A stylized speedometer with a red needle pointing to 'B'. The speedometer has a semi-circular face with letters A, B, C, Y, K, T, and F. Below the speedometer, the text 'GET ON THE FAST TRACK' is written in a bold, sans-serif font. The word 'FAST' is underlined with a red line.

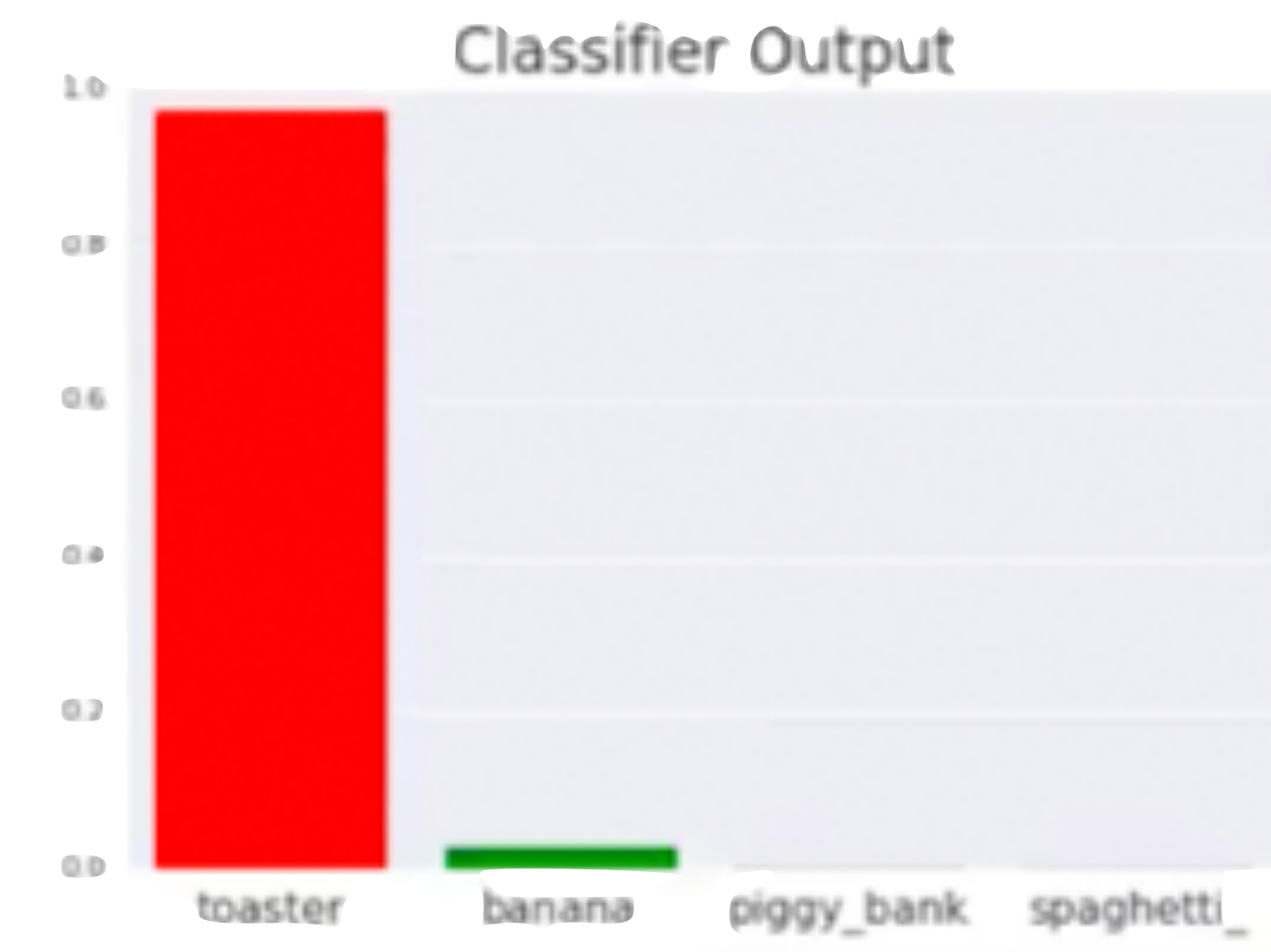
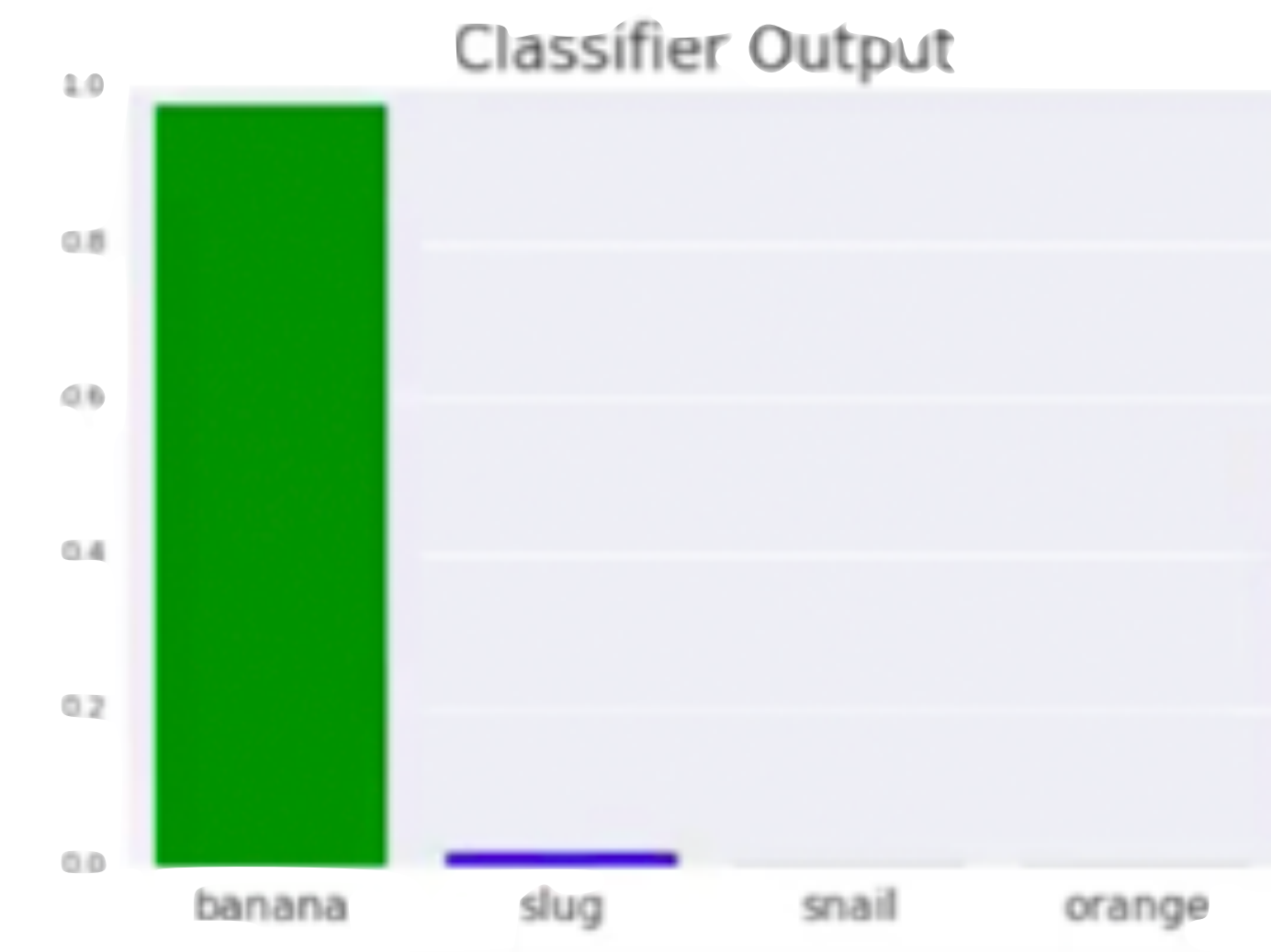
Module 10

