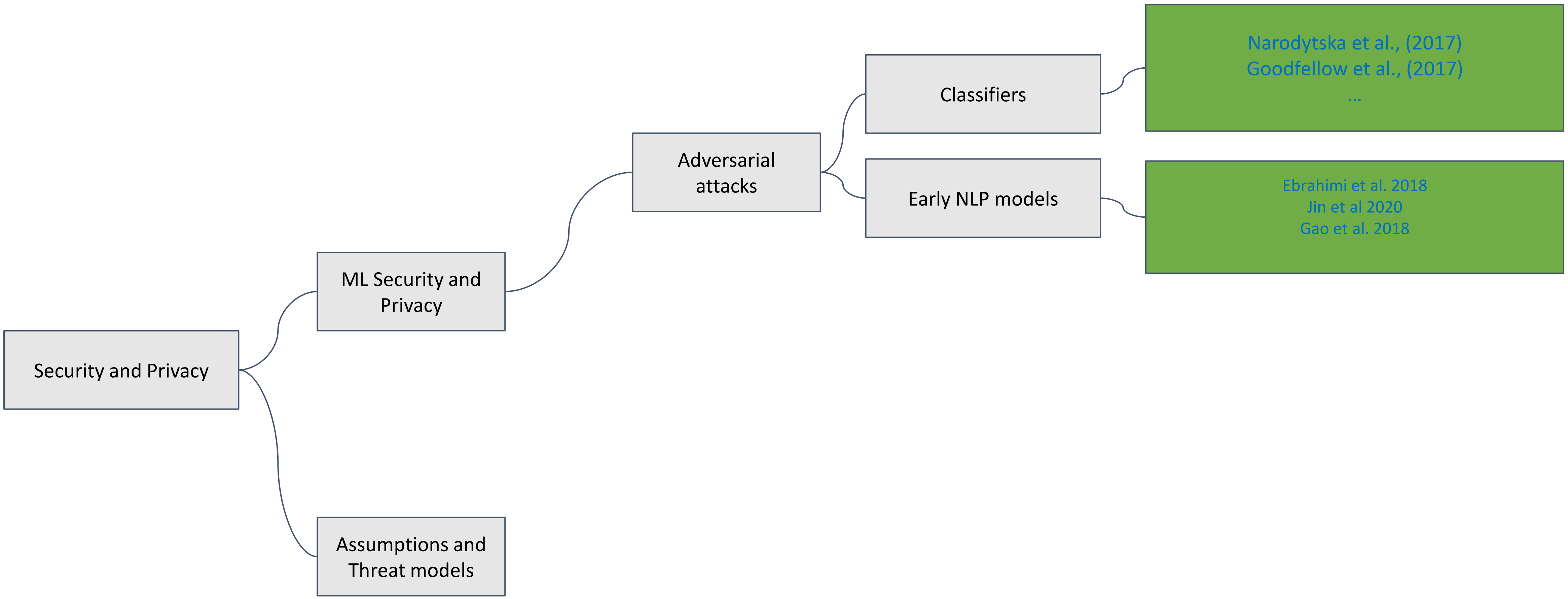
UC RIVERSIDE

Thinking like a Hacker

Slides: Md Abdullah Al Mamun and Nael Abu-Ghazaleh

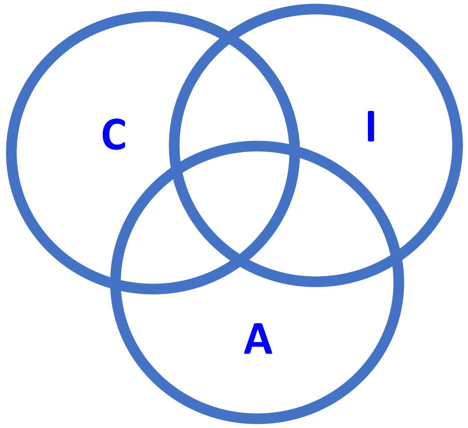
Roadmap



Security and Privacy
ML Security and Privacy
Assumptions and Threat models
Adversarial attacks
Classifiers
Early NLP models
Narodytska et al., (2017)
Goodfellow et al., (2017)
...
Ebrahimi et al. 2018
Jin et al 2020
Gao et al. 2018

Early Security: Information Security

- CIA Triad
  - Confidentiality: Who is authorized to use data?
  - Integrity: Has the data been modified?
  - Availability: Can access data whenever need it?



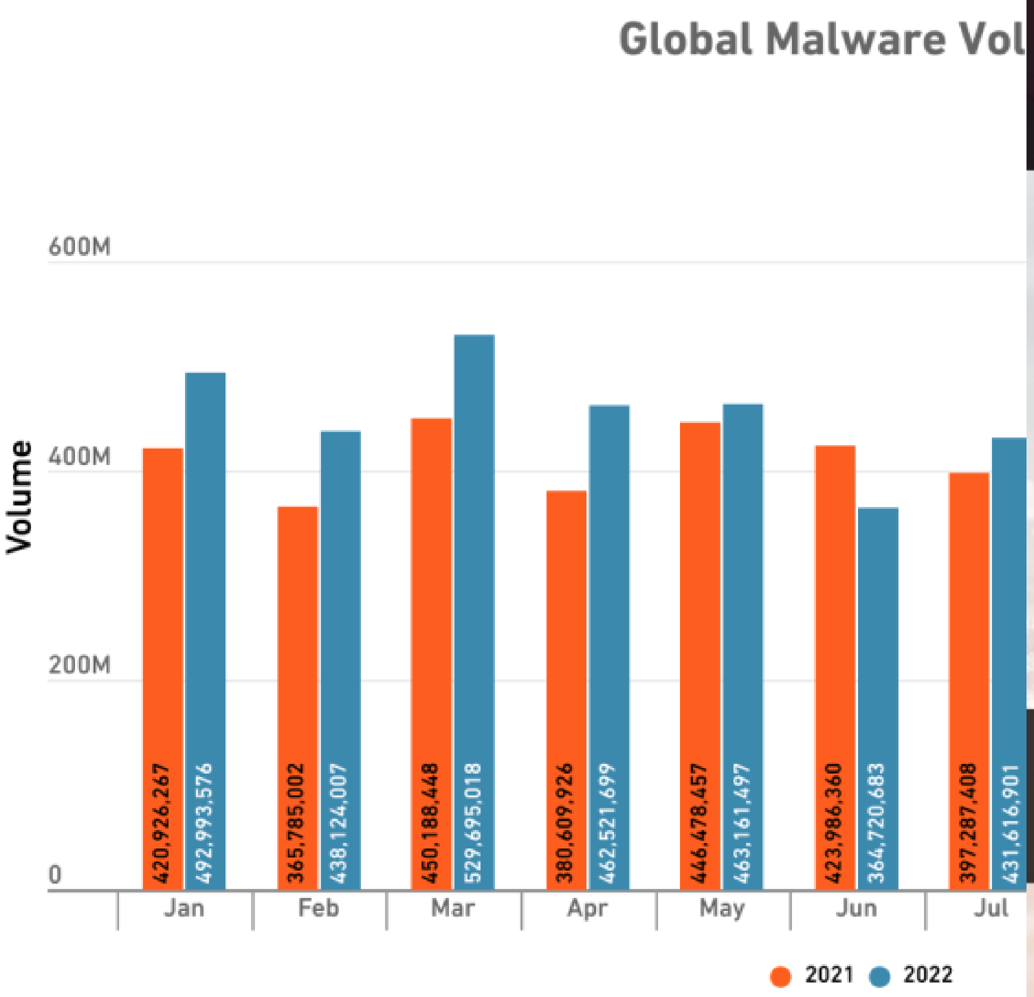
- Other components often added
  - Authentication
  - Authorization
  - Non-repudiation
  - …

