Adversarial Machine Learning

Lucy Jiang and Daniel Zhu - CSE 484 Final Project

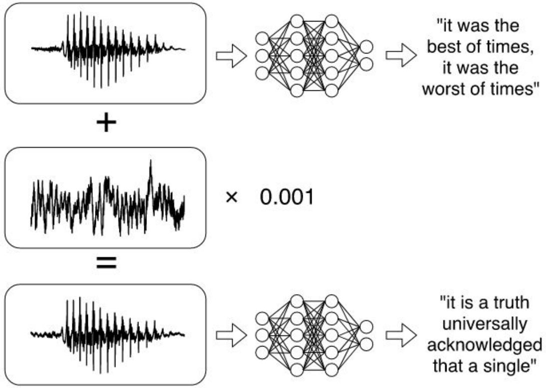
What is Adversarial ML?

Exploiting the vulnerability of machine learning systems to incorrectly evaluate manipulated inputs (adversarial examples) that are engineered by attackers

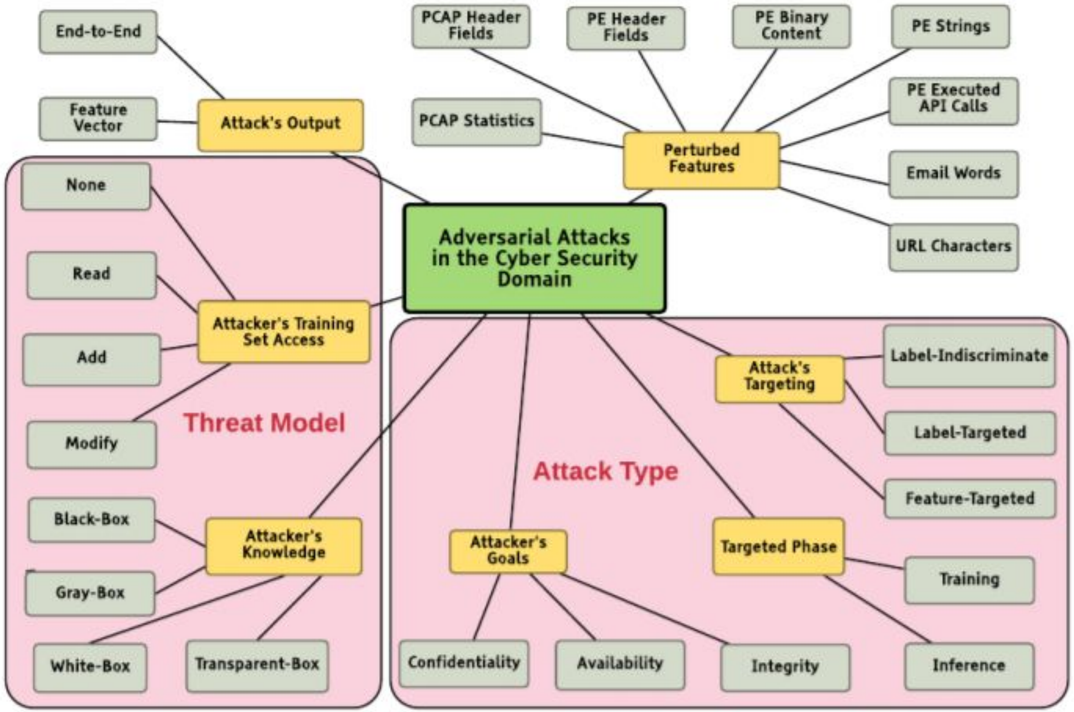
Adversarial Attacks in Practice

- Confuse autonomous vehicles with adversarial road signs
  - Stop sign can easily be physically or digitally modified to appear like a speed limit or yield sign / be ignored entirely
  - Life-threatening consequences that are not detectable to the human eye
- Manipulate automated speech recognition systems with adversarial audio
- Trick medical imaging systems to be certain about incorrect predictions
- Exploit NLP text classifiers
  - Spam filters, sentiment analysis, etc.
- Microsoft's Tay Chatbot

Adversarial Attacks in Practice, Cont.



