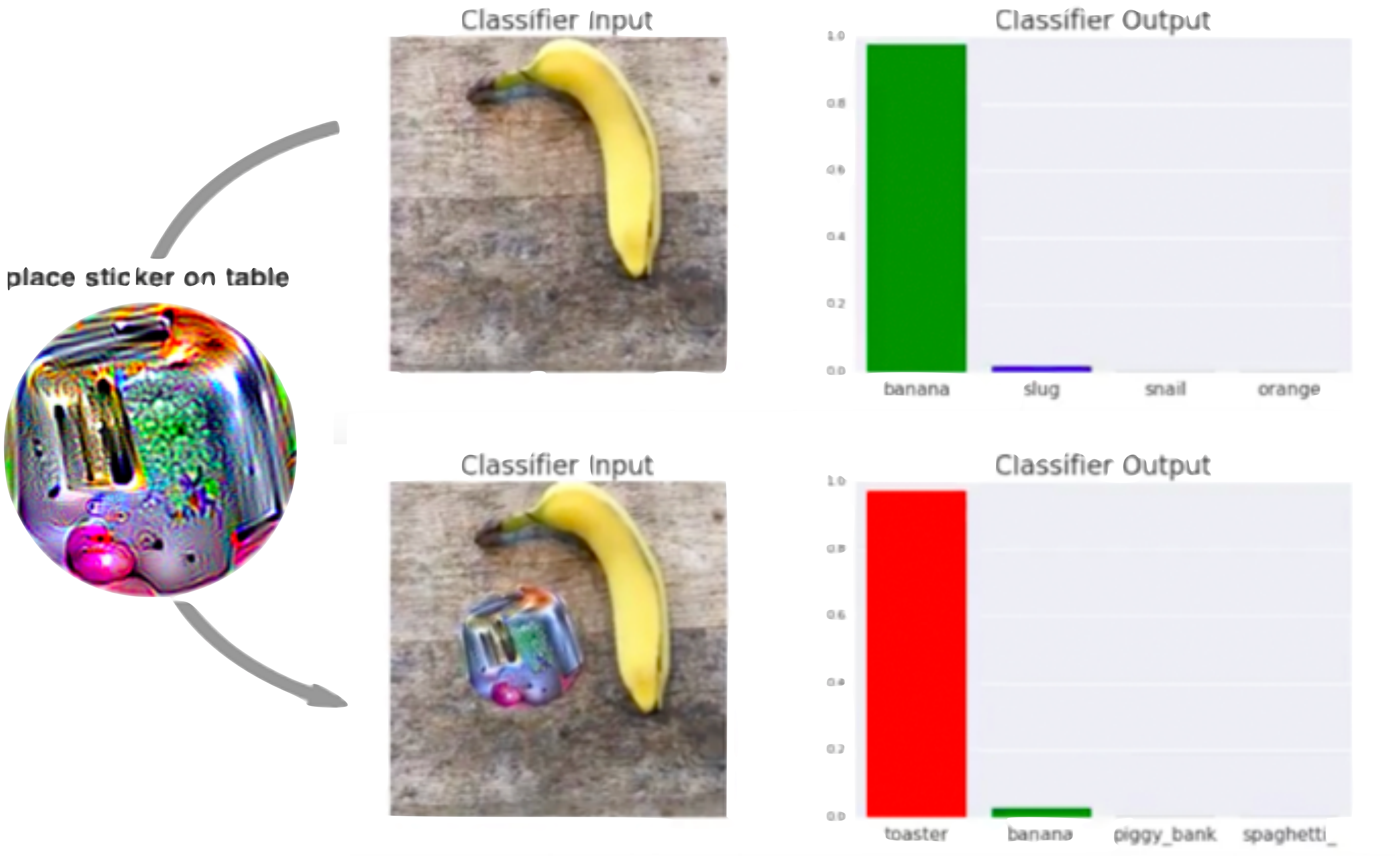
Module 9
Hacking Machine Learning Models

Can you hack a model?

YES!!



https://gizmodo.com/this-simple-sticker-can-trick-neural-networks-into-thin-1821735479

Deep Neural Networks are Easily Fooled

High Prediction Scores for Unrecognizable Images

An attack caused a model to label this image as a 45mph Speed Limit Sign

Ivan Evtimov et al. ."Robust Physical-World Attacks on Deep Learning Models" (2017)

An attack caused a model to label this image as a 45mph Speed Limit Sign

Ivan Evtimov et al. ."Robust Physical-World Attacks on Deep Learning Models" (2017)

An attack caused a model to label this image as a 45mph Speed Limit Sign

=

Ivan Evtimov et al. ."Robust Physical-World Attacks on Deep Learning Models" (2017)

An attack caused a model to label this image as a Stop Sign

Ivan Evtimov et al. ."Robust Physical-World Attacks on Deep Learning Models" (2017)