An adversarial strategy that compromises any of these is an attack.

New application spaces

- Security no longer just about information
- Cloud is ubiquitous
- Machine learning everywhere
- New threat models
- Increasingly motivated and resourced attackers

What motivates attackers?



Trustworthy ML

- *Trust vs. Security*
- *CIA in the context of ML*
- *New concerns emerge*
  - *Fairness and Inclusiveness*
  - *Toxicity*
  - *Safety*
  - *Sustainability*
  - *Explanability*
  - *...*

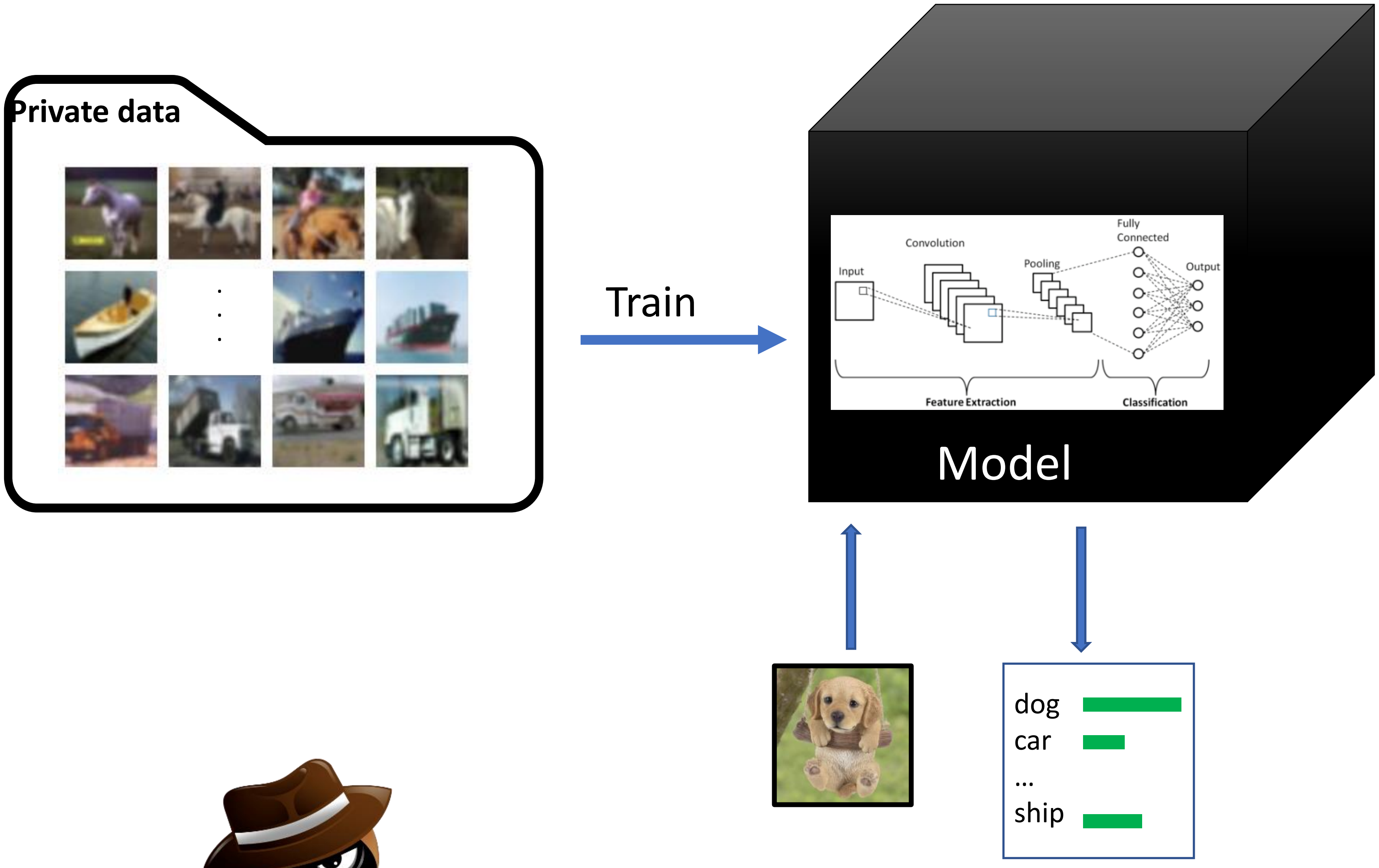
Threat models

- *What are our assumptions with respect to the attacker?*
  - *How does the attacker access the system?*
  - *What are they able to observe?*
  - *What is their goal?*
  - *Any other assumptions about the system?*

