QuantifiCare

A.I. seminar

Challenges of A.I. : integrity, confidentiality and availability

QuantifiCare

Adversarial Machine Learning Overview

QuantifiCare

Summary

I) ML is ubiquitous

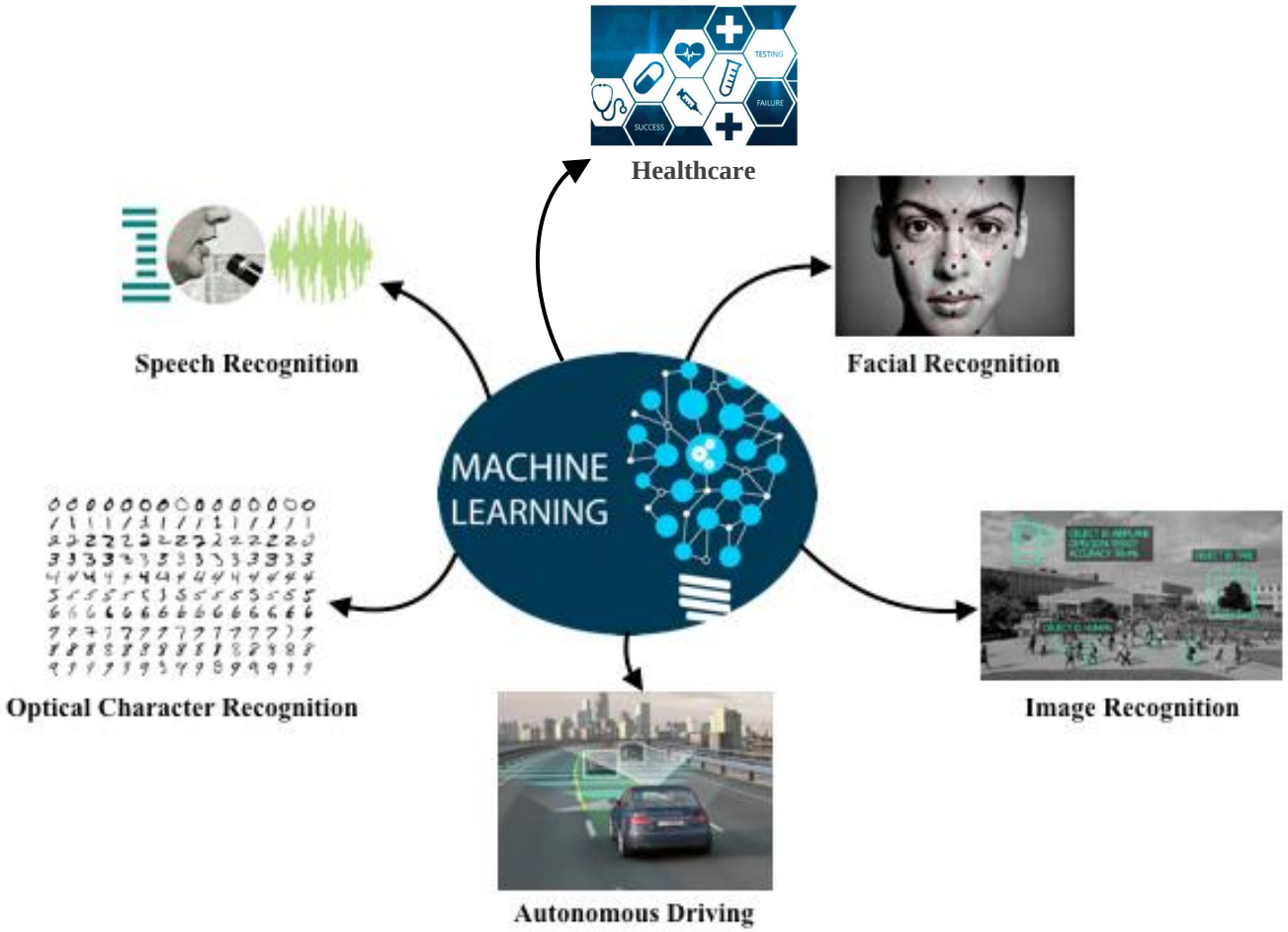
II) Context: Security of ML models

III) Threat model

IV) Different types of attacks

ML is ubiquitous

ML is ubiquitous



ML is ubiquitous

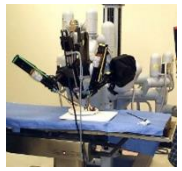
QuantifiCare

1) Will for ubiquitous Machine Learning

aws

Azure Machine Learning

Google Cloud Vision API



2) Wide range of attacks against *integrity, availability or confidentiality*

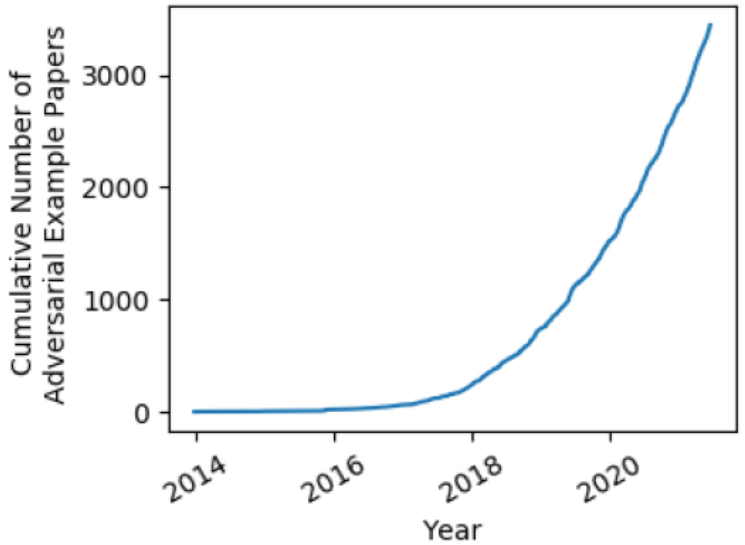
Security consequences of embedded models:

- Increased attack surface for an adversary: *exploit physical implementation flaws*

Emerging laws related to the security of ML systems:
New Regulatory Framework on Artificial Intelligence (April 2021)

Since 2014:

- The scientific community tackles these threats very seriously
- Beginning of standardization procedures



Context: security of ML models

Context: Security of ML models

QuantifiCare

Integrity

Protection against any malicious modification that aim at compromising the correct achievement of a task. When related to a machine learning model, integrity refers to the integrity of the inference process. An adversary mounting an attack against the integrity of the model targets its inference process, in that he aims at compromising the integrity of the output.

