УНИВЕРСИТЕТ ИТМО

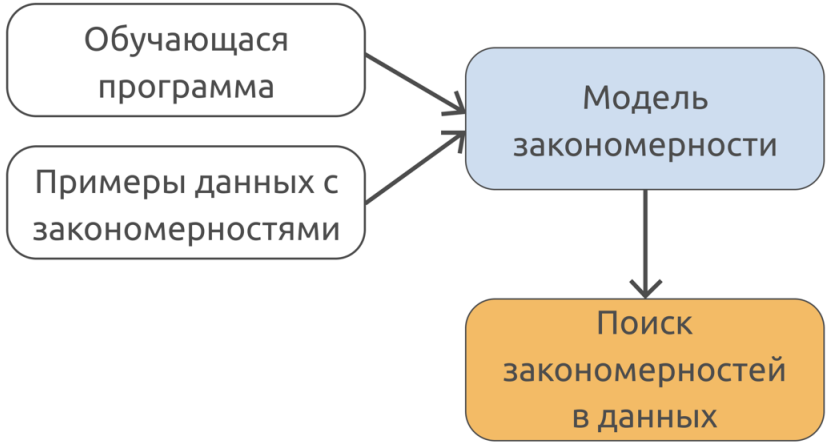
Атаки на системы искусственного интеллекта
Лекция 2. Таксономия атак на системы искусственного интеллекта

Воробьева Алиса Андреевна
vorobeva@itmo.ru
23.05.2022

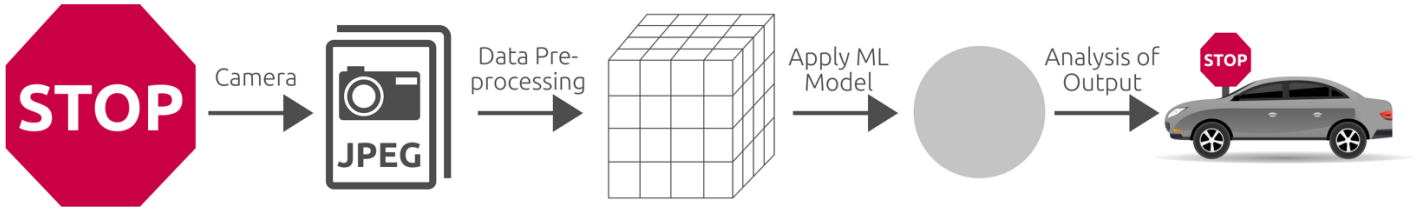
ОБРАЗОВАТЕЛЬНЫЕ ПРОГРАММЫ В ОБЛАСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Последовательность операций во время эксплуатации системы МО

1. сбор входных данных от сенсоров или из хранилищ данных;
2. передача данных в цифровой домен;
3. обработка трансформированных данных с помощью модели машинного обучения для получения выходного сигнала;
4. действия, предпринятые на основе выходных данных.