- *The less the assumptions, the more dangerous the attack*

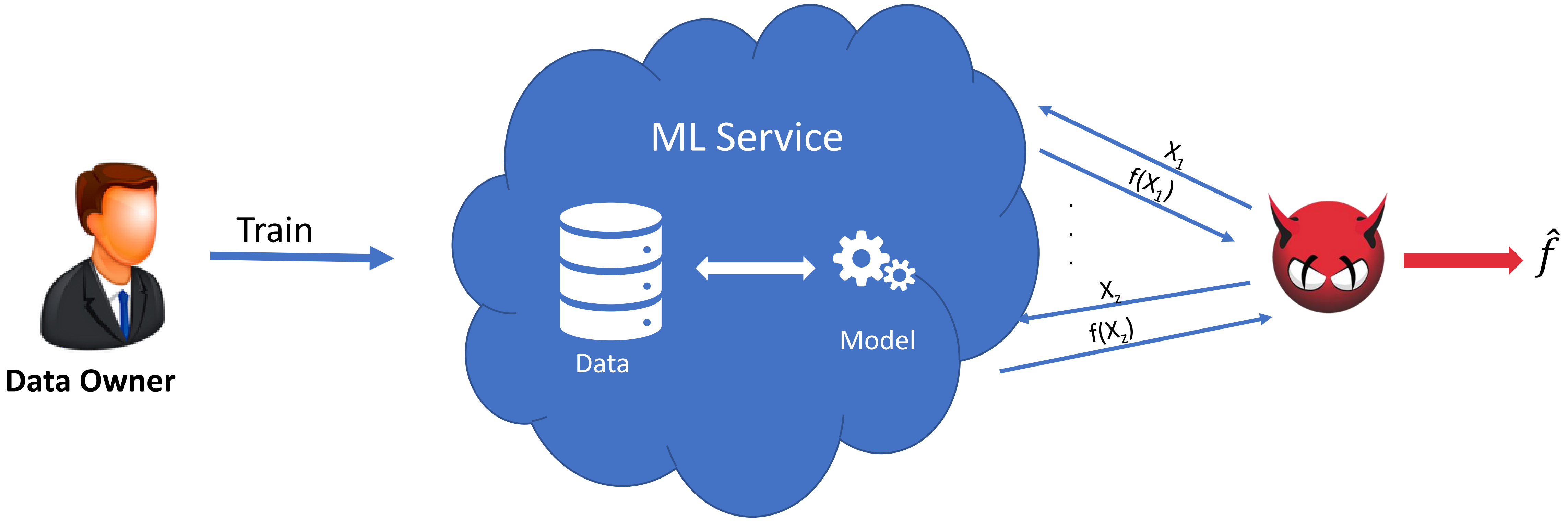
What are some common ML threat models/attacks?

Membership Inference attacks

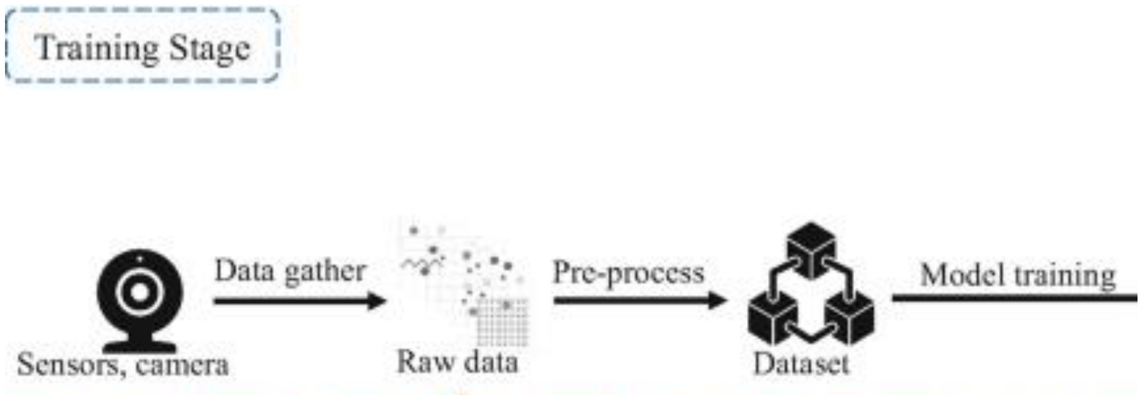


Was this specific picture in the training set?