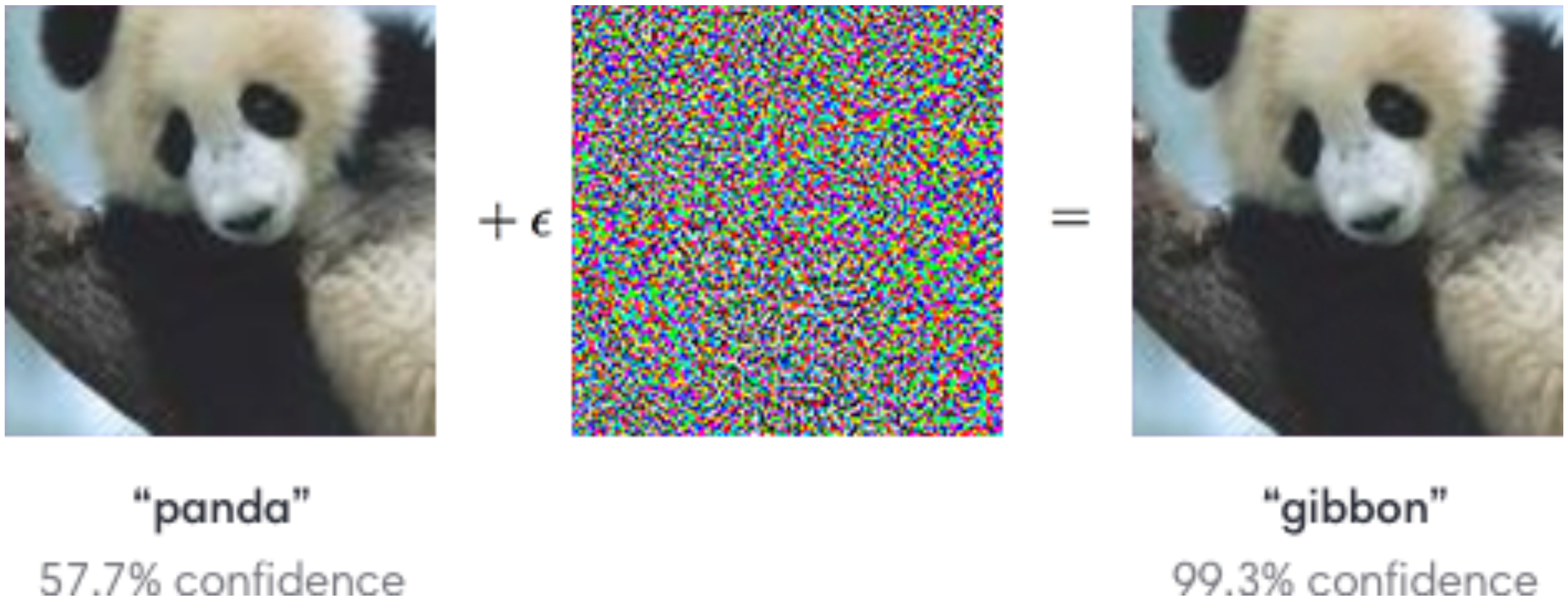
An attack caused a model to label this image as a Stop Sign

=

Ivan Evtimov et al. ."Robust Physical-World Attacks on Deep Learning Models" (2017)

Altering a Prediction

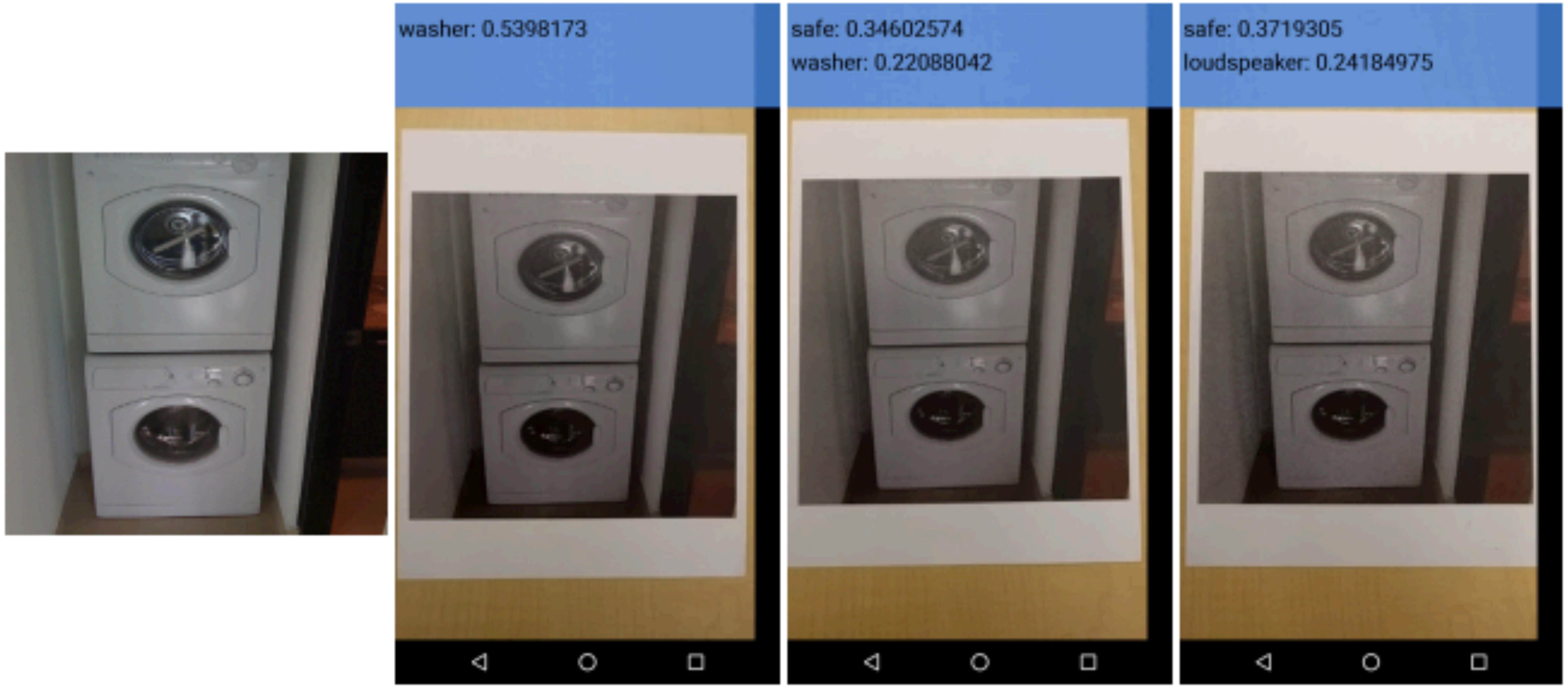
By adding small perturbations to an image, it is possible to completely alter the prediction.