Система распознавания дорожных знаков беспилотного автомобиля



Исходные данные = основа МО

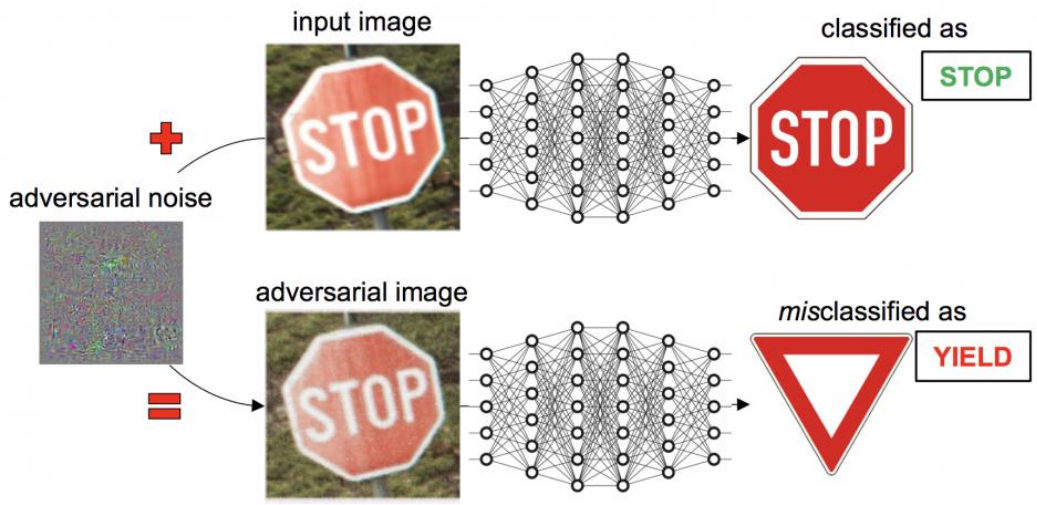
Но так ли легко получить доступ к обучающим выборкам? На самом деле, да. Многие вендоры обмениваются информацией об угрозах через специальные потоки данных об угрозах (threat intelligence feeds), и уже известны случаи, когда злоумышленники манипулировали такими механизмами. Например, агрегатор данных об угрозах VirusTotal атаковали специально созданными чистыми файлами с признаками вредоносов. После того, как один из антивирусных сканеров определяет такой файл как вредоносный, эту ошибочную классификацию начинают использовать другие системы безопасности, что вызывает цепную реакцию ложных срабатываний по всему миру – схожие чистые файлы детектируются как вредоносы.

kaspersky

Adversarial machine learning

Adversarial machine learning – это боковая ветвь машинного обучения, ставшая основой для разработки инструментов, которые могут создать помехи в работе системы, основанной на алгоритмах машинного обучения.

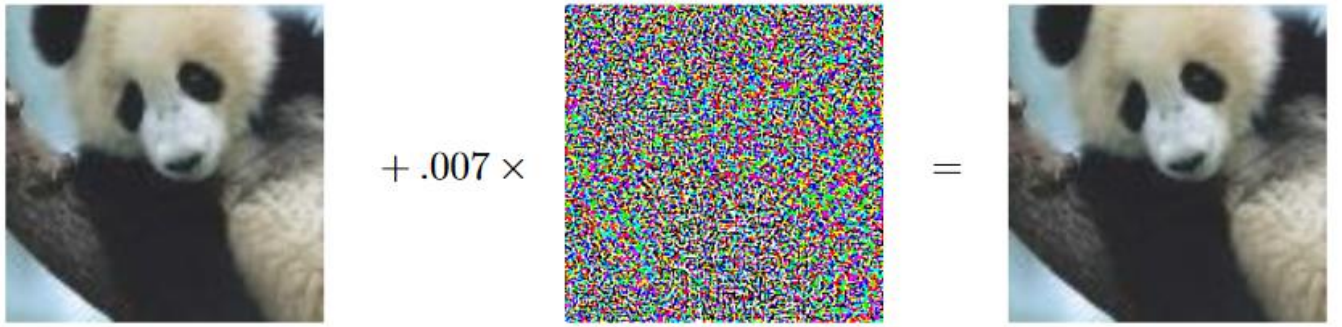
Adversarial machine learning is the study of the attacks on machine learning algorithms, and of the defenses against such attacks.



Adversarial machine learning

Состязательная атака (adversarial attack) —способ обмануть нейронную сеть с целью изменения «ответа» системы на необходимый злоумышленнику.

Состязательный пример (adversarial sample) — некий пример тестовых данных, в который внесены искажения, приводящие к некорректному распознаванию.



Злоумышленник в системах ИИ

Злоумышленники могут иметь **различный уровень доступа** к системе и разный уровень знаний о ней.

Злоумышленники могут преследовать **различные цели:**

- снижение точности работы модели, которая ведет к снижению достоверности классификации;
- неправильная классификация, при которой модель будет неверно определять классы;
- целевая неправильная классификация, которая заставляет модель определять объекты как класс, заