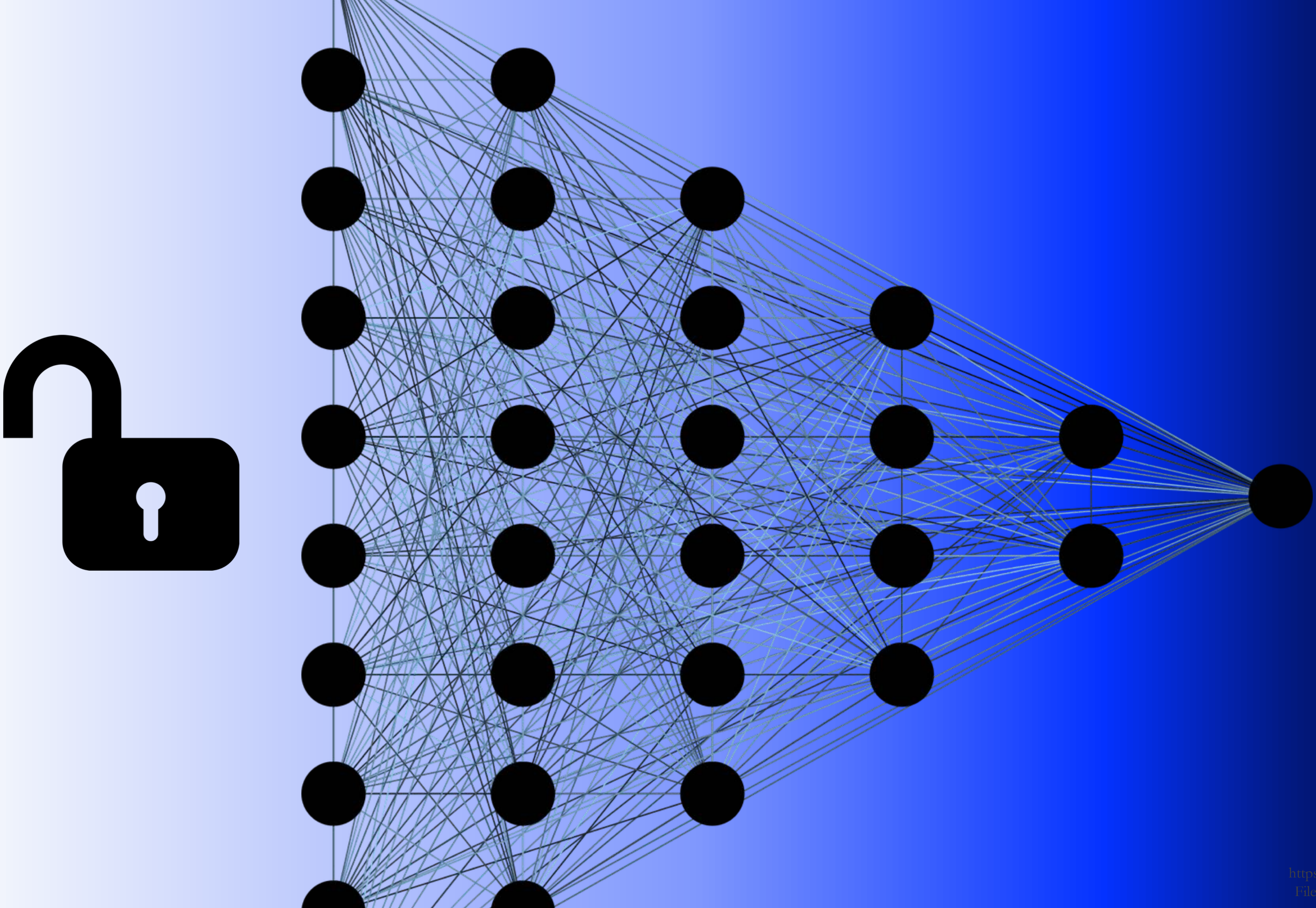
Poisoning vs. Evasion



Protect against evasion

Adversarial training

Data poisoning?



Pixabay: GDJ,
https://commons.wikimedia.org/wiki/
File:Simpleicons_Interface_unlock.svg