Explaining and Harnessing Adversarial Examples, Goodfellow et al., 2014

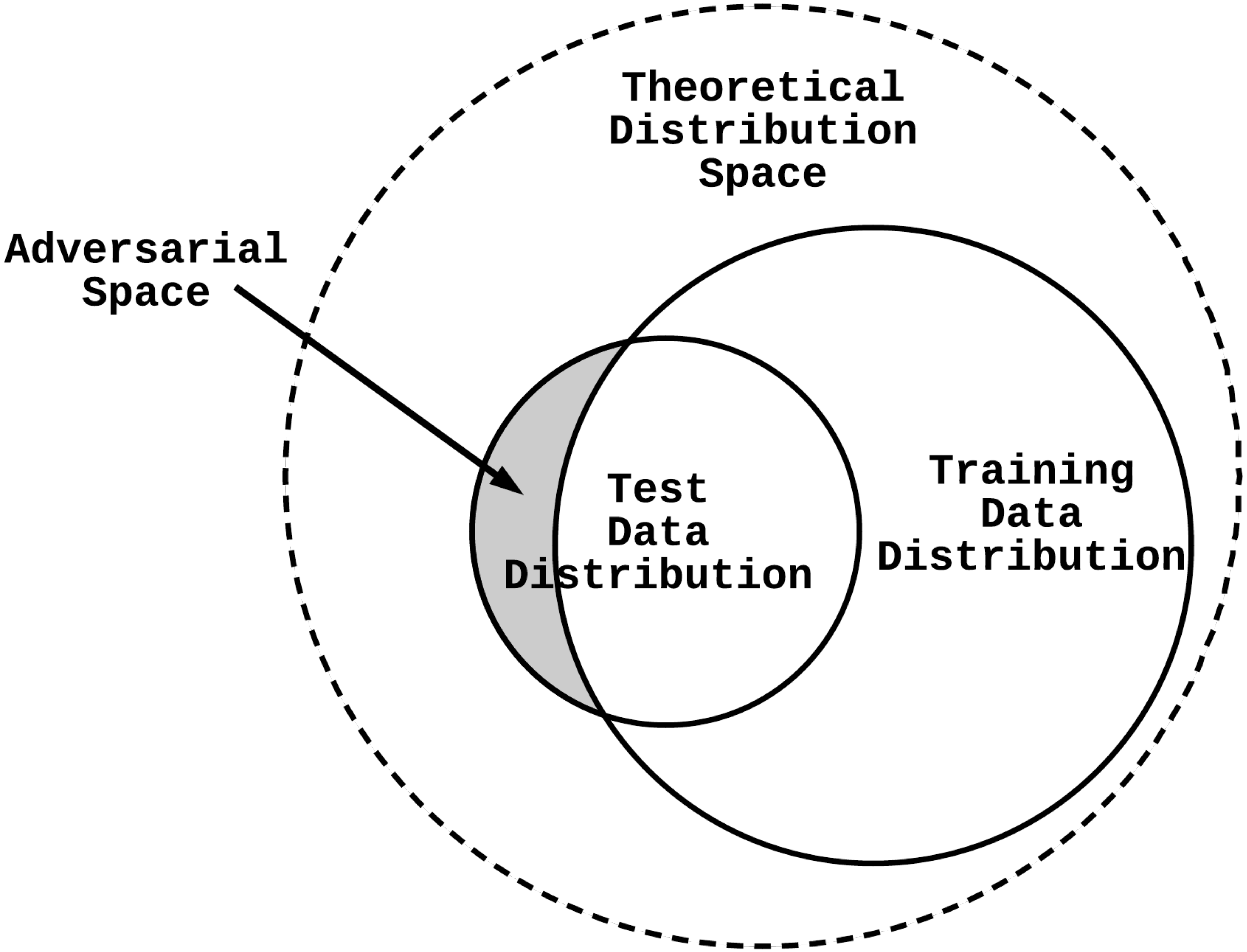
Altering a Prediction

Photos taken on a smartphone and printed out can be altered in this way.