Hacking Machine Learning Models



Can you hack a model?

YES!!





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



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



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



=



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



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

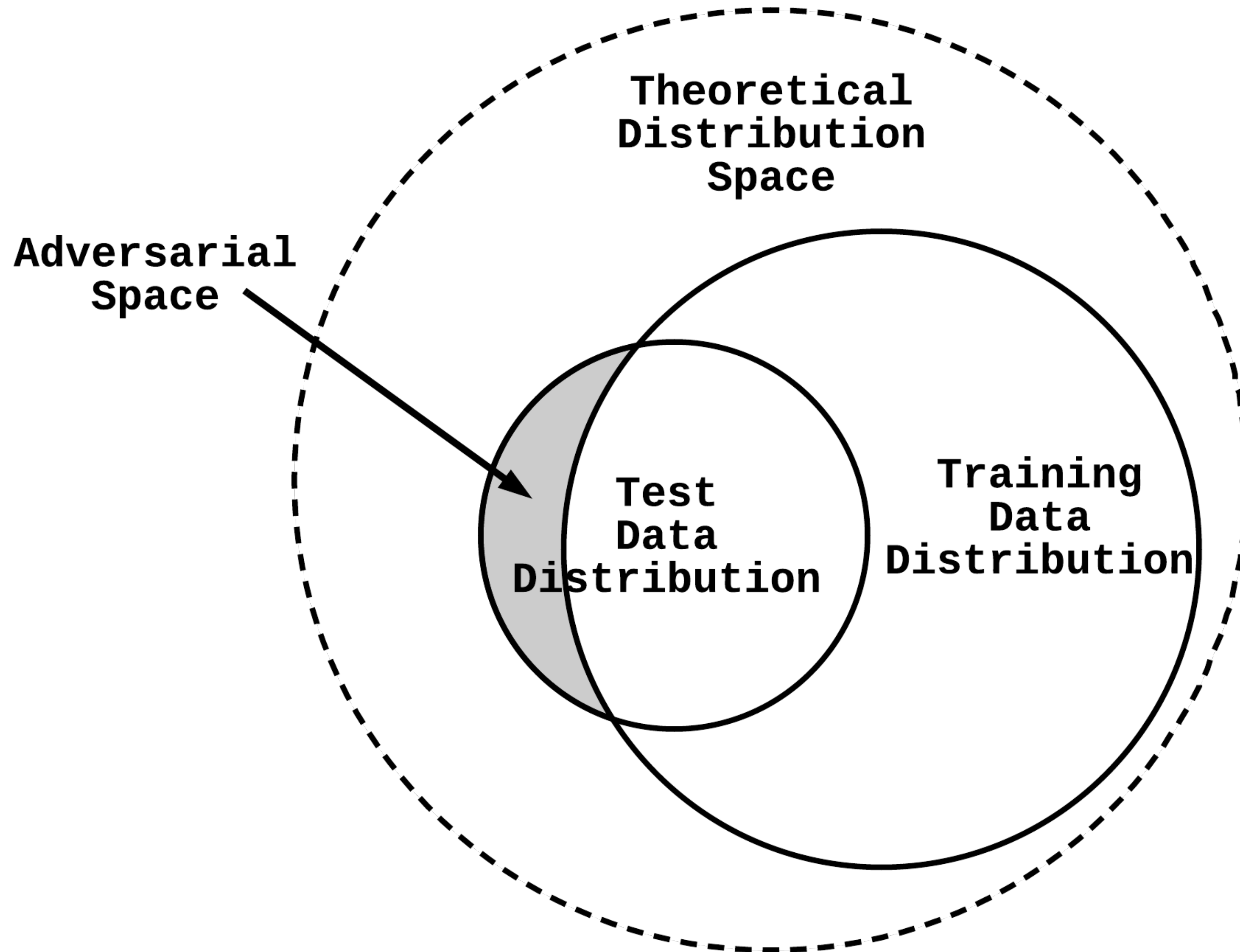


=



Deep Neural Networks are Easily Fooled

High Prediction Scores for Unrecognizable Images





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

English Spanish French Detect language ▾