Context: Security of ML models

QuantifiCare

Confidentiality

The ability to protect confidential or private data from unauthorized access, in order to prevent a malicious disclosure of it. Private data may refer to confidential personal data such as medical or financial records, as well as to data protected by intellectual property legislation, such as copyrights or patents

Context: Security of ML models

QuantifiCare

Availability

The ability to keep a system accessible to authorized users. The goal of an adversary targeting the availability of a system is to prevent access to it, or significantly degrade its quality, with respect to specific requirements.

Attacks against availability => not in terms of a model only, but in terms of the overall system containing the model

Context: Security of ML models

QuantifiCare

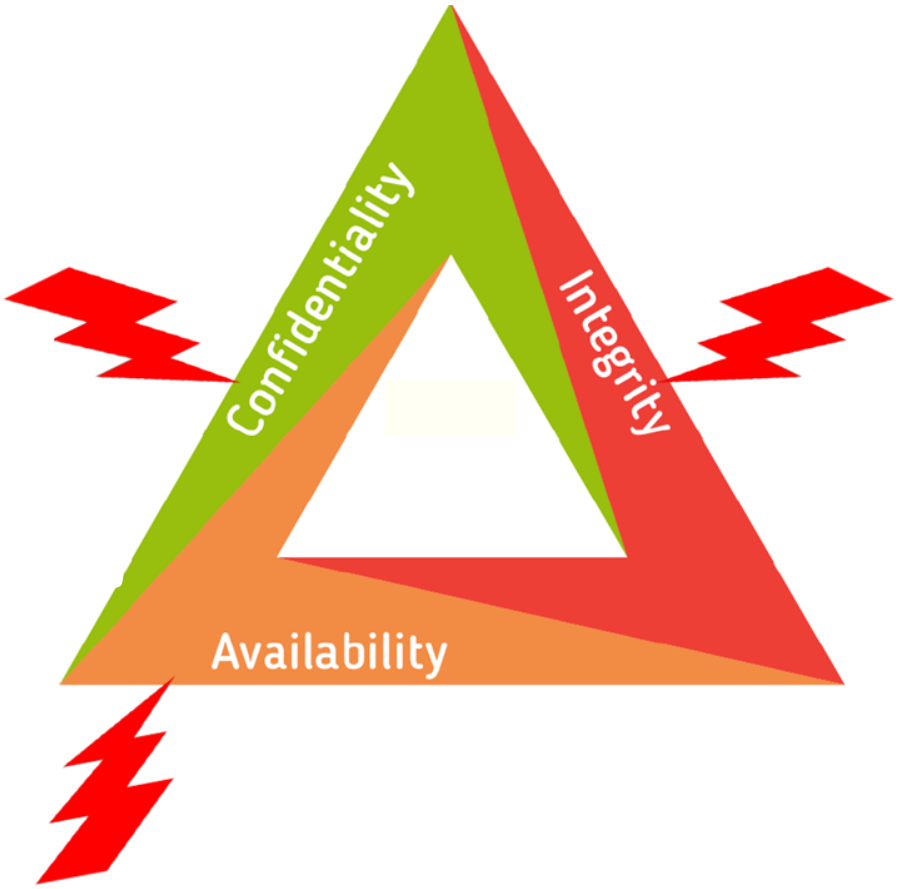
Availability

Attacks against availability => not in terms of a model only, but in terms of the overall system containing the model

1. Latency requirements: the availability can be threatened with denial of service attacks, which completely block the access to the system, or other attack methods increasing the response time
2. Quality requirements: an adversary can target memory constraints or processing time requirements
3. Task-based requirements:
   i) Integrity attacks (adversarial examples): reduce the accuracy on specific examples
   ii) Availability attacks (data poisoning): impact the behavior of the system