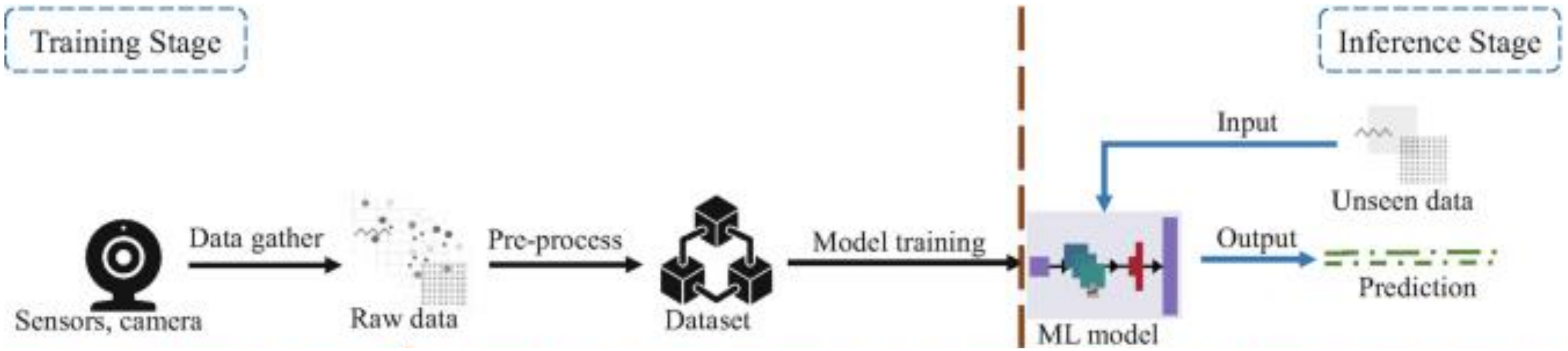
Model Extraction attacks



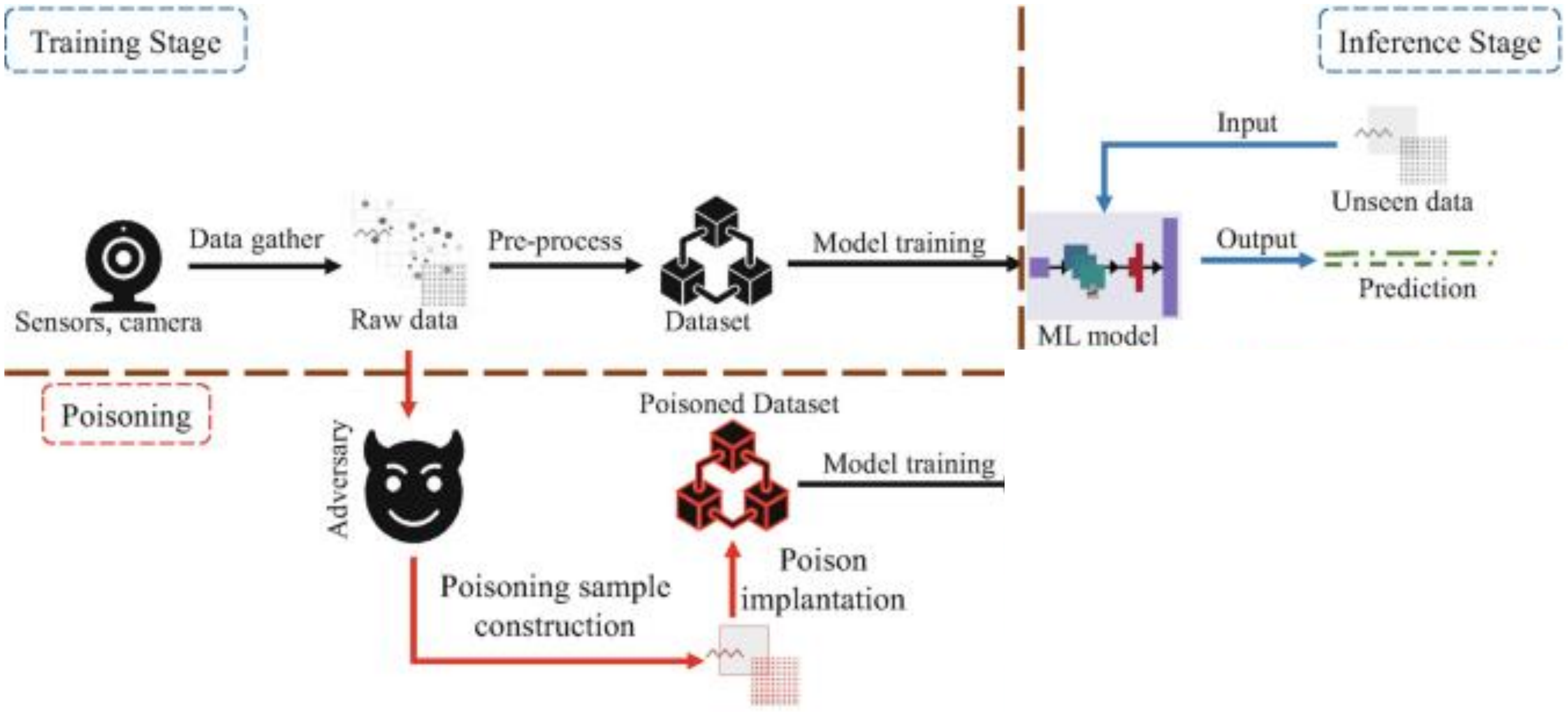
Poisoning attacks



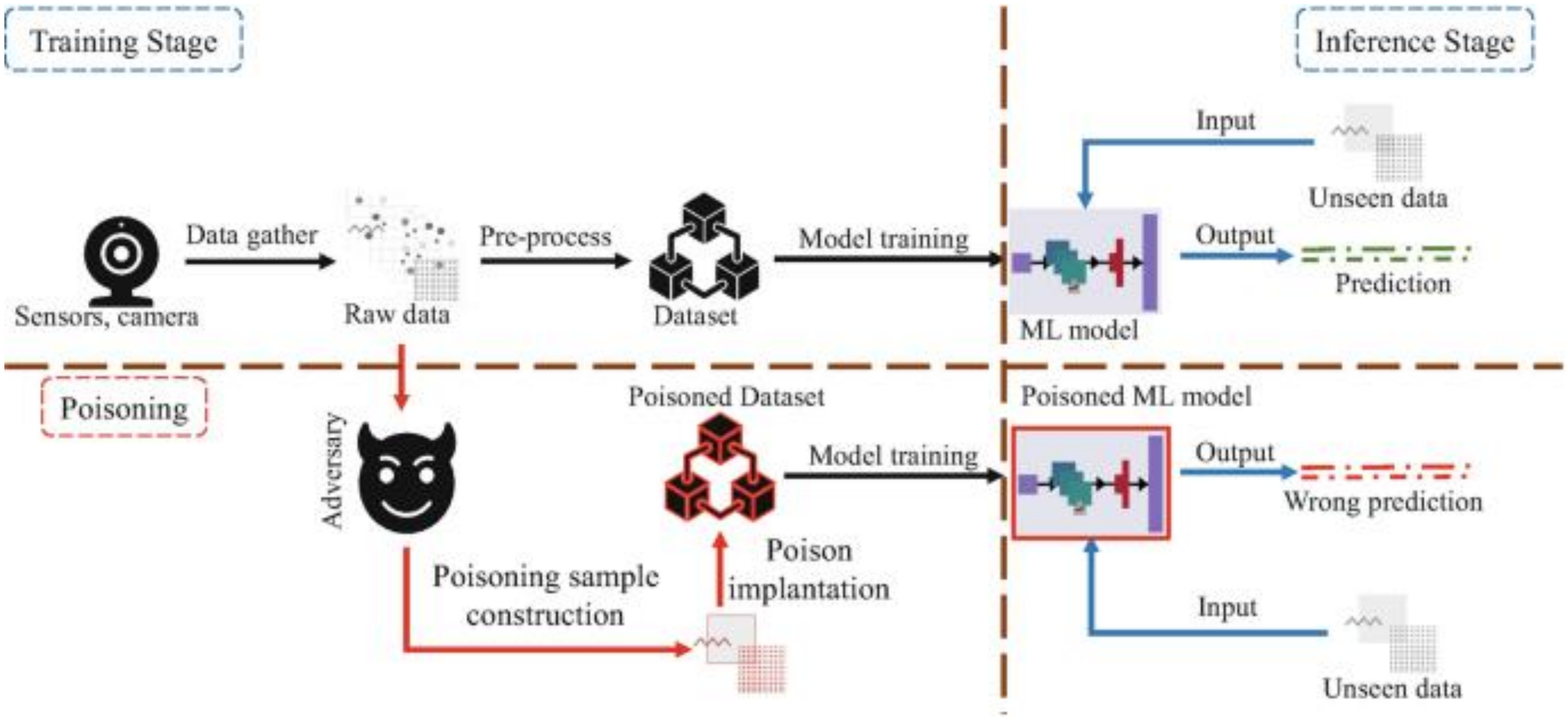
Poisoning attacks



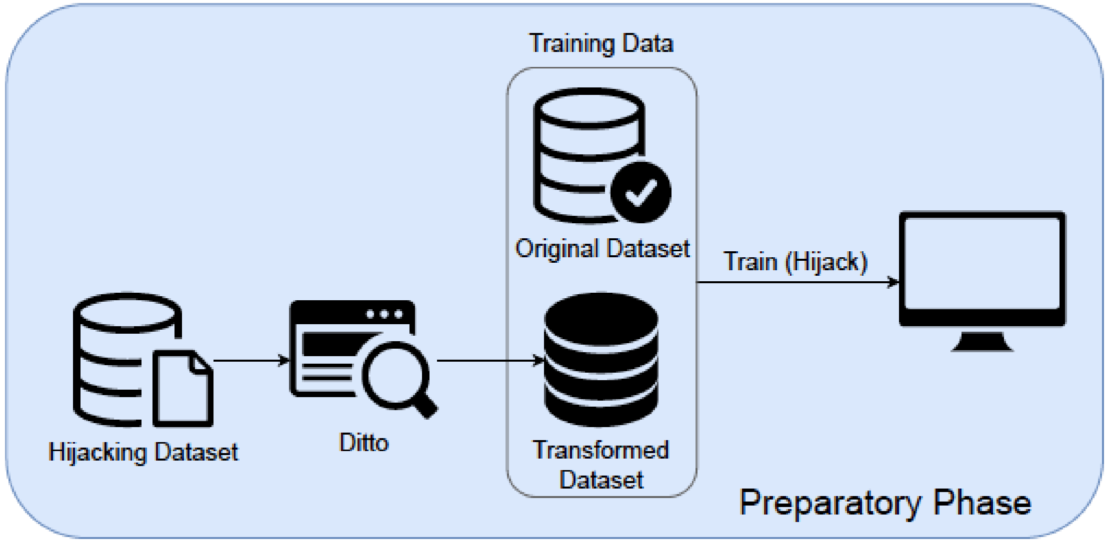
Poisoning attacks