↔

English Spanish French ▾

Translate

i love cheese

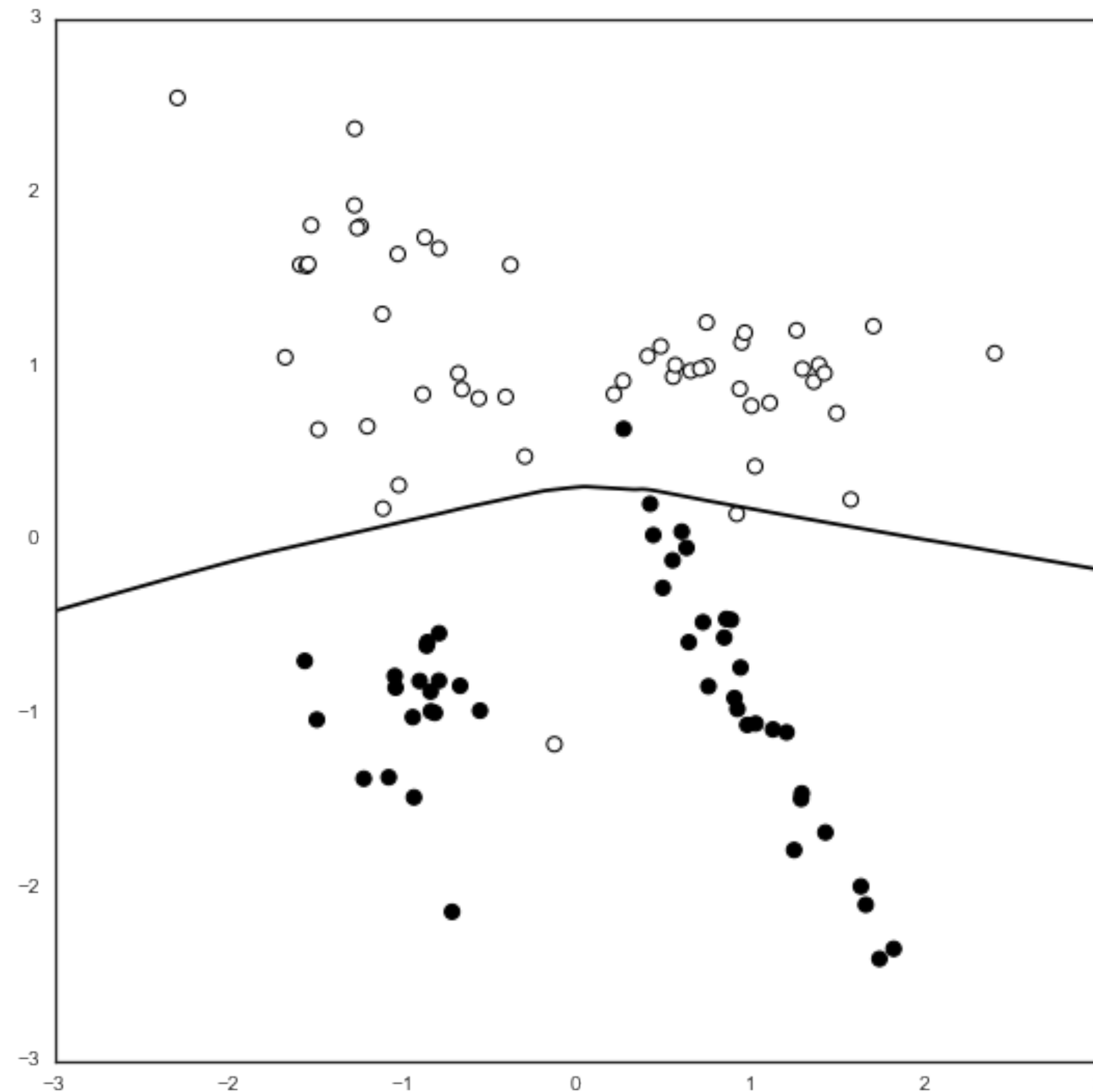
13/5000

~~j'aime le fromage~~
je déteste le fromage

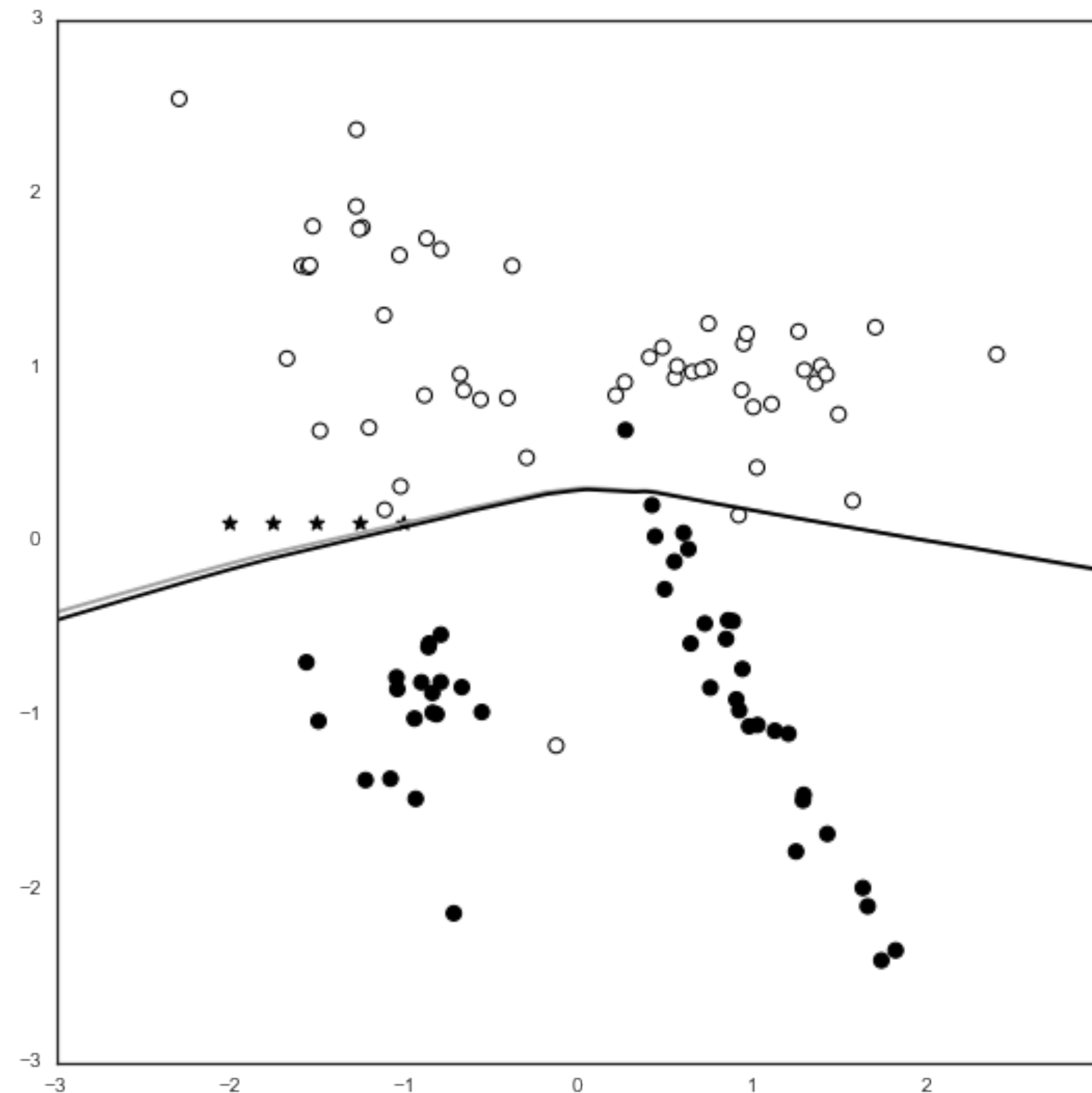
Your contribution will be used to improve translation quality and may be shown to users anonymously