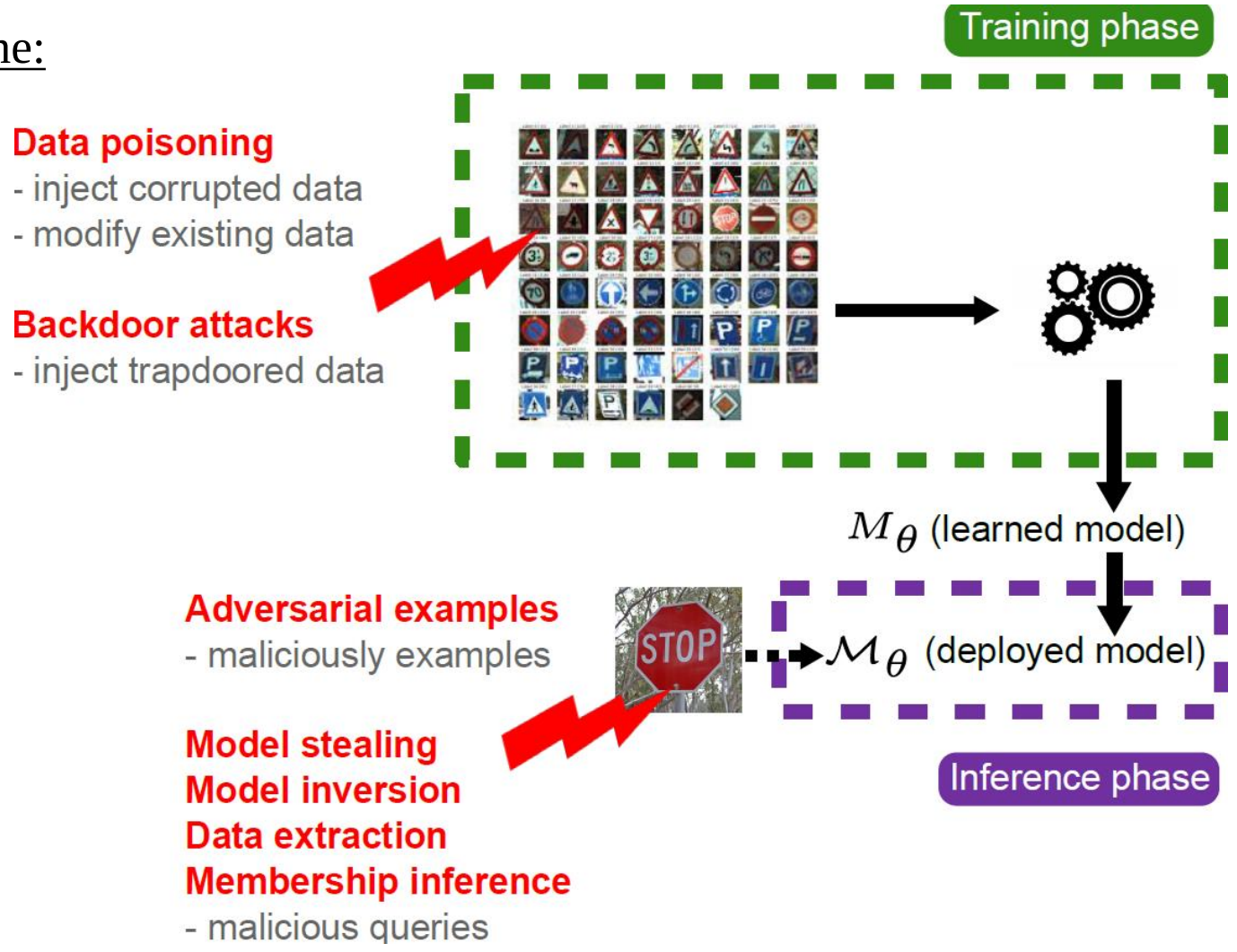
Context: Security of ML models

C.I.A. triad



Context: Security of ML models

Machine Learning Pipeline:



Threat model

Threat model

Adversary goal

What the adversary wants to do

Adversary knowledge

What the adversary wants to do

Adversary capacity

How much the adversary can intervene

Threat model

Adversary goal

Formally defines the objective of the adversary when attacking a machine learning model.

Adversary capacity

sets a bound on the ability of the attacker to cause harm. For some attacks, given some adversary goal and adversary knowledge, an unconstrained adversary would result in an attack that would be both unrealistic and without possible way of countering it.

Threat model

Adversary knowledge

Specifies what the adversary is able to know about the model under attack

For a neural network:

White-box setting: complete access to the target model.
This includes architecture, parameters, and ***the ability to compute gradients*** *(loss w.r.t. inputs, output w.r.t inputs, etc.)*

Black-box setting: limited access to the target model. **No ability to compute gradients.**
Only access to output label, or confidence scores, etc.

Threat model

e.g. adversarial examples:

Adversary goal

What the adversary wants to do

Fool the model (make it predict not the correct label)

Adversary knowledge

What the adversary knows about the target

Architecture/Parameters/Possibility to compute gradients

Adversary capacity

How much the adversary can intervene

How much can he distort the original clean image

Different types of attacks

Different types of attacks

Training time attacks:

Poisoning attacks
Backdoor attacks

The adversary modifies training data
=> Powerful adversary (requires permissive access to the training pipeline)
=> Often, one of the stakeholders of a project is involved

Inference attacks:

Membership Inference attacks
Adversarial examples
Model stealing
Model inversion
Data extraction

The adversary modifies test data
=> Can be anyone even if very limited access

Training-time attacks

Target:

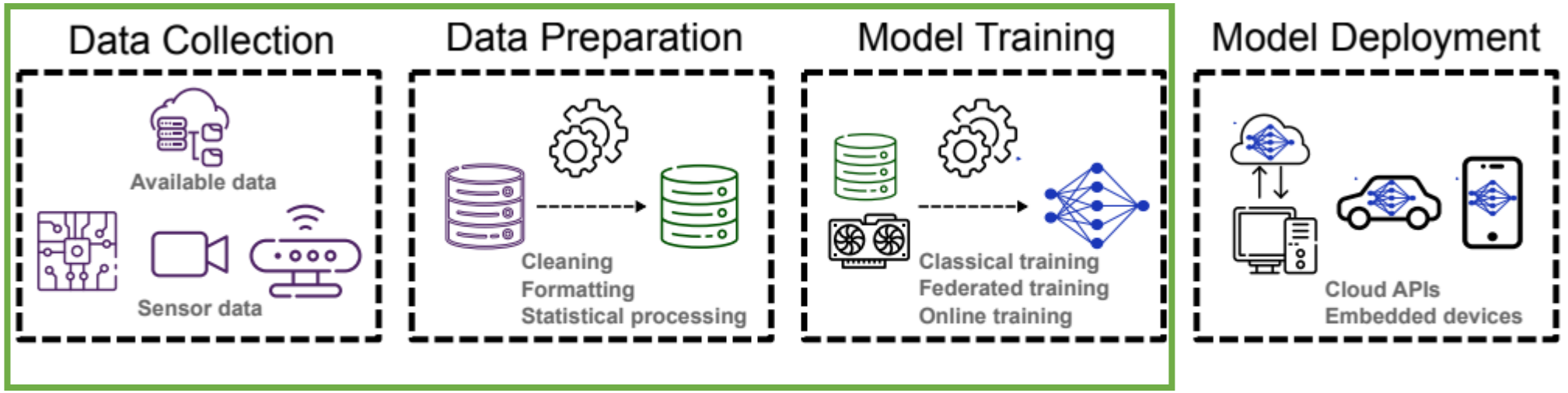


Figure 2.1: Complete machine learning pipeline. Different phases of the development of a model.

Training-time attacks

Goal:

Maliciously influence the training stage, in order to eventually have some specific influence on the final model.

Two scenarios:

1) Training phase is outsourced on the cloud for example.
=> An adversary that manages to penetrate the system responsible for the training phase.

2) The adversary is one of the stakeholders of the project dedicated to the development of the machine learning model.

=> He can interfere in two ways

Training-time attacks

QuantifiCare

Scenario: The adversary is one of the stakeholders of the project dedicated to the development of the machine learning model.

First way: Federated learning
(a model is trained in a decentralized way among different collaborative actors)

Each actor participates to the training phase with his own data, training a local model, and then sharing model updates with other actors on a common server to update a global model, which then sends updates to each participant. During this process, the data of each actor is not communicated to the other actors, thus addressing privacy and confidentiality requirements

=>

The adversary is limited to modify his own training data. After performing the attack on his data, he will update his local model accordingly, and share the compromised updates with other collaborative actors

Training-time attacks

Scenario: The adversary is one of the stakeholders of the project dedicated to the development of the machine learning model.

Second way: Online learning
(External users are requested (voluntarily or not) to provide meaningful data, which will be later used to fine-tune the model)

Parameters of the model are continuously updated to allow it to be more efficient for future inputs. For example, users are asked about the pertinence of the output of a recommendation system, to be able to later provide them with more pertinent ones

=> The adversary can then be one of these users, and provide false feedback on the recommendations

Poisoning attacks

QuantifiCare

Principle

Goal: deteriorate the quality of a deployed model

Decrease the accuracy of the target model on specific test set examples

or

for the whole test set

How ?