Poisoning attacks



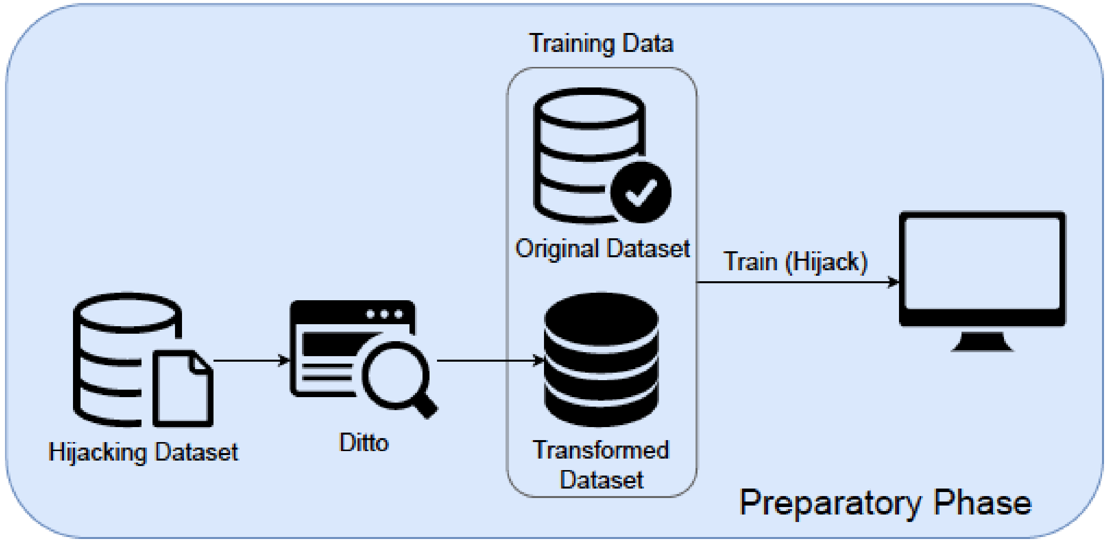
Model Hijacking



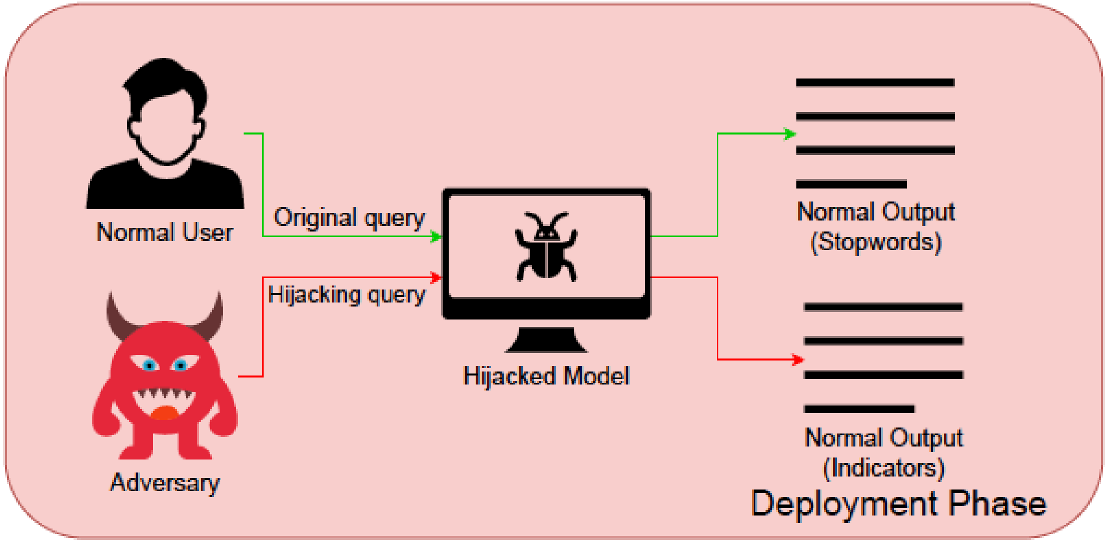
(a) Preparatory phase

Two-in-One: A Model Hijacking Attack Against Text Generation Models, Usenix security 2023