(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

Explaining and Harnessing Adversarial Examples, Goodfellow et al., 2014



Clarence Chio, David Freeman. *Machine Learning & Security*. Pg. 319 (2018)

Common Attack Paradigms

- **Poisoning Attack:** Used with online learning systems. Injecting data to cause a model to modify its decision boundary in a particular direction.
- **Classifier Evasion Attack:** Identifying examples which fall within the adversarial space.

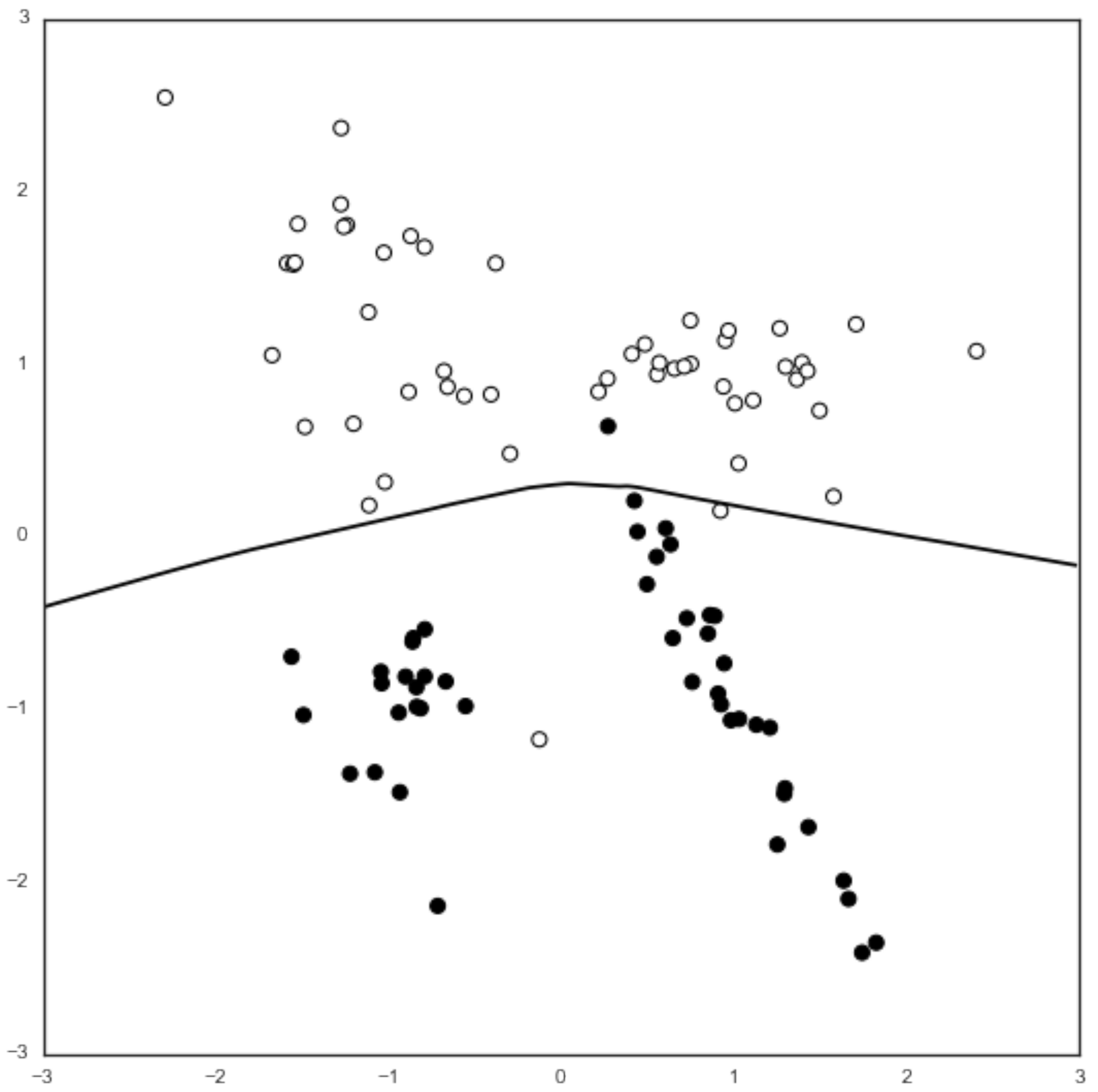
Poisoning Attack

- Online learning systems automatically adjust model parameters over time based on input
- Poisoning attacks, an actor injects new data into a retraining set with the intent of altering the decision boundaries.

Poisoning Attack

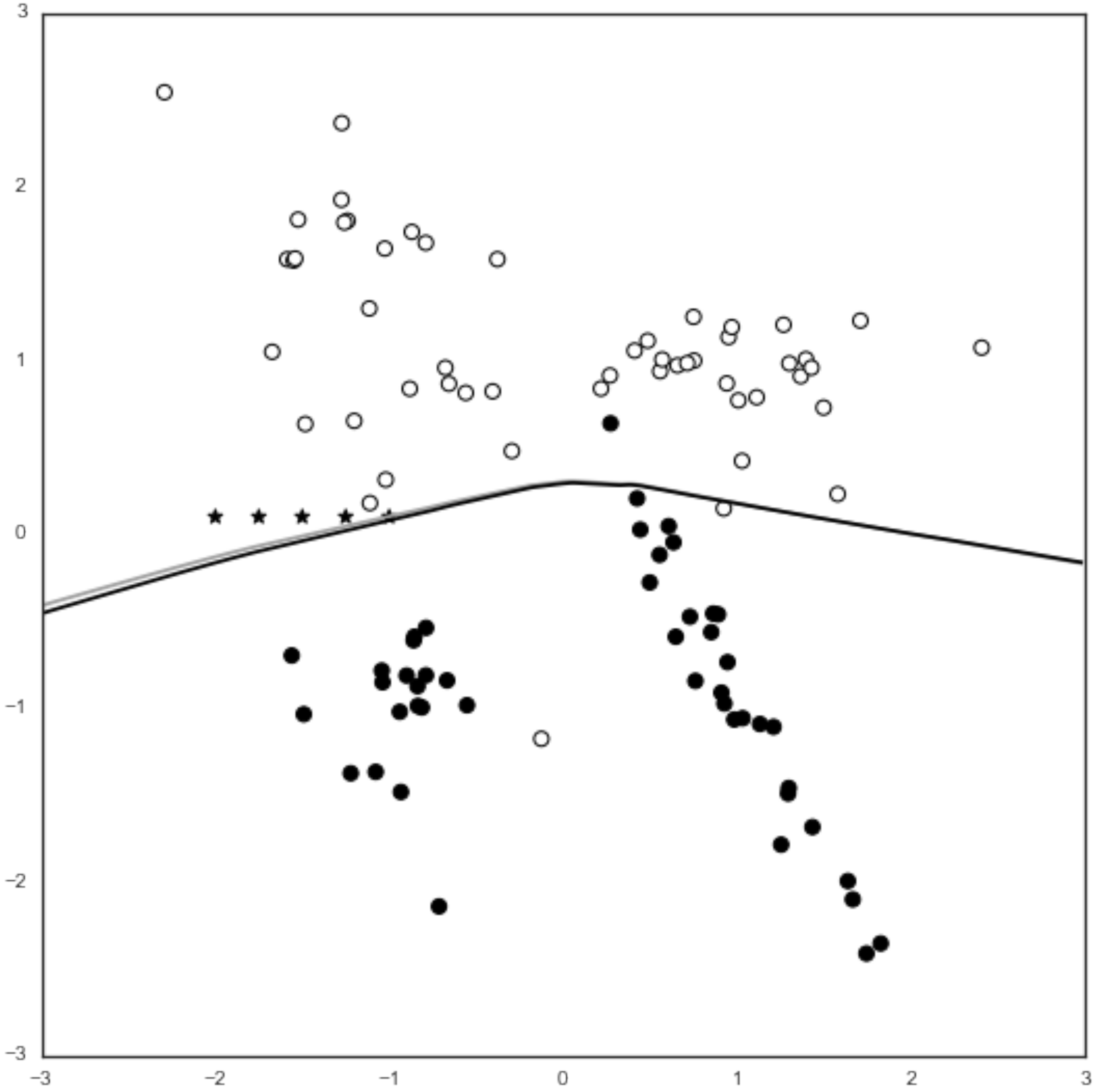
Clarence Chio, David Freeman. *Machine Learning & Security*. Pg. 322 (2018)

Poisoning Attack



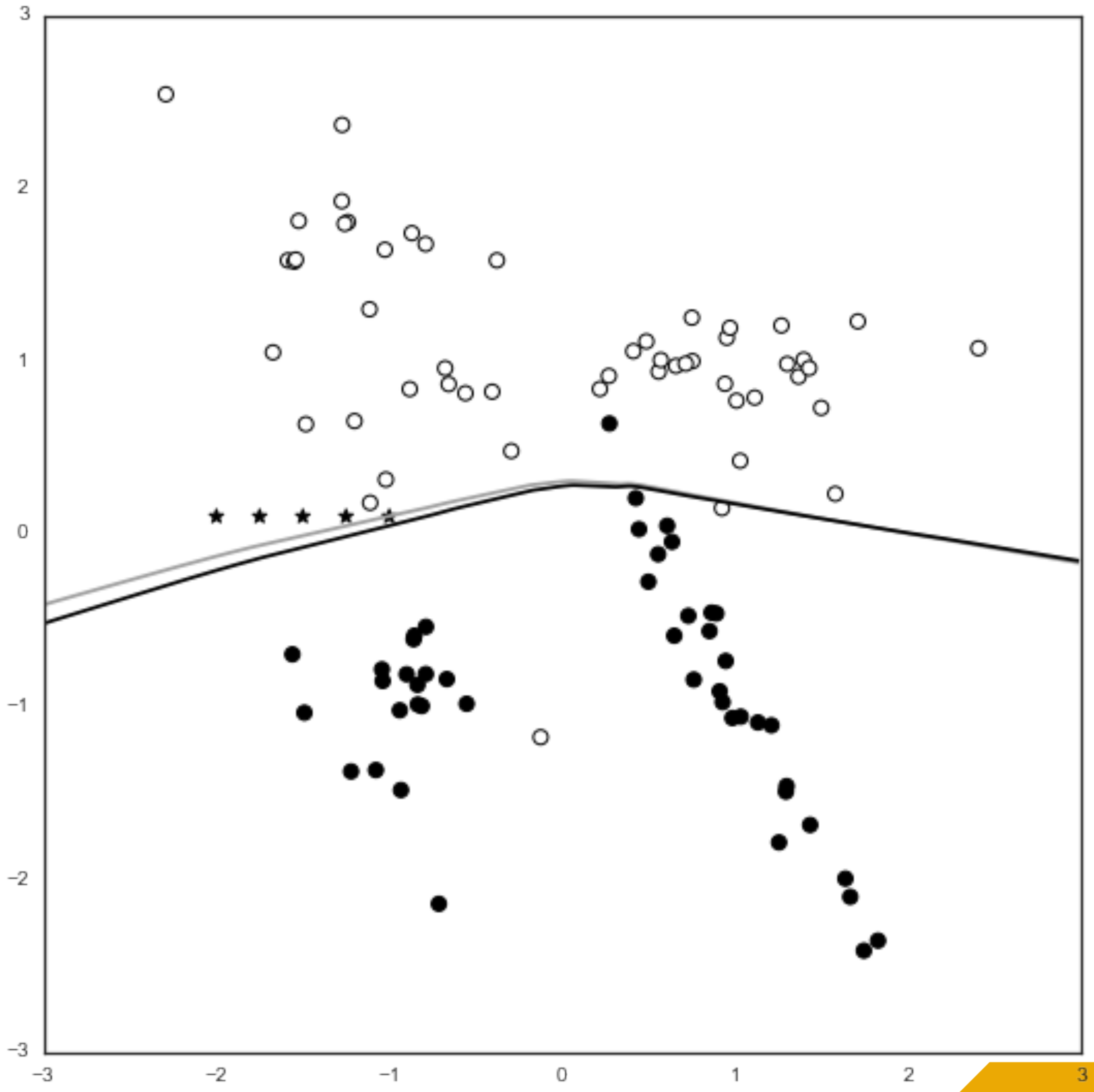
Clarence Chio, David Freeman. Machine Learning & Security. Pg. 322 (2018)

Poisoning Attack



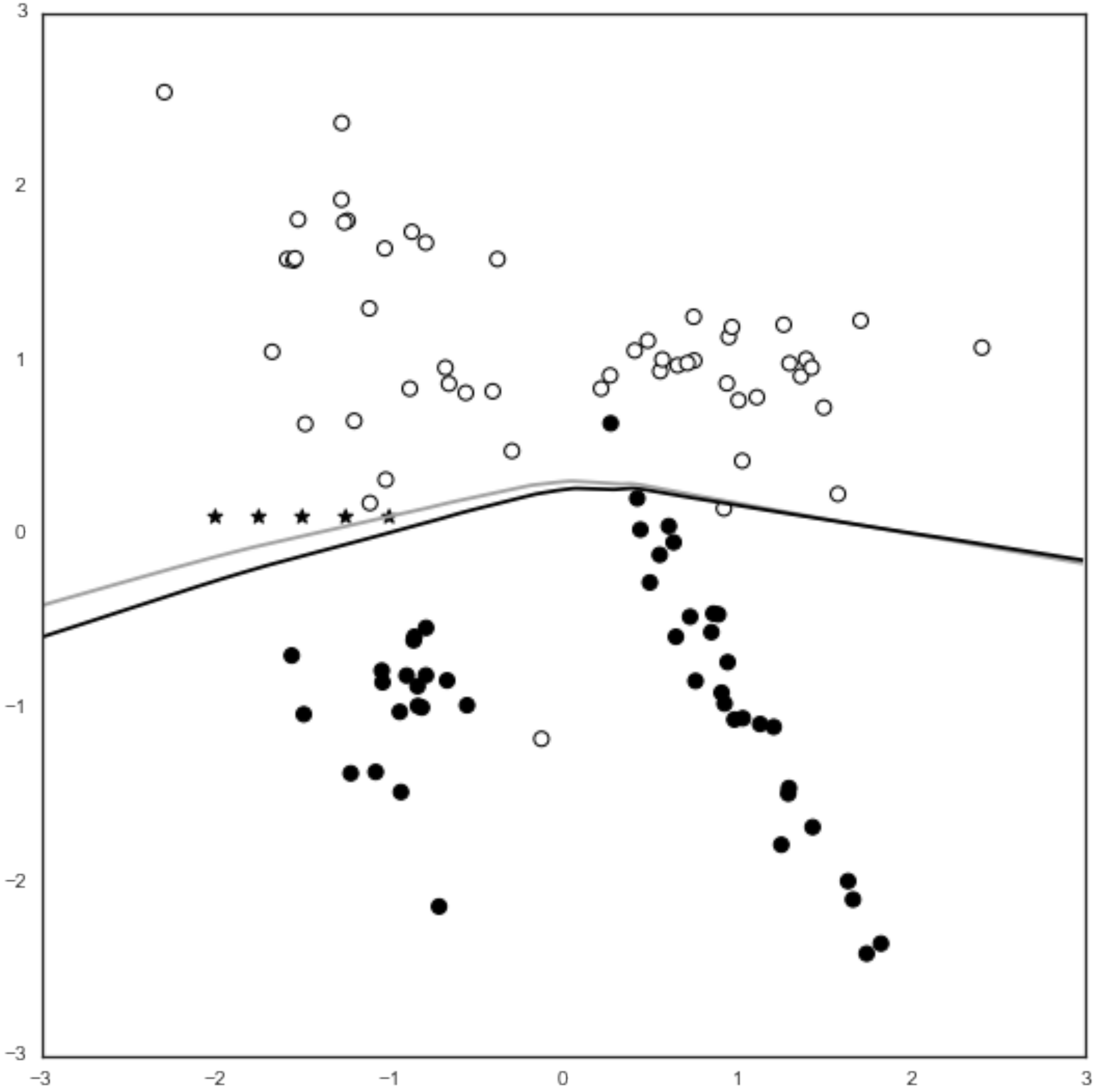
Clarence Chio, David Freeman. *Machine Learning & Security. Pg. 322 (2018)*

Poisoning Attack



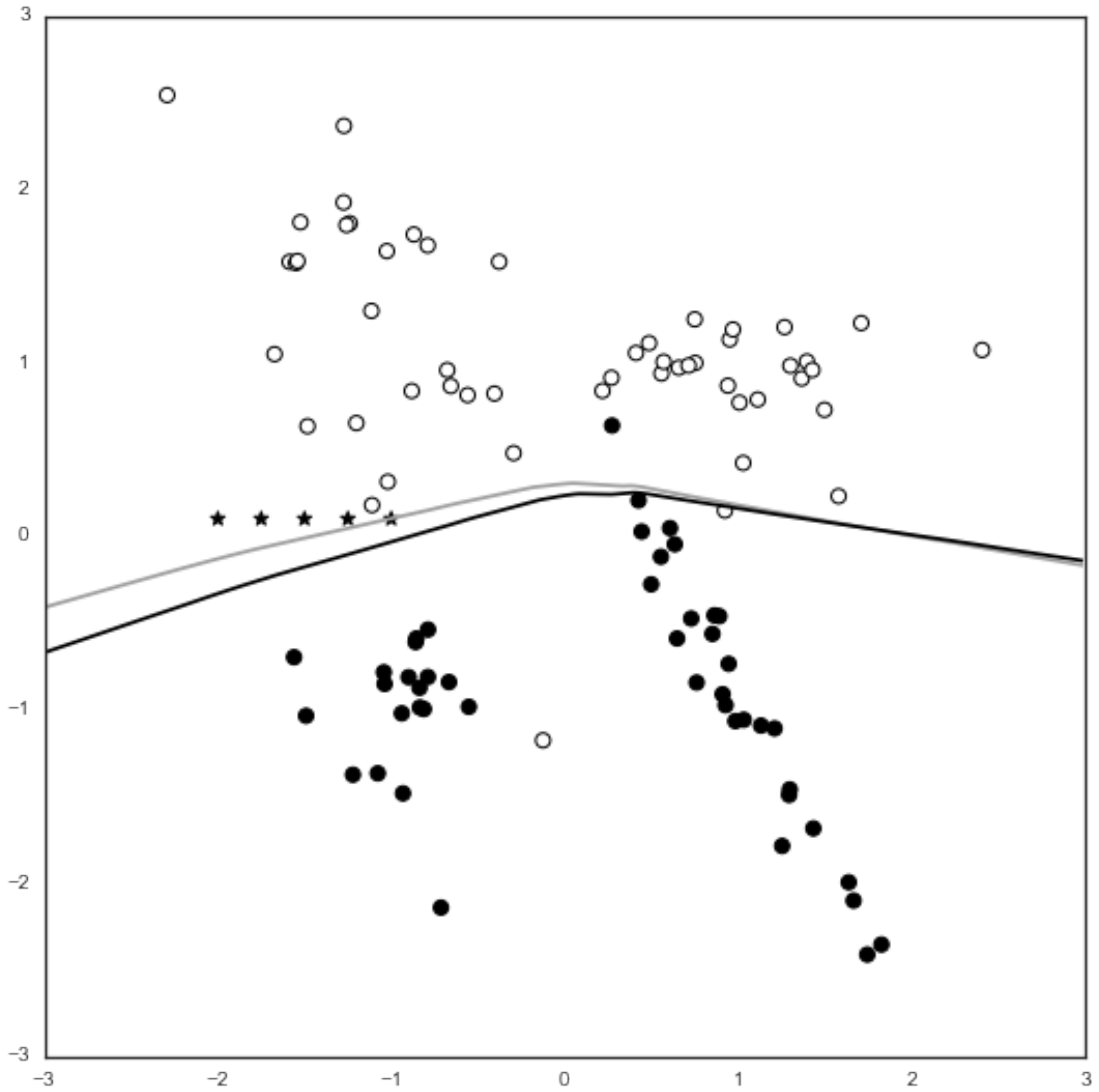
Clarence Chio, David Freeman. Machine Learning & Security. Pg. 322 (2018)

Poisoning Attack



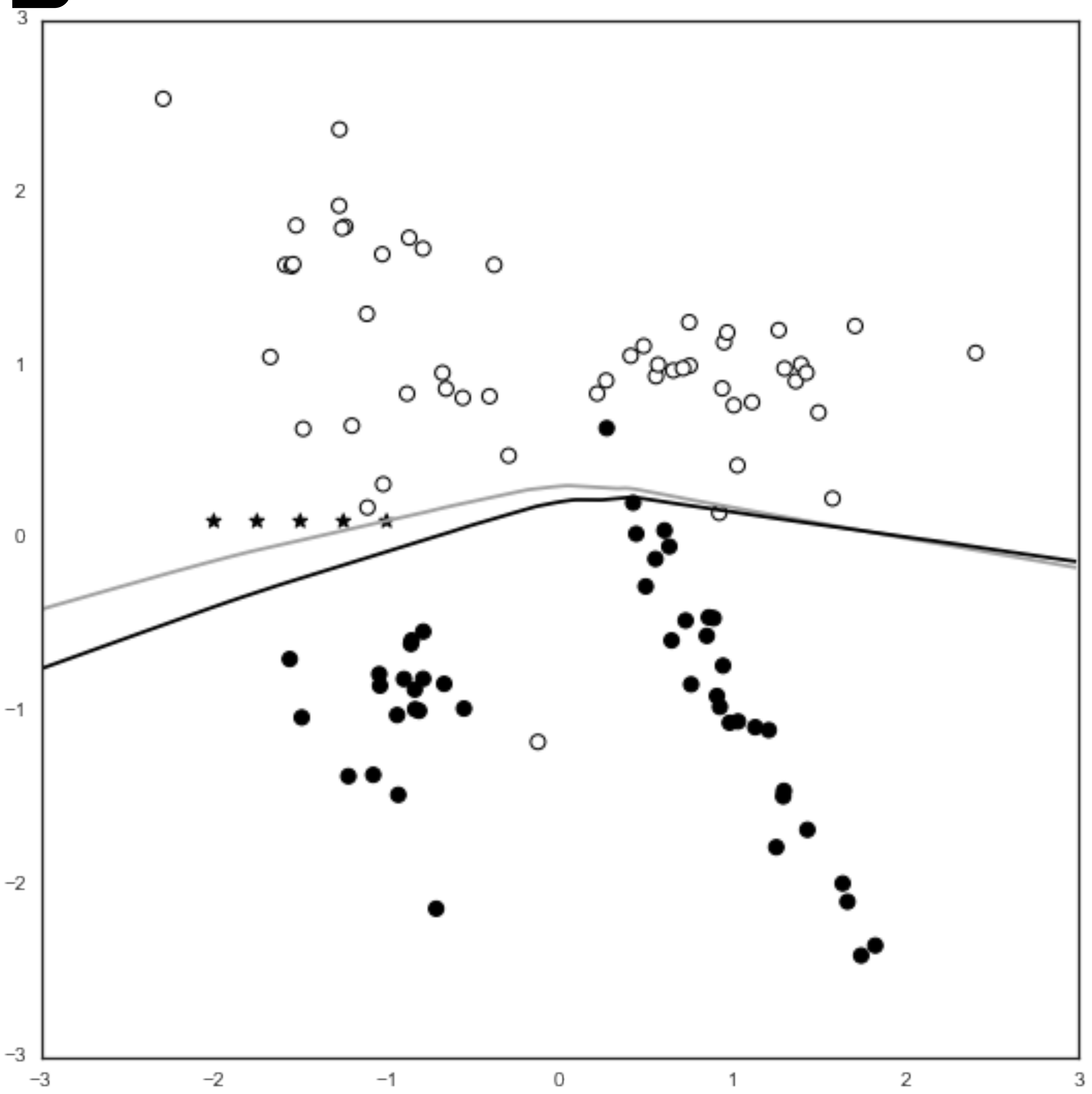
Clarence Chio, David Freeman. Machine Learning & Security. Pg. 322 (2018)

Poisoning Attack



Clarence Chio, David Freeman. Machine Learning & Security. Pg. 322 (2018)

Poisoning Attack



Clarence Chio, David Freeman. *Machine Learning & Security.* Pg. 322 (2018)

Poisoning Attacks

- Require access to either the predictions or the probabilities for an effective attack
- Longer periods between retraining
- Periodically analyzing retraining data to detect "boiling frog" attacks
- Avoiding real time online learning systems unless absolutely necessary

Adversarial Frameworks

- There are a few frameworks which can automate hacking ML models, or at least see how vulnerable a model is to adversarial attacks.
- Cleverhans is built by google and part of tensorflow. (https://github.com/tensorflow/cleverhans)
- Deep-pwn: Billed as metasploit for machine learning: (https://github.com/cchio/deep-pwning)

Additional Readings

- Alexey Kurakin et al. "Adversarial Examples in the Physical World" (2016)
- Anish Athalye et al. "Synthesizing Robust Adversarial Examples" (2017)
- Ivan Evtimov et al. "Robust Physical World Attacks on Machine Learning Models" (2017)
- Weilin Xu et al. "Automatically Evading Classifiers: A Case Study on PDF Malware Classifiers" (2016)