Injecting new poisoned data in the data set used for training the model

or

Maliciously modifying existing training data

Poisoning attacks

QuantifiCare

Consequence

1. Accuracy decreased on the whole test set => the model is particularly not accurate, preventing it being used in a system.

The model does not satisfy some requirements in terms of task-based performances (e.g. accuracy), defined by requirements of the system it is deployed in,
=> Declared unreliable

2. Accuracy decreased on specific examples => severely harm the decision process of the model

If the model is responsible for identifying anomalies in a process, or detect unauthorized users
=> Anomalies fall between the cracks, or allow an unauthorized user to bypass the detection system

Poisoning attacks

QuantifiCare

Examples

1. Shahafi et al. [19] show that 50 poisoned examples are sufficient to mount an attack on the CIFAR10 data set [24], which makes a model misclassify a specific target example, with 70% success rate.

2. On the same data set, and without access to the target model, Zhu et al. [20] show that, with only poisoning 1% of the training set, an adversary can mount an attack to make the target model misclassify a test set example.

3. Biggio et al. [21] decrease by up to 15% the accuracy of a Support Vector Machine trained to distinguish two classes of the MNIST data set [25], introducing only one poisoned example.

Poisoning attacks

Examples

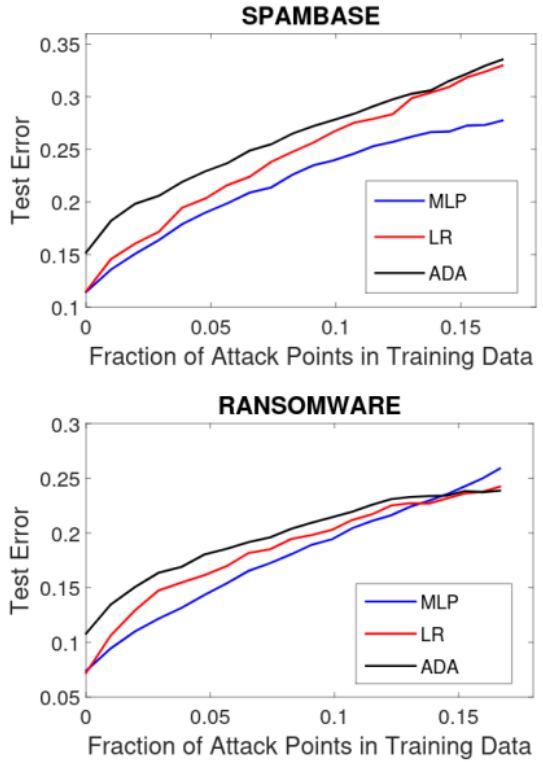


Figure 2.3: Data poisoning attack results (from [22]). For two data sets and different machine learning algorithms is presented the error induced on the test set by perturbing a portion of the training set.

Poisoning attacks

Threat model

Adversary goal

Deteriorate the accuracy on the whole test set or on specific examples

Adversary knowledge

The adversary can intervene at training time

=> Very broad adversarial knowledge

Adversary capacity

How many inputs he can modify / add

Backdoor attacks

QuantifiCare

Principle

Goal: Trick the model into having specific behaviors on some inputs

The model behaves particularly at inference time on inputs in which some specific trigger is included

How ?

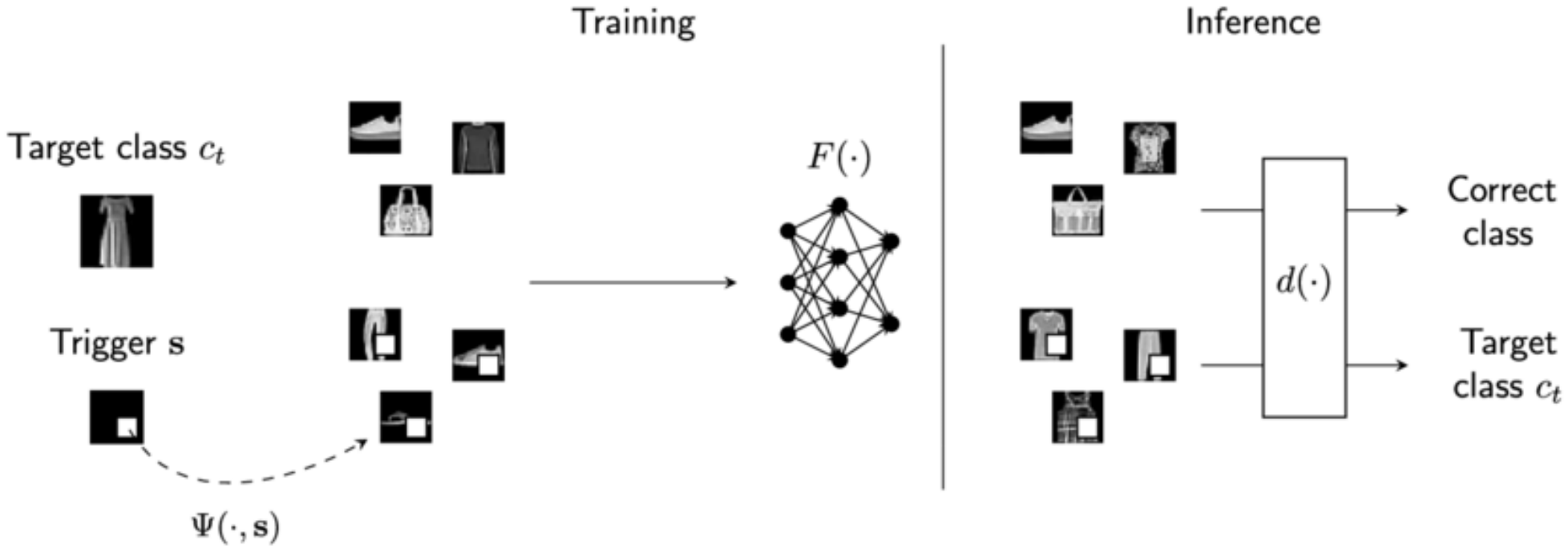
Including in the training data so-called trapdoored input-label pairs

Trapdoored inputs consist of clean inputs containing specific signals on which the model is trained to produce a malicious behavior

Backdoor attacks

Principle

How ? Including in the training data so-called trapdoored input-label pairs



Backdoor attacks

QuantifiCare

Consequence

1. An adversary can precisely induce misclassification at test time

2. When the model is deployed, as long as no malicious input is given to the model, its behavior seems standard

Scenario: a model distinguishes between benign and malicious inputs.
With backdoor attacks, an adversary can simply include a trigger in an input to make the model classify it as benign, and therefore bypassing detection.

Backdoor attacks

Examples

1. Gu et al. [30] perform a backdoor attack against a U.S. road sign detector, which allows at inference that a simple sticker on a "stop" road sign makes the object detection model to mistake it as a "speed limit" road sign.

2. Liu et al. [32] take advantage of the natural reflection phenomenon to create stealthy backdoor inputs. The malicious cue takes the 19 form of a reflection on the image

3. Guo et al. [29] introduce a backdoor attack, which allows an adversary to impersonate a authorized user, for a face recognition system

Backdoor attacks

Examples

Figure 2.4: Backdoor attack (from [30]). The clean image (left), the clean image with the trigger (middle), and the resulting misclassification (right).

Backdoor attack (from [32]).

Backdoor attacks

Threat model

Adversary goal

Trigger a specific inference-time behavior on specific inputs

Adversary knowledge

The adversary can intervene at training time

=> Very broad adversarial knowledge

Adversary capacity

How many inputs he can modify / add

Inference-time attacks

Target:

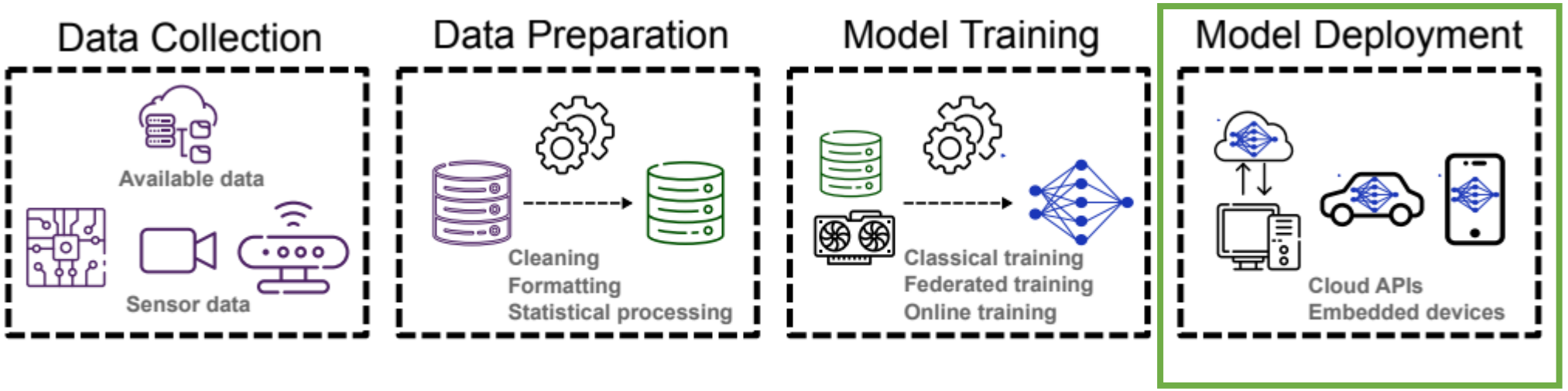


Figure 2.1: Complete machine learning pipeline. Different phases of the development of a model.

Inference-time attacks

Goal:

Provide the deployed model with malicious inputs to induce a specific behavior at inference time.

Very diverse adversaries:

Every actor having access to the deployed model

(Depending on the type of access, details about the machine learning model may be or not available to the adversary)

Inference-time attacks

Adversary knowledge strongly linked with the real-life scenario :

1. An API relying on a machine learning model
   => everyone who is able to use this API is a potential adversary.

   Distant APIs (Google Cloud Vision API[1] or Clarifai[2], internal details of the underlying models are not communicated. => limited to queries

2. A model implemented without obfuscation method on a mobile device
   => any individual having access to the hardware platform (e.g., a 32-bit microcontroller, a typical platform for many IoT domains)

1. https://cloud.google.com/vision/
2. https://www.clarifai.com/

Model inversion

QuantifiCare

Principle

Goal: Retrieve hidden information about sensitive features of input data

How ?

The adversary only needs to be able to get the output response of the machine learning model

and/or

already have some auxiliary knowledge of some non-sensitive parts of the input data, to retrieve a very good approximation of the target sensitive information

=> Malicious queries to the target model

Model inversion

Consequence

Severe threat to confidentiality and privacy

A machine learning model may be trained on sensitive data to perform some task, and then be released on the market

The sensitive-features may concern personal privacy, such as information about the identity of a person, or some of his behaviors

Model inversion

QuantifiCare

Examples

1. Fredriskon et al. [37] consider a model, trained on the picture of people under various conditions (exposure, brightness, with/without glasses etc.) to predict their names. With only the confidence scores corresponding to an input image, an adversary is able to recover a highly resembling image to the one of the input image.

2. Fredrikson et al. [33] consider a linear regression model trained to predict the dose of warfarin based on personal information. Given auxiliary knowledge of the input data, as well as the predicted dose of warfarin, the authors manage to recover sensitive knowledge about genetic markers.

Model inversion

Examples

3. Mehnaz et al. [34] show that an adversary can mount a model inversion attack with high accuracy and precision to reveal personal privacy related information.

General Social Survey (GSS) data set, used to train models that predict the level of happiness of an individual in his marriage
=> an adversary targets the input information "Have you watched X-rated movies in the last year ?", and the proposed attack reaches 61% accuracy.

Adult data set [28], used to train model that predict if an individual earns more than $50K a year
=> an adversary targets the marital status of the individual, and the proposed attack reaches more than 70%

Model inversion

Examples

Figure 2.5: Model inversion attack (from [37]). The clean input (left) and the retrieved image with the model inversion attack (right).

Model inversion

Threat model

Adversary goal

Retrieve information about training data

Adversary knowledge

At worst, the adversary has access to the output of the model.
He may have access to non-sensitive parts of the input data

Adversary capacity

Limit number of queries to the target model

Data extraction

QuantifiCare

Principle

Goal: Retrieve exact parts of the data the model was trained with

How ? Exploit the memorization phenomenon of machine learning models that tend to memorize exact information of the training data

=> Malicious queries to the target model

Memorization occurs at an early stage of the training phase, and label memorization, for example, is needed for good generalization.
Moreover, this overfitting is not a necessary consequence of memorization, as language models, which do not exhibit overfitting, memorize exact text sequence of the training data.