Adversarial Machine Learning Taxonomy



Source: Rosenberg et. al. 2021

Adversarial Attacks

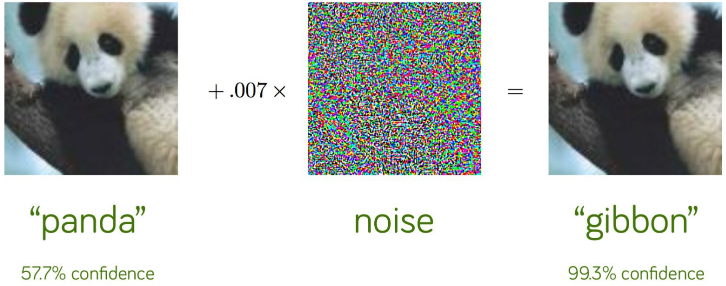
Data Poisoning

Contaminating the training data with misrepresentative samples during model training

Ex. Internet users feed Microsoft's Tay chatbot with offensive tweets

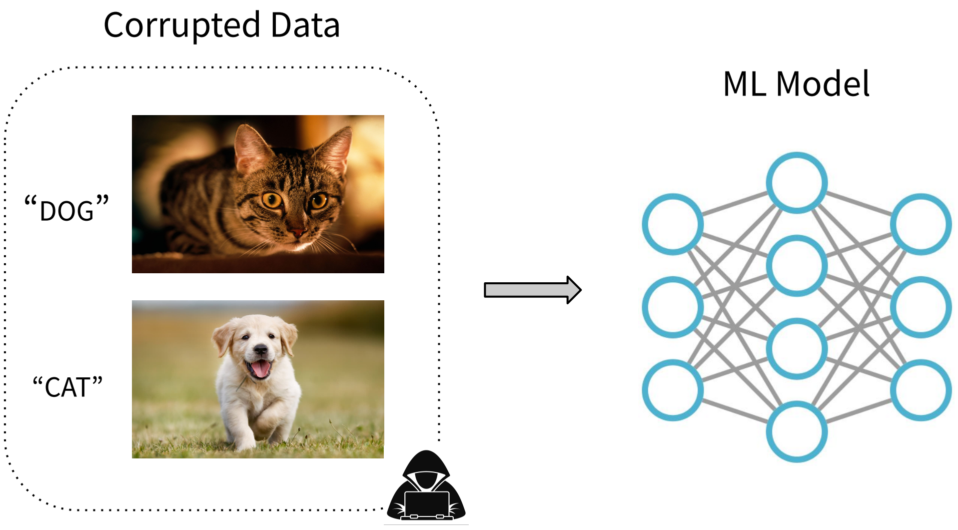
Adversarial Perturbation

Engineering malicious inputs that fool a model to make incorrect decisions at inference time



Ex. Researchers add a perturbation to an image causing a panda to be classified as a a gibbon