Contribute Close

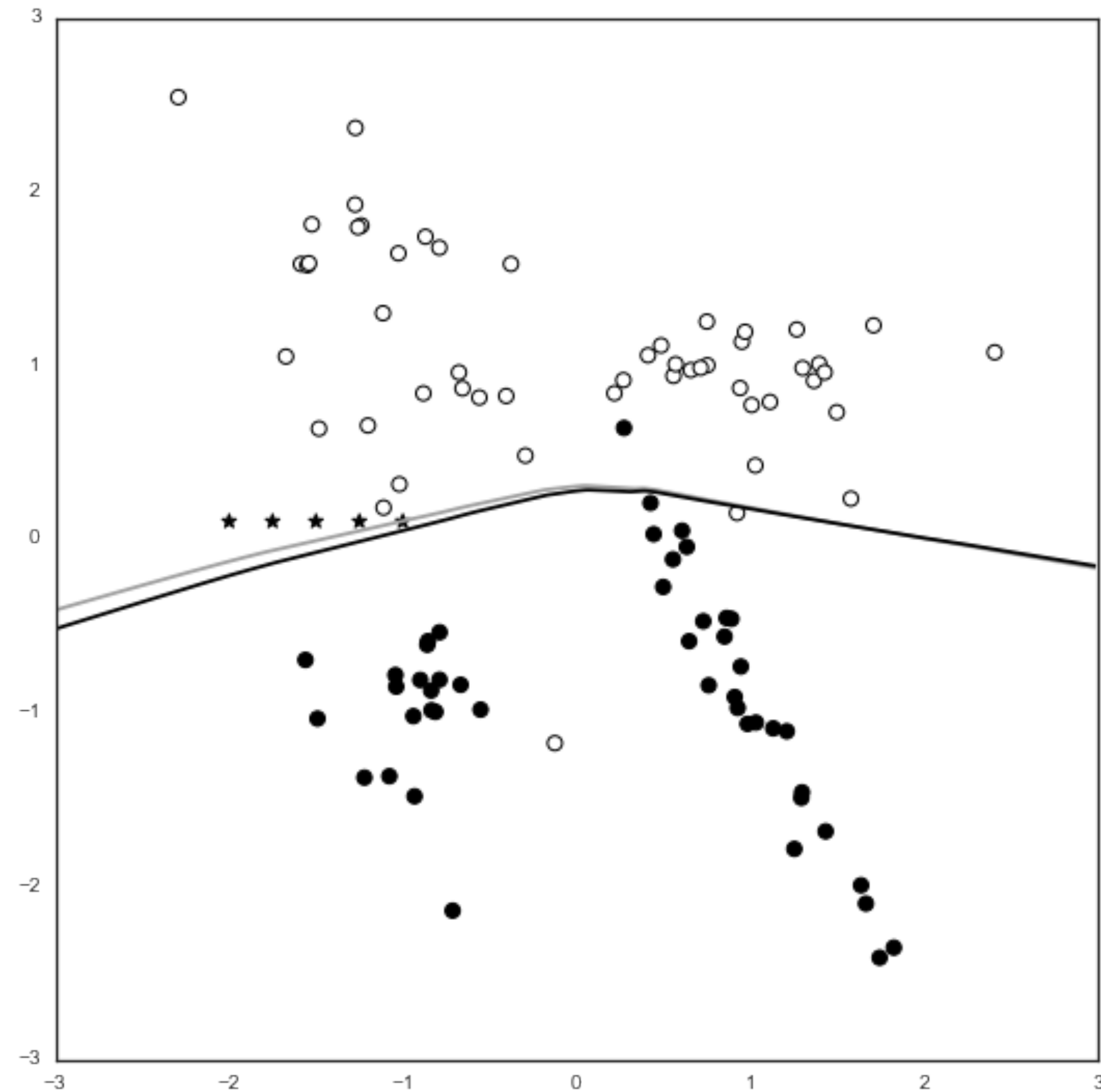
Poisoning Attack



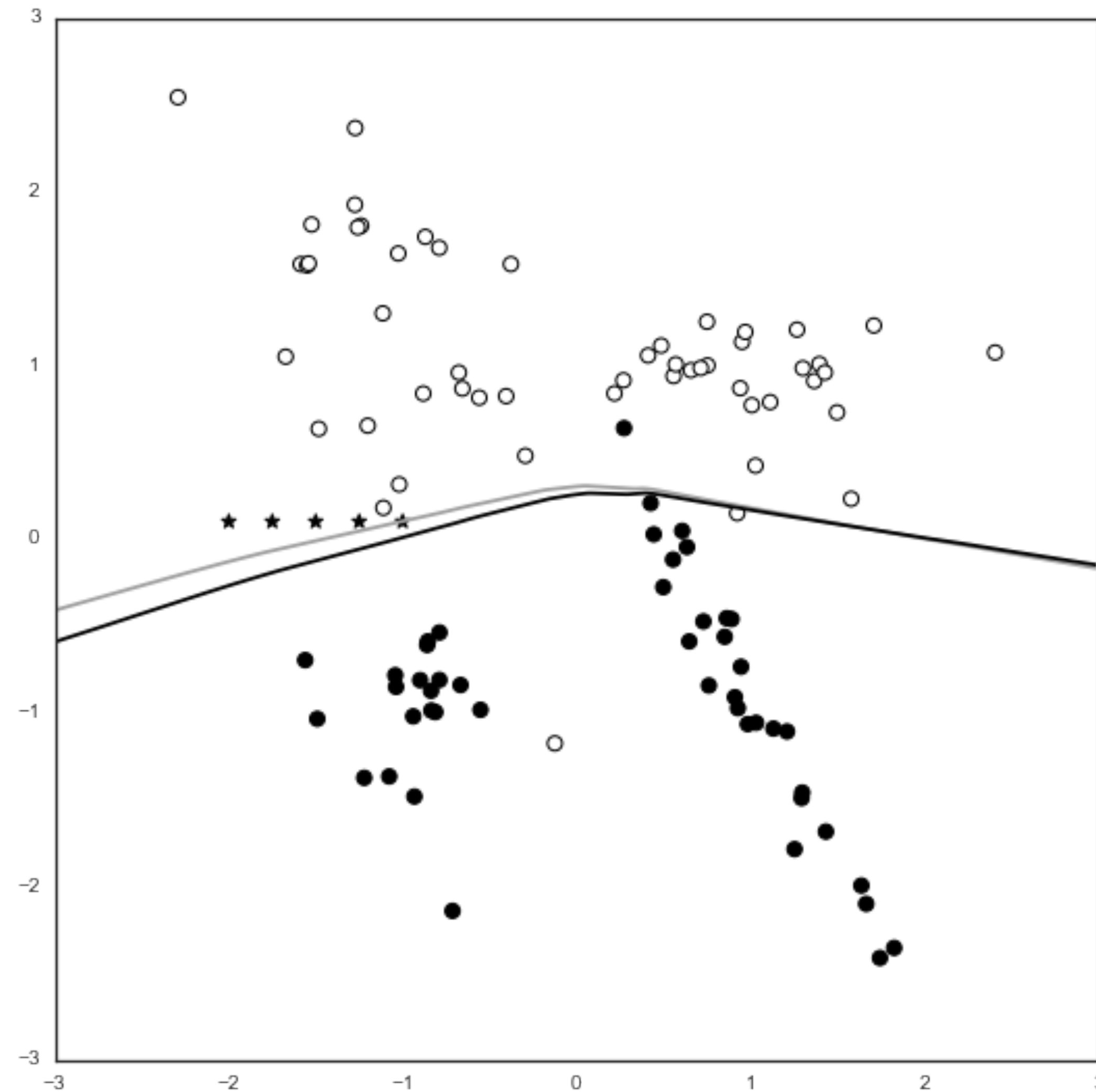
Poisoning Attack