Data extraction

Consequence

Severe threat to confidentiality and privacy

A machine learning model may be trained on sensitive data to perform some task, and then be released on the market

The sensitive-features may concern personal privacy, such as information about the identity of a person, or some of his behaviors

Data extraction

Examples

1. Carlini et al. [41] show that machine learning tend to memorize infrequent data encountered during training. Exploiting this phenomenon, they manage to extract credit card numbers from a data set of emails sent between employees of the Enron Corporation

2. Carlini et al. [42] attack the GPT-2 language model, trained on data scraped from the internet [44], and manage to extract the name, the phone number, the fax number, and the physical address of individuals

Data extraction

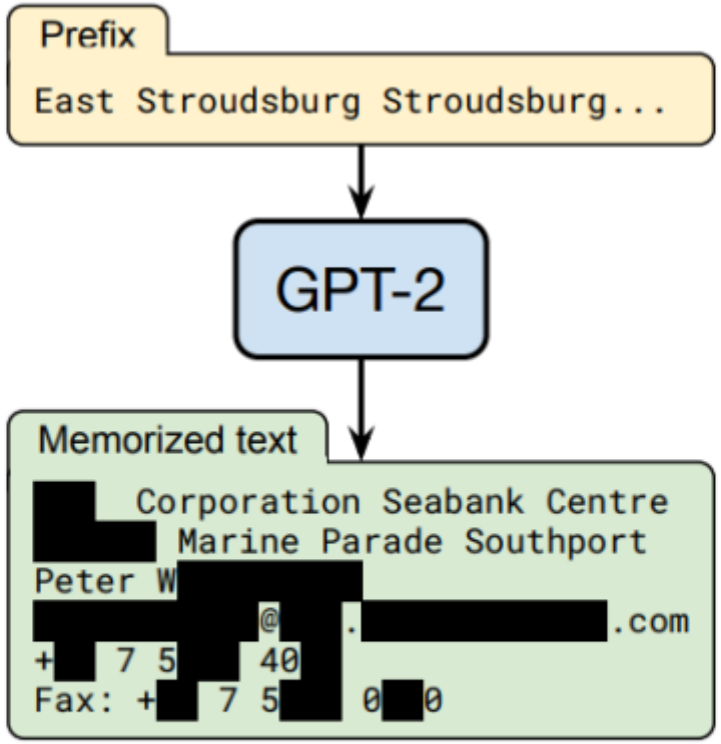


Figure 2.6: Data extraction attack (from [42]). A particular query allows to retrieve the name, the phone number, the fax number, and the physical address of some individual.

Data extraction

Threat model

Adversary goal

Retrieve exact information about training data

Adversary knowledge

At worst, the adversary has access to the output of the model

Adversary capacity

Limit number of queries to the target model

Model stealing

Principle

Goal: Retrieve a machine learning model with high fidelity

How ?

Take advantage of a purely mathematical reasoning

and/or

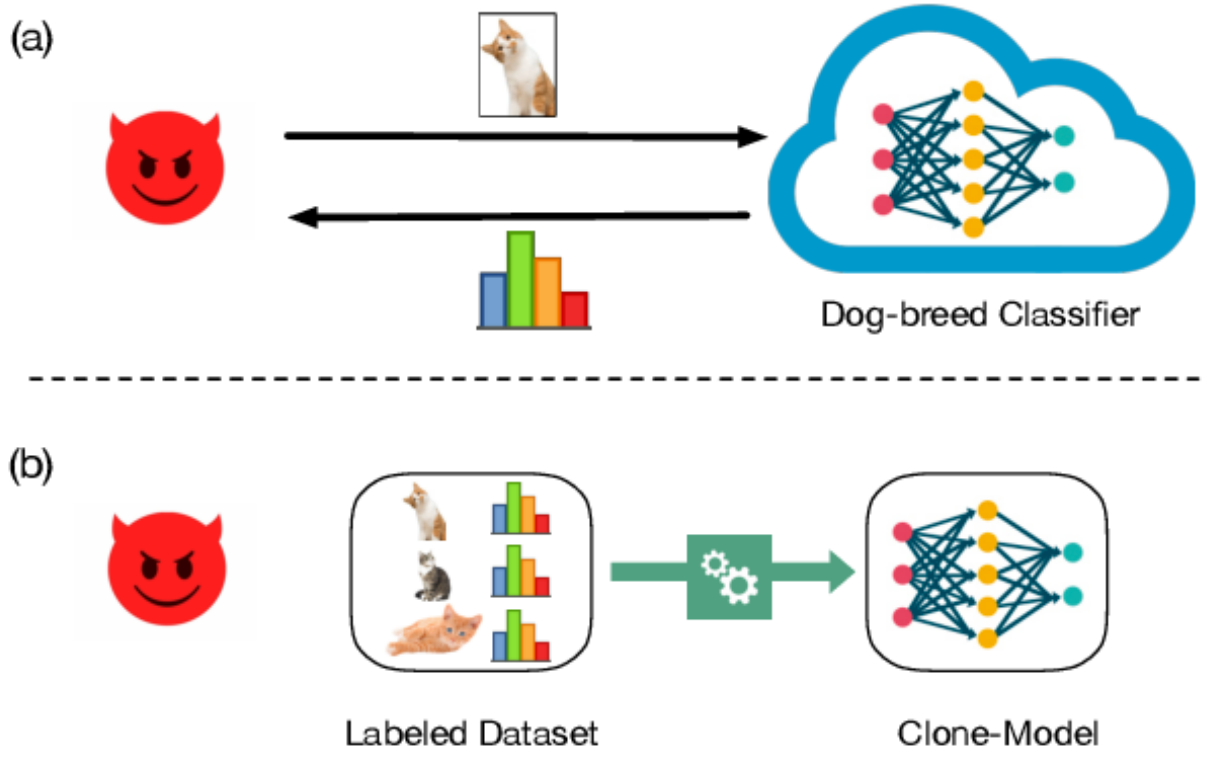
Exploit leakages related to the physical implementation of a model with Side-Channel Analysis (SCA)

(any kind of attack exploiting information collected from the implementation of a system)

Model stealing

Principle

How ? Take advantage of a purely mathematical reasoning



Model stealing

QuantifiCare

Consequence

1. Private property and copyright issues

An adversary can:

i) copy an existing model without bearing the same costs as the model issuers, such as research and development costs and training related expenses
(some actual models come with hundred of millions of parameters which have to be fine-tuned)

ii) use a model for his own purpose without paying for it, in case of the initial model usage was a paid service, or release his copy on the market, causing intellectual property breaches and unfair competition

Model stealing

Consequence

2. Ease the mounting of various types of attacks against a machine learning model

Being able to have a more detailed knowledge of the target model allows to mount more dangerous attack

=> In the case of the target model being hidden to the adversary, model stealing attacks allow him to obtain a local substitute model on which he will fine-tune his attack.
Once he has derived the best way to perform some attacks, he can use it against the distant target model

Model stealing

Examples

1. Jagielski et al. [45] target a two-layer neural network model using the ReLU activation function. Exploiting the way the ReLU function change sign, they manage to completely recover weight values of the first layer of a model. The weights of the second layer can then be determined algebraically.