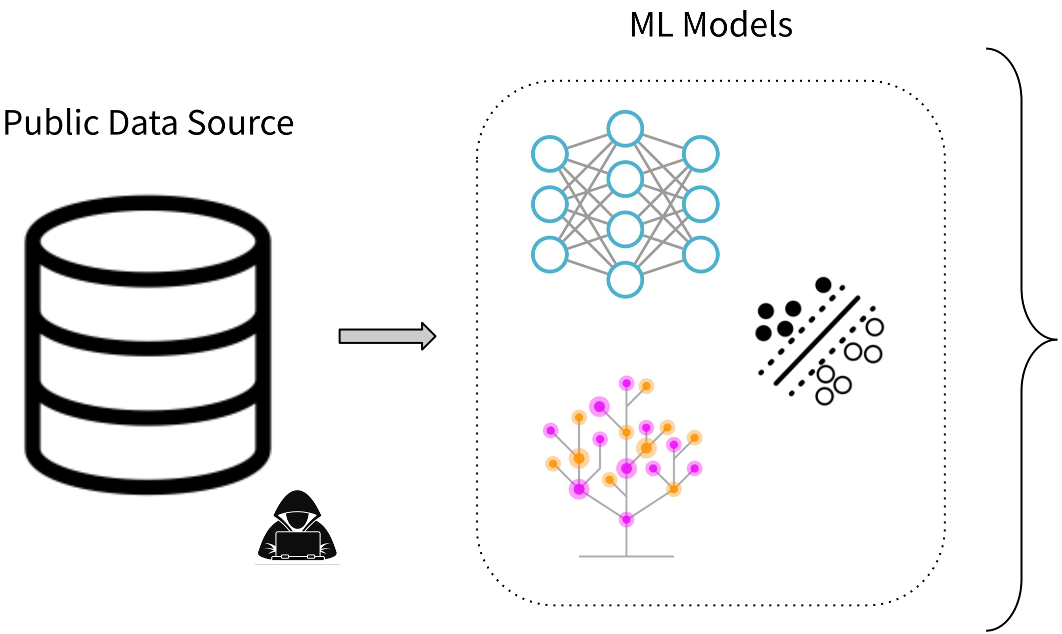
Targeted Data Poisoning



Impact

- Reduce confidence in predictions
- Misclassify specific examples
- Cause specific actions

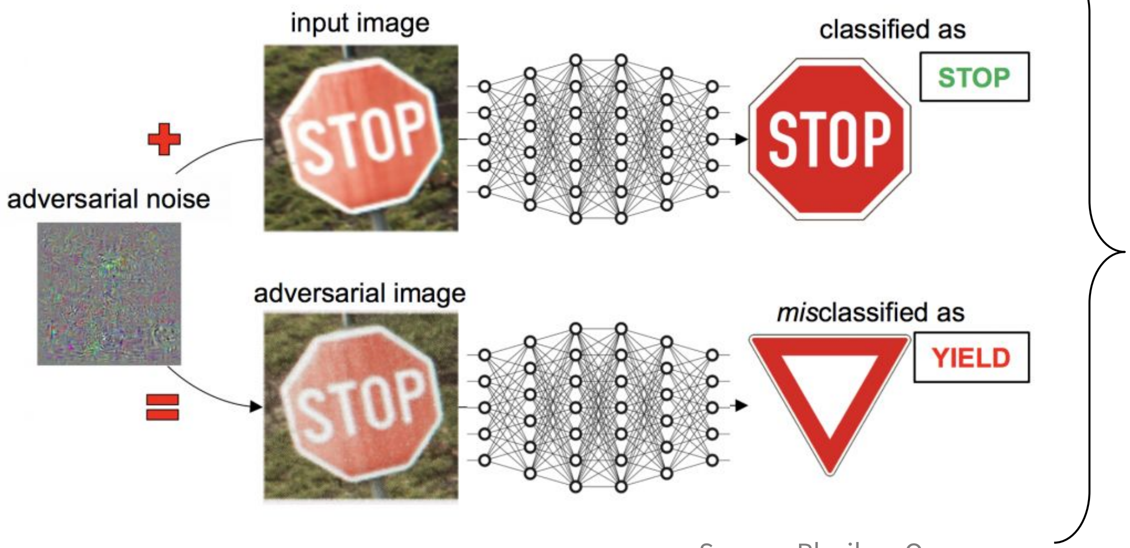
Indiscriminate Data Poisoning



Impact

- Ruin the quality and integrity of the data set
- Garbage-in = Garbage -out

Adversarial Perturbation



Source: Pluribus One

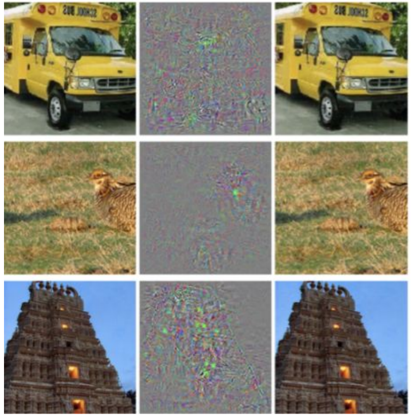
Impact

- Classify noise as a legitimate class
- Misclassify inputs
- Reduce the confidence of correct classifications