Model Hijacking



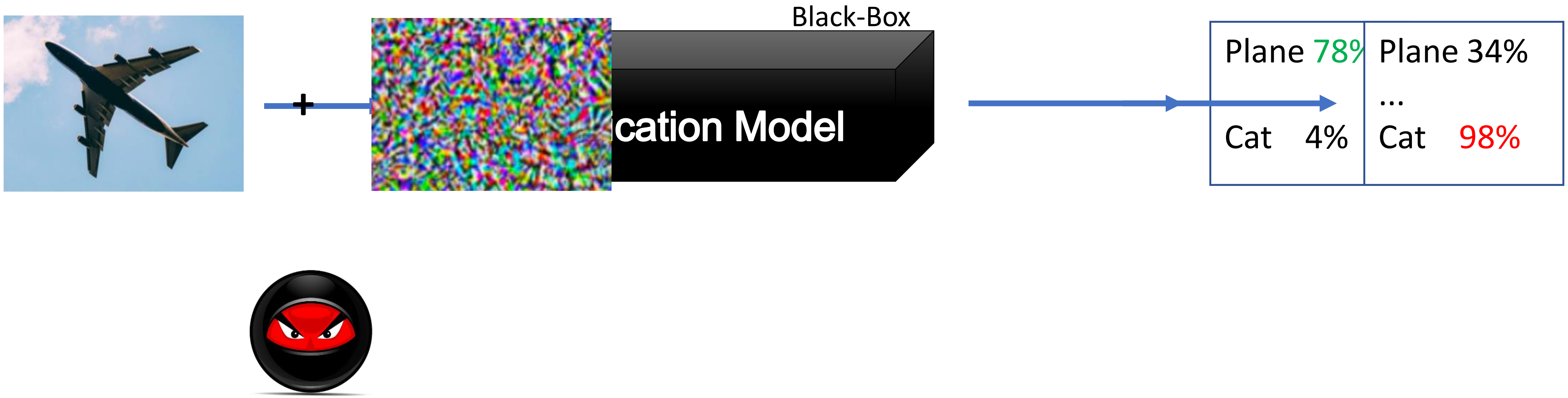
(a) Preparatory phase



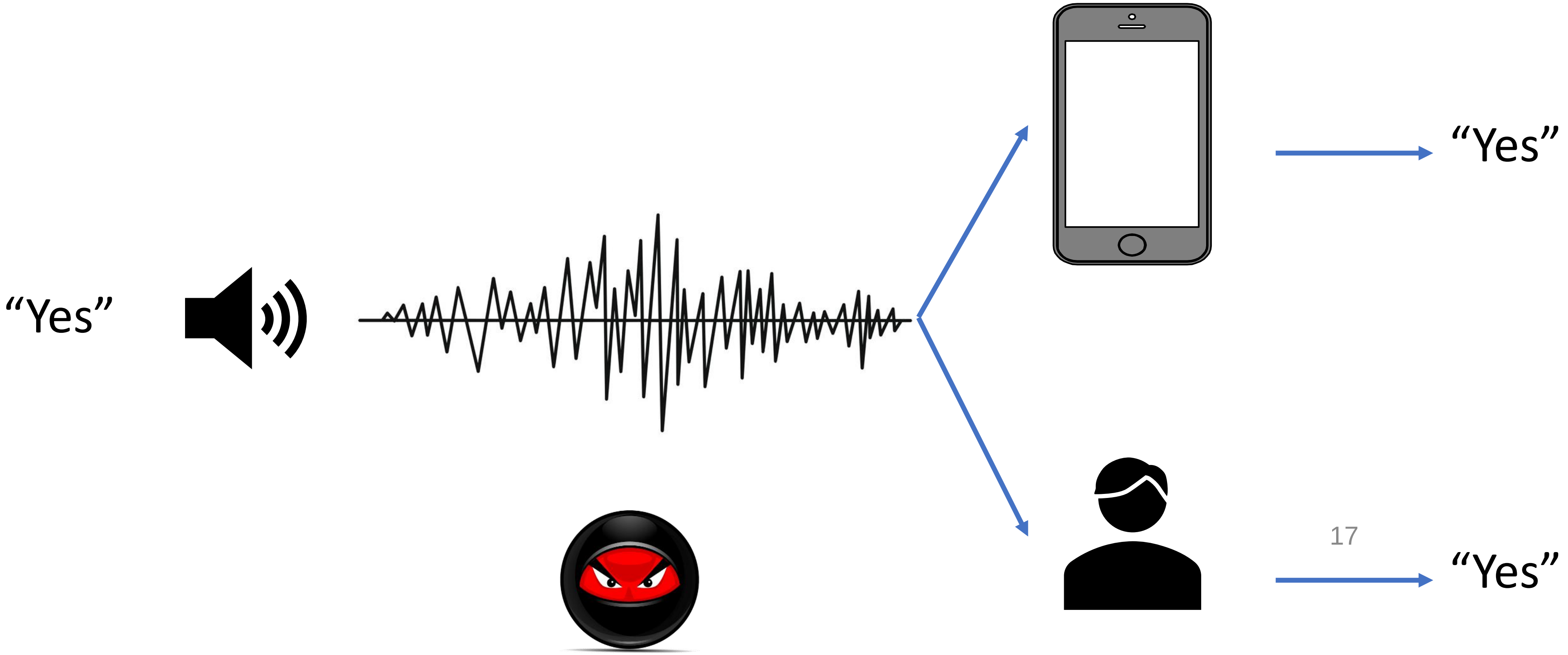
(b) Deployment phase

Two-in-One: A Model Hijacking Attack Against Text Generation Models, Usenix security 2023