2. Carlini et al. [46] manage to mount a fast and precise model stealing attack by seeing the model extraction as a cryptanalytic problem.

Model stealing

Examples

3. Batina et al. [14] use SCA methods such as the Correlation Power Analysis (CPA: analysis of power traces of execution) to exploit electromagnetic emanations from an ARM Cortex-M3 microcontroller on which a neural network model is implemented, in order to retrieve important information such as the activation functions used, the number of layers and the weight values, as 32-bit floating point information.

4. Maji et al. [48] also use SCA to exploit timing, electromagnetic and power leakages to perform a model stealing attack for models implemented on two very common micro-processors (ATmega328P, ARM Cortex-M0+) and a RISC-V platform.

Model stealing

Examples

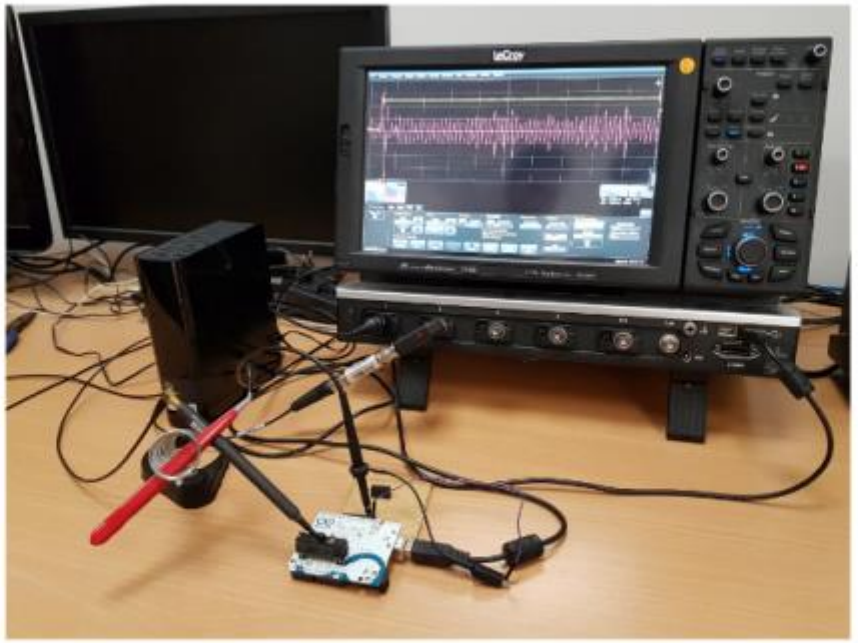


Figure 2.7: Setup of a model stealing attack based on side-channel analysis (from [14]).

Model stealing

Threat model

Adversary goal

Retrieve a machine learning model with high-fidelity

Adversary knowledge

Various types of knowledge.
From strong black-box (SCA) to output only

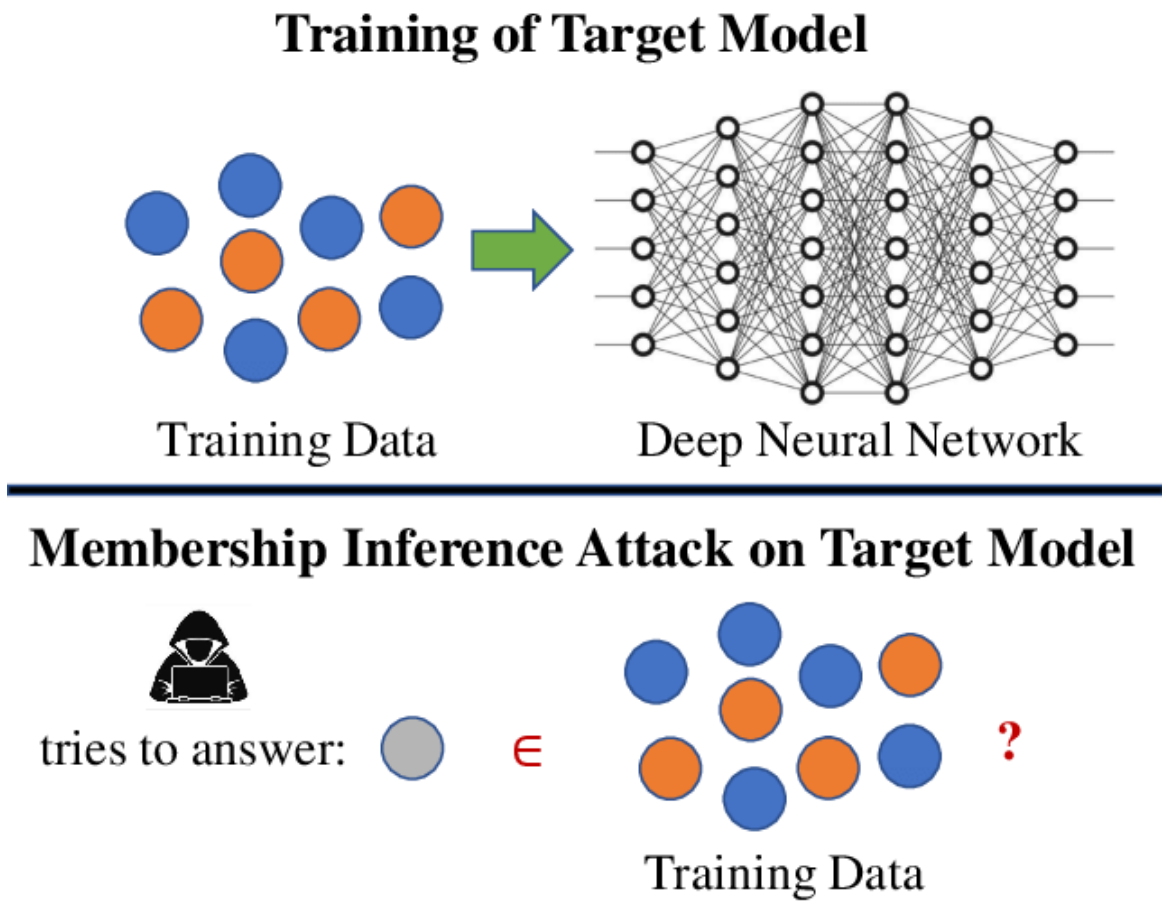
Adversary capacity

Various types of adversarial capacities

Membership Inference

Goal: Determine if a particular record of data was part or not of the training set

Principle



Membership Inference

QuantifiCare

Principle

Goal: Determine if a particular record of data was part or not of the training set

How ? Exploit a model behavior difference between train and test set

Due to overfitting, …

=> Analyze model responses (confidence score, loss value, etc.)