Defenses against Adversarial Attacks

Adversarial Training

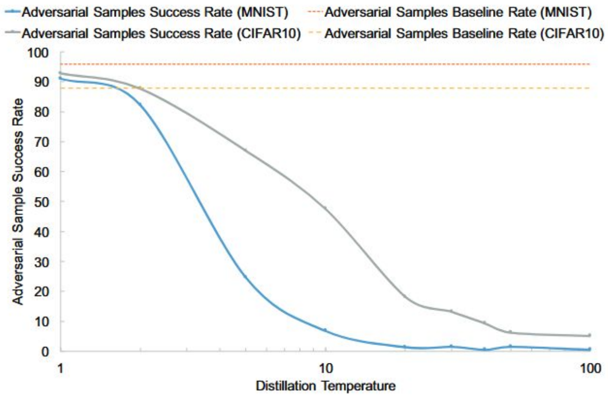
Augmenting the training data with artificially generated adversarial examples during training



Ex. Artificially add noise to training data to proactively train the model against potential adversarial inputs

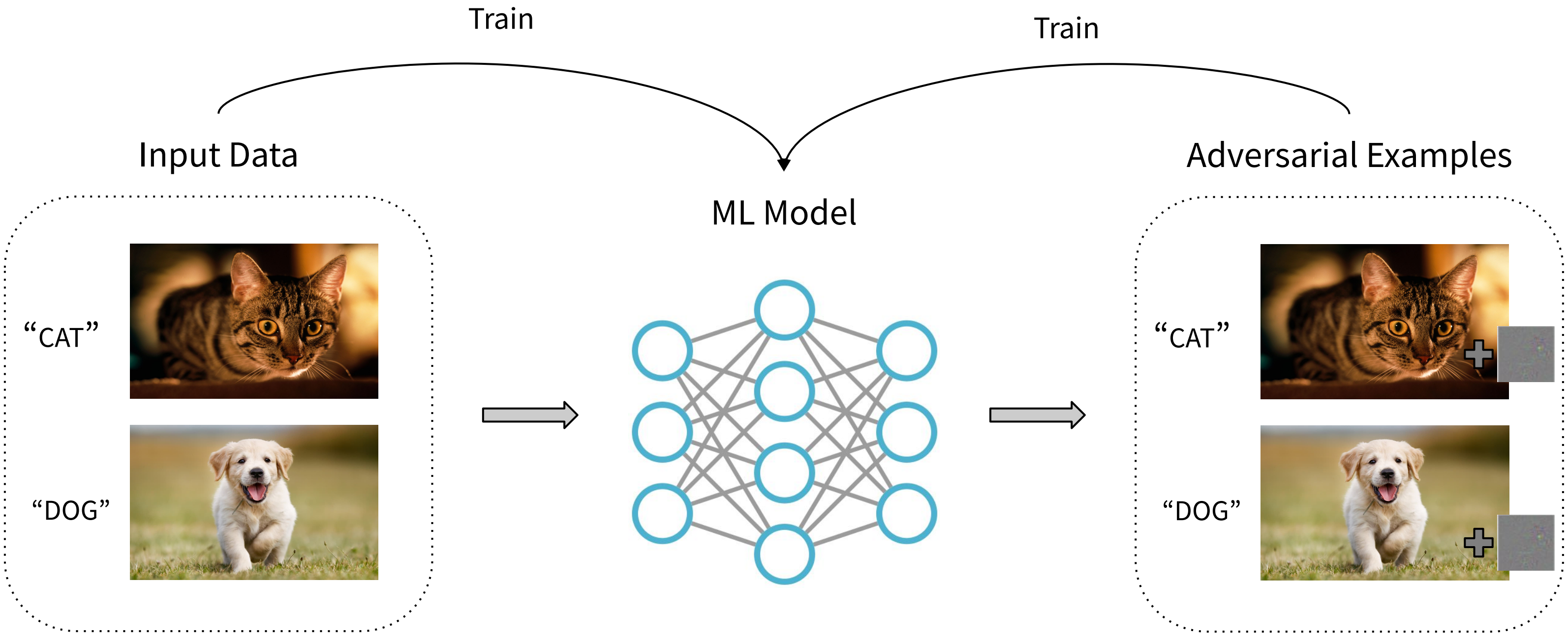
Defense Distillation

Training a model with the probabilities of different classes rather than on hard class labels

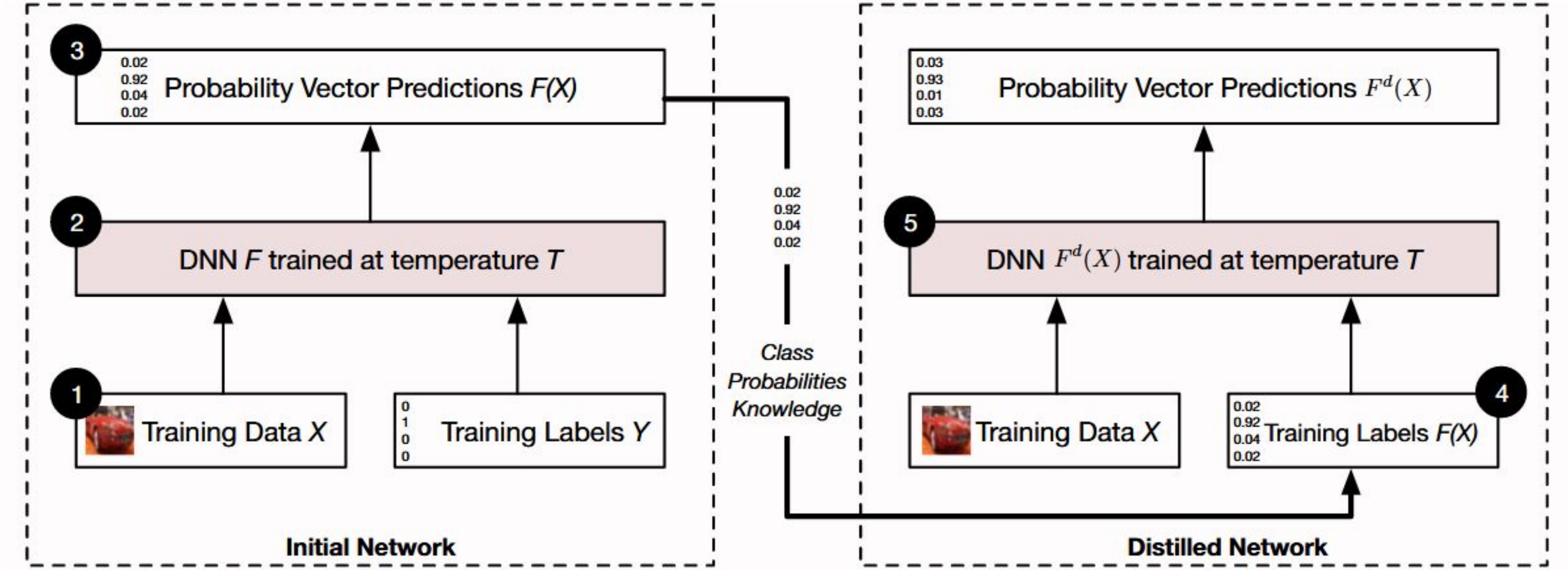


Ex. Plot of the percentage of targets achieved by crafting an adversarial sample while altering 112 features