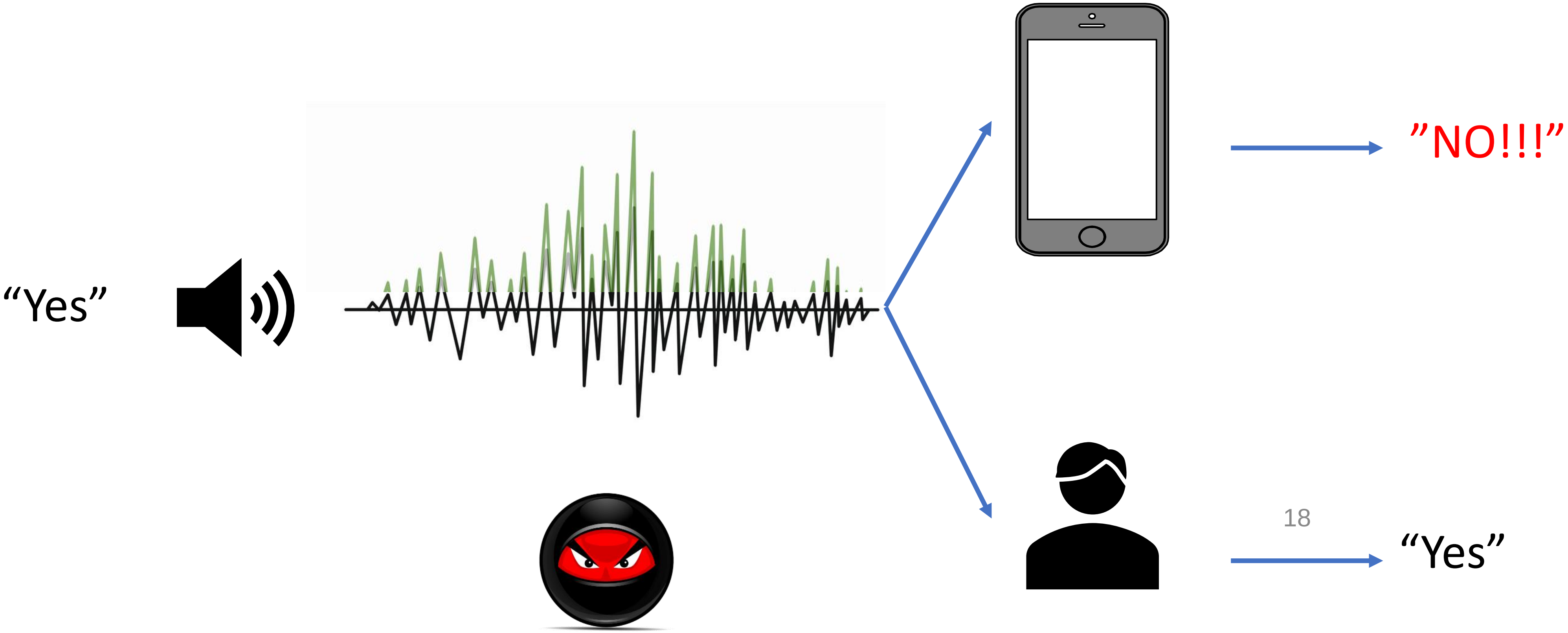
Adversarial attacks



Adversarial attacks

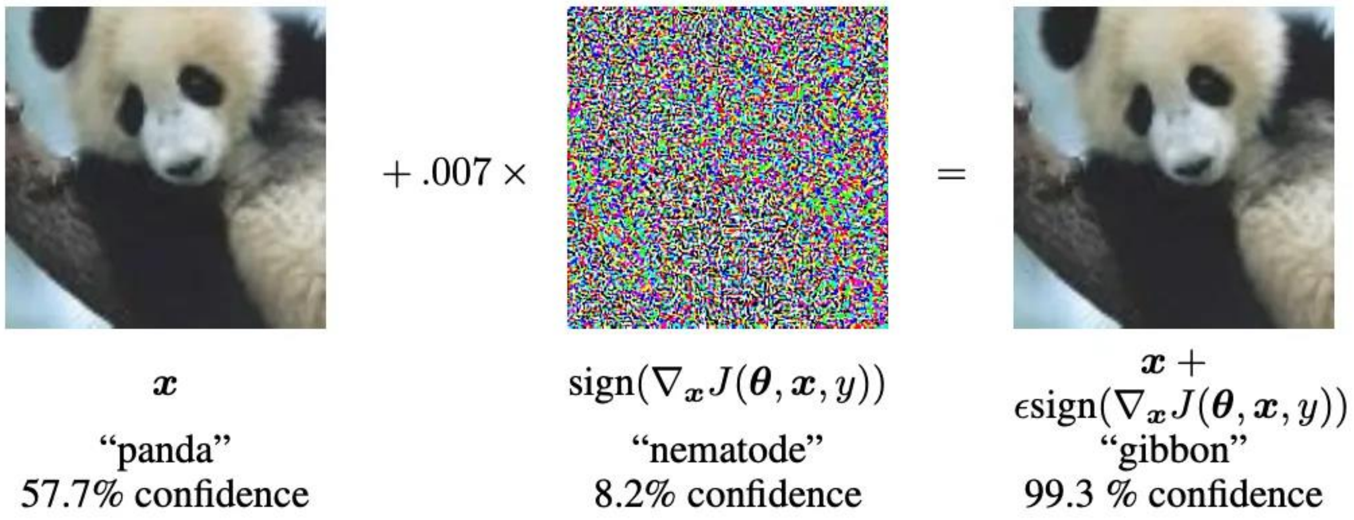


Adversarial attacks



Adversarial Attack

- **Adds noise to the input data which is imperceptible to human but alter the model's prediction**

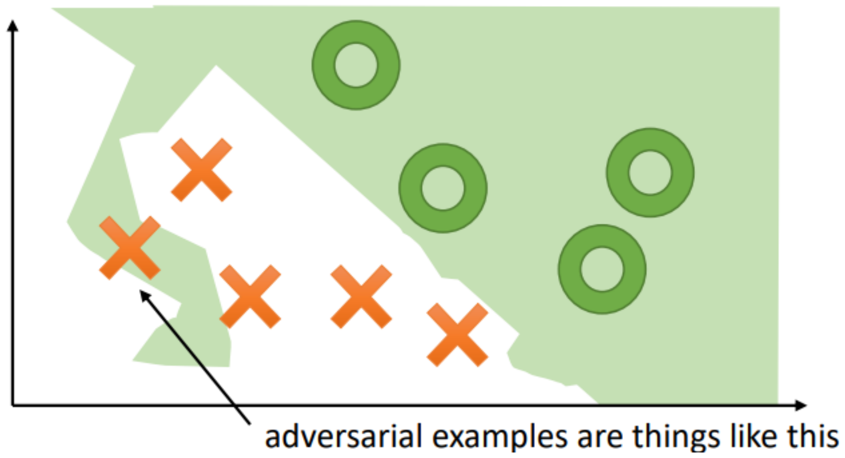


Cause was initially a mystery: extreme nonlinearities of the model? Insufficient regularization?

Szegedy et al., (2014): Explaining and Harnessing Advesarial Examples

Interesting properties

- Likely explanation: linear changes in high dimensional models
  - Tension between building models that are easy to train and vulnerability to adversarial attacks



Adversarial examples on the decision boundary

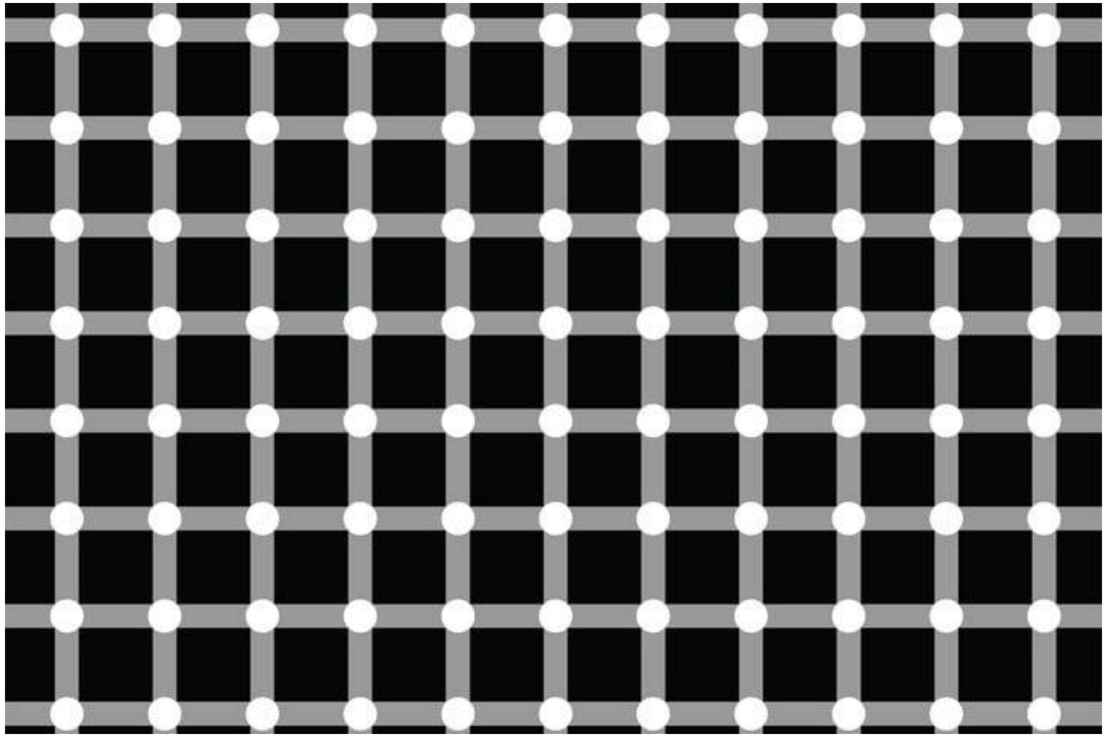
- Attacks can be performed from any class to any class using a constrained noise budget