Membership Inference

QuantifiCare

Consequence

Severe privacy disclosure

=> Knowing that an individual was part of the training set may reveal sensitive information about him

A model may have been trained to predict some medication dosage value based on samples of medical data of people suffering from a particular disease. Considering a medical data sample, determining that this sample was used to train the model implies that the person is suffering from the disease

Membership Inference

Examples

1. Yeom et al. [52] establish a strong link between overfitting and the vulnerability of machine learning models against membership inference attacks. Consequently, for many models, an example well predicted may be sufficient to ascertain that is was part of the training set.

2. Leino et al. [54] take advantage of the fact that a model learns features which are only predictive of the training data. Considering an adversary having a complete access to the target model, they mount a membership inference attack by investigating if a model relies on such type of features.

Membership Inference

QuantifiCare

Examples

3. Sablayrolles et al. [55] derive formally the optimal strategy to perform a MIA, and show that the adversary only requires to have access to the loss function value for the target example. Therefore, an adversary with limited information, i.e. only able to access loss function values, is as powerful as an adversary having a complete access to the target model.

4. Shokri et al. [53] perform MIAfor a classification task, which are based on the confidence score vectors given by the target model. In fact, this attack is performed by training a machine learning model for each class, which predicts if an example is part of the training set of the target model, by taking as inputs the confidence score vector given by the target model

Membership Inference

Examples

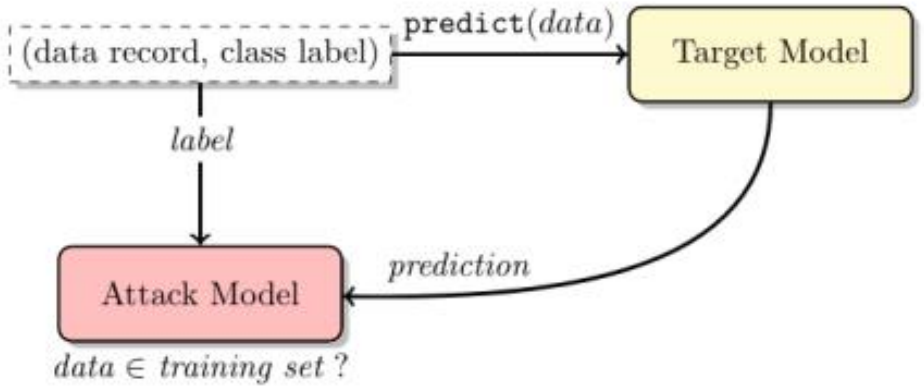


Figure 2.8: Illustration of the attack process from [53]. From the confidence score of the target model, and the ground-truth label of the record, the attack model predicts if the record was part of the training set or not.

Membership Inference

Threat model

Adversary goal

Determine if a particular record of data was part or not of the training set

Adversary knowledge

From strong black-box (label-only) to white-box (architecture, parameters, gradients, …)

Adversary capacity

Number of queries…

Adversarial examples

QuantifiCare

Principle

Goal: Fool the inference process of a machine learning model

How ? Maliciously modifying a clean input, by adding an adversarial perturbation to it, in order to deceive the prediction of a target model

ML models leverage features different than those used by humans when performing predictions, Notably, models can rely on specific signals in an image, to which humans are not sensitive to.

Importantly, this adversarial perturbation has to be imperceptible or appear as benign, in order not to raise suspicion from a human observer.

Adversarial examples

Consequence

Severe menace to the integrity of the prediction scheme of a wide range of systems.

Adversarial examples

QuantifiCare

Examples

1. Szegedy et al. [11] firstly introduce a method to craft adversarial examples against a neural network model by searching for a small perturbation maximizing the classification loss.

2. Engstrom et al. [60] show that an adversary can exploit malicious rotations and translations in an image classification task to craft adversarial examples

3. Shamsabadi et al. [61] propose an attack method where adversarial examples are crafted by identifying parts of an image, where color modification does not raise suspicion, and subsequently maliciously modify their colors to fool a machine learning model.

Adversarial examples

Examples

4. Su et al. [62] leverage differential evolution strategies to craft adversarial examples by only modifying one pixel in an image.

5. Papernot et al. [64] introduce the phenomenon of transferability, which constitutes a powerful tool for an adversary with a constrained access to the target model.
=> An adversarial example crafted on a substitute (source) model can fool another machine learning model (target). Importantly, the machine learning models do not need to be of the same type.

Adversarial examples

Examples

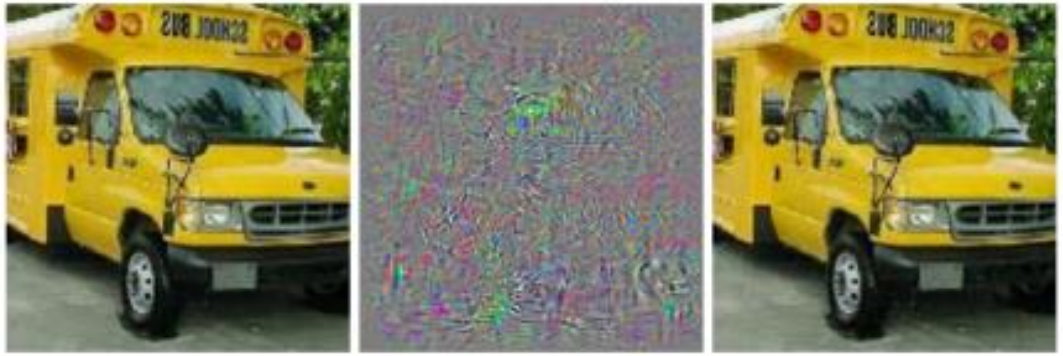


Figure 2.9: Illustration of adversarial examples from [11]. Clean image classified as a "school bus"(left), magnified adversarial perturbation (middle), and resulting adversarial example misclassified by the target model as "ostrich" (right).

Adversarial examples

QuantifiCare

Threat model

Adversary goal

Fool the inference process of a machine learning model

Adversary knowledge

From strong black-box (label-only) to white-box (architecture, parameters, gradients, …)

Adversary capacity

How and how much the adversary can distort the clean input