Adversarial Training



Defense Distillation



Source: Papernot et. al. 2016

Ethical and Legal Implications

Case Study: Microsoft's Tay

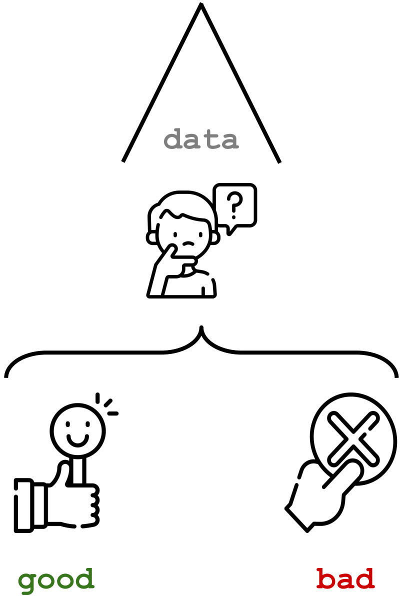
- Designed to interact like a human and trained with data from other Twitter users
    - Intention: to develop better "conversational understanding" for products
- Taken down within 24 hours of release for **spewing hate speech**
    - Racist, sexist, and anti-Semitic language
    - Faced extreme backlash on Twitter

"The more you talk the smarter Tay gets"

Ethical and Legal Implications, Cont.

Case Study: Microsoft’s Tay

- Training a model solely with public, user-supplied input makes them very vulnerable to adversarial data attacks
  - Should **monitor user-provided data** heavily, especially early on
  - Should **classify training data** as morally good or bad to encourage ethical outcomes (Candid)
- Should be better regulations to **review emerging technologies and prevent release**
  - Similar to an Institutional Review Board at the university level



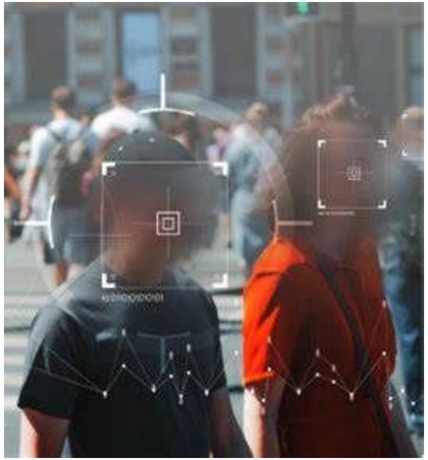
Designing Secure ML / AI

- How would you know if your data has been poisoned?
- Are you training from user-supplied inputs?
- Does the model only output results necessary to achieving its goal?
- What is the impact of a false negative or a false positive?
- How sensitive is your training data in case it is recovered from your model?
- Where does your training data come from?

Future Implications

- AI / ML systems are becoming prevalent in society (IoT, autonomous vehicles, etc.)
  - Must be vigilant against adversarial attacks
- **Proactive, not reactive**
- Legislation to prevent untested / undertested systems from harming the public
  - **King County recently banned government usage of facial recognition** software due to privacy threats and biases against certain demographics
- Embed ethical considerations into engineering culture and education
  - CSE 492E - Computer Ethics Seminar

King County



References

- Adversarial Machine Learning (https://cltc.berkeley.edu/aml/)
- Attacking Machine Learning with Adversarial Examples (https://openai.com/blog/adversarial-example-research/)
- Threat Modeling AI / ML Systems and Dependencies (https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml)
- Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications (https://core.ac.uk/download/pdf/231074604.pdf)
- Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain (https://arxiv.org/abs/2007.02407)
- Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks (https://arxiv.org/abs/1511.04508)
- Towards the Science of Security and Privacy in Machine Learning (https://arxiv.org/abs/1611.03814)
- Learning from Tay's Introduction (https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/)
- Audio Adversarial Examples: Targeted Attacks on Speech-to-Text (https://arxiv.org/pdf/1801.01944.pdf)
- Adversarial attacks on medical machine learning (https://science.sciencemag.org/content/363/6433/1287)
- Finally, progress on regulating facial recognition (https://blogs.microsoft.com/on-the-issues/2020/03/31/washington-facial-recognition-legislation/)