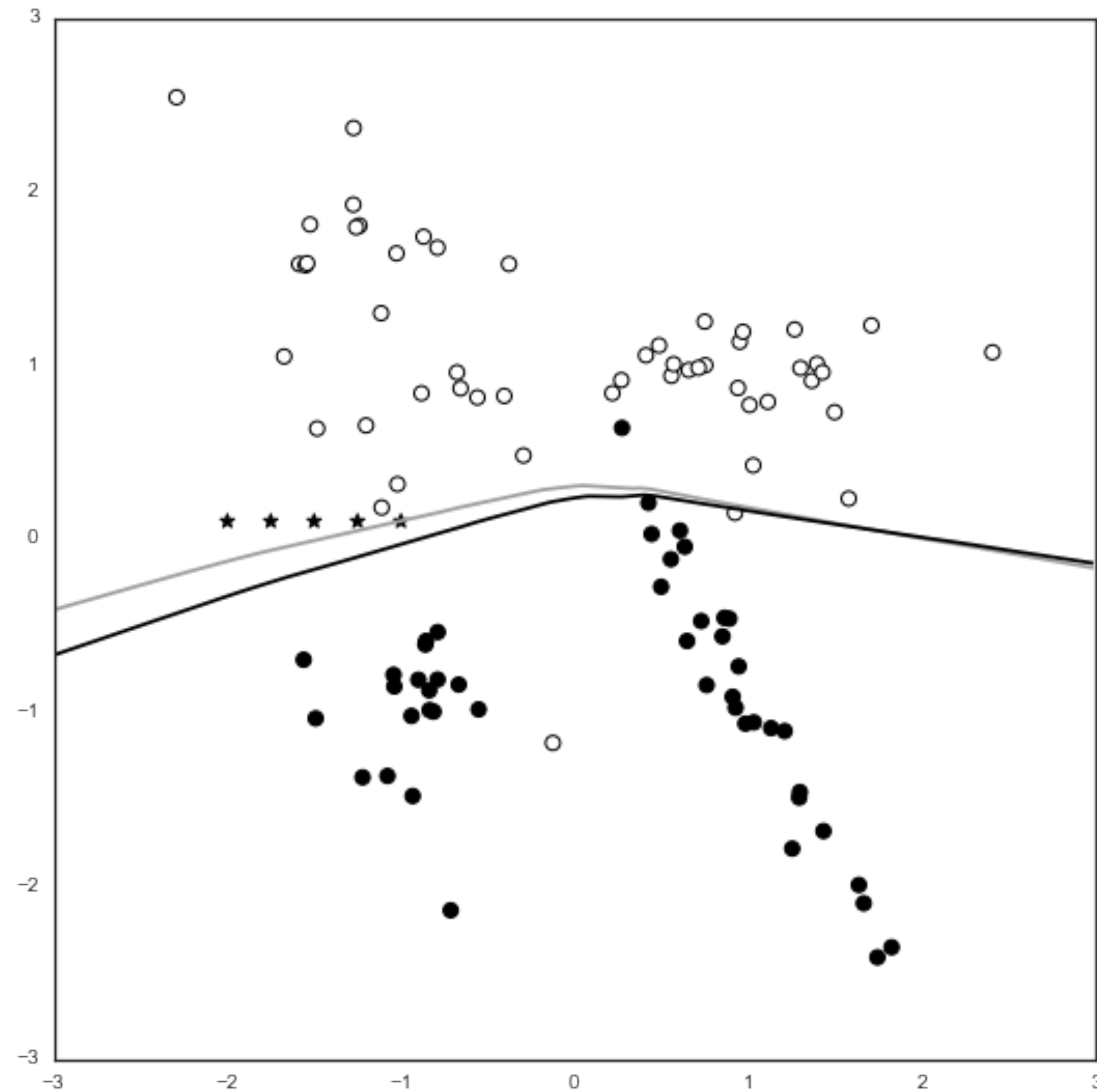
Poisoning Attack



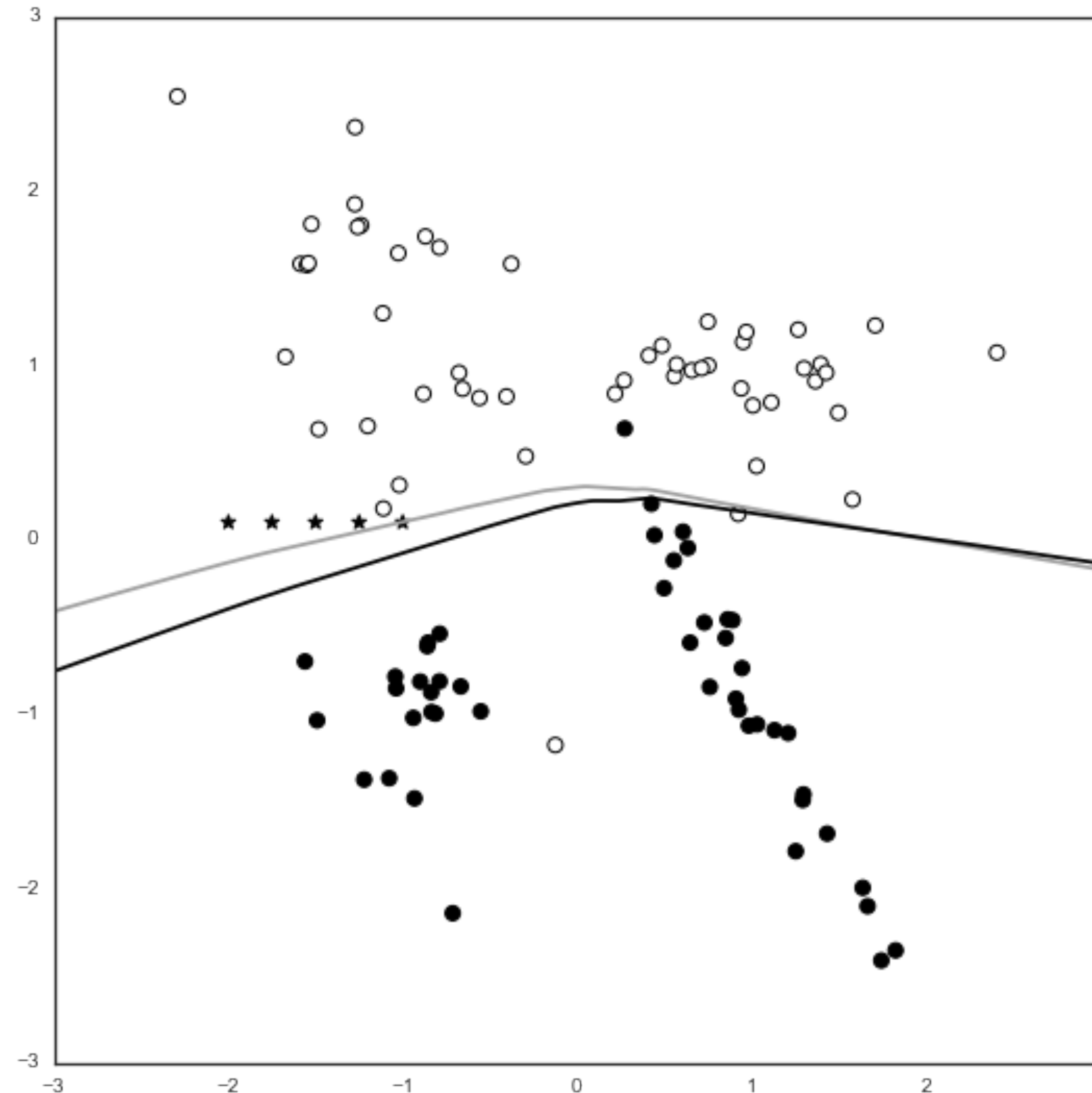
Poisoning Attack



Poisoning Attack



Poisoning Attack





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



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)