UC RIVERSIDE

Interesting properties

- Attacks transfer!
  - They are not a function of the model, but capture something fundamental

- Training on adversarial examples can regularlize models
  - Conventional regularization approaches do not work



Source: Reader's digest (https://www.rd.com/article/optical-illusions/)

White-box Adversarial Attack

- Fast Gradient Sign Method (FGSM)
- Hotflip
- TextFooler

Fast Gradient Sign Method (FGSM)

- *Each pixel can change by at most a small amount* ϵ

- **Move the dimensions of x in direction of $\nabla_x \mathcal{L}$ by ϵ**

$$x^* = x + \epsilon \operatorname{sign}(\nabla_x \mathcal{L})$$

Hotflip

- Requires access to the internal gradients of the model
- *Single character changes* (substitution, insertion, or deletion)
  - Sets up an expression for the loss
- Identifies the most influential characters or tokens in the input text
  - Uses beam search to identify multiple character perturbations

Hotflip

Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the opposition Conservatives.

Original Prediction: 75% World

Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the oBposition Conservatives.

Altered Prediction: 94% Business

Table 1: **Adversarial examples with a single character change cause misclassification by a neural classifier**

TextFooler

- Generates adversarial examples by synonym substitution
- preserves the original meaning but changes the model's prediction
- Identifies and substitute the important words in input text using gradients
- Iteratively substitute the words until the model's prediction changes.

TextFooler

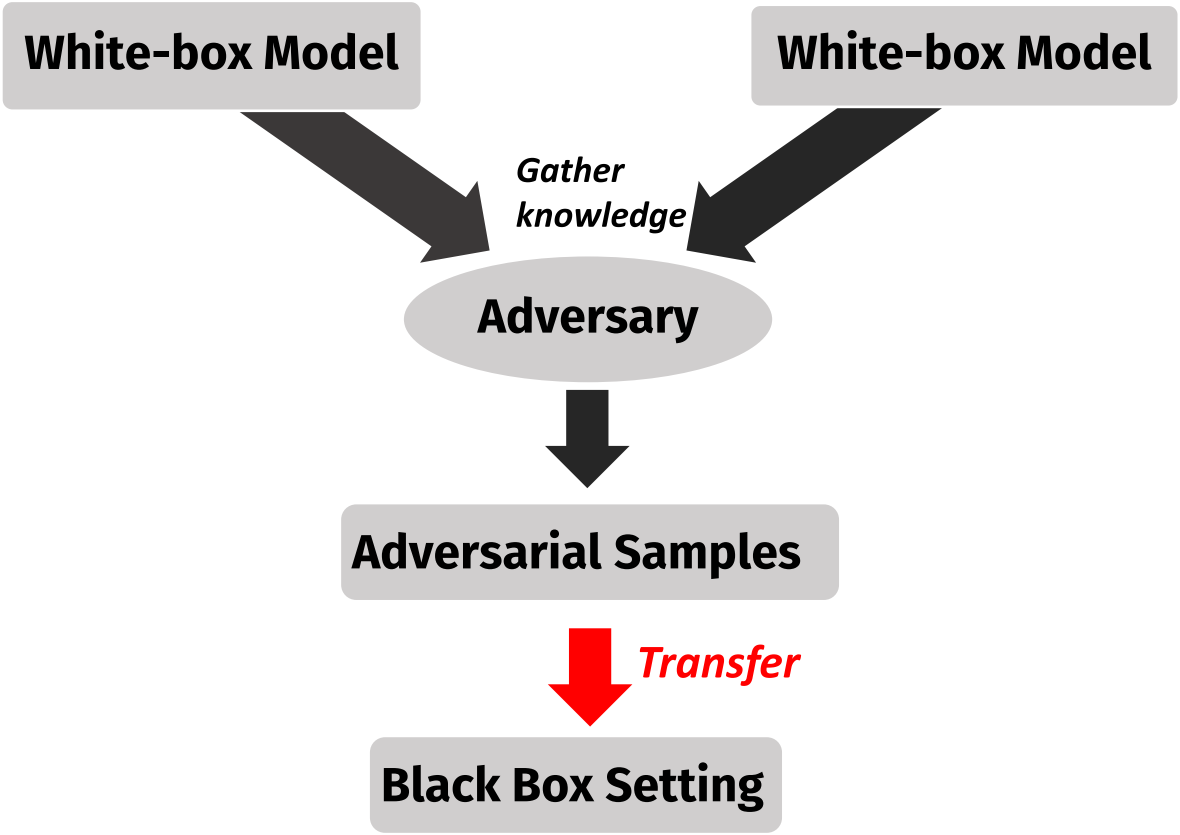
Original Sample: The characters, cast in impossibly **contrived situations**, are **totally** estranged from reality. (Label: NEG)

Adversarial Sample:The characters, cast in impossibly **engineered circumstances**, are **fully** estranged from reality. (Label: POS)

Table 2 : **Examples of original and adversarial examples (Movie Review (Positive (POS) ↔ Negative (NEG)))**

Black Box Adversarial Attack

- *Generate adversarial samples and transfer to the black-box setting*



DeepWordBug

- By comparing the prediction before and after a word is removed reflects which word influences most to the classification result
- Targets the characters in the most important words
- Perturbations include character swapping, insertion, deletion, or replacement

DeepWordBug

Original Sample: *The film has a special place in my heart*

Deep Learning Model

Positive Review

Adversarial Sample: *The film has a special plcae in my herat*

Deep Learning Model

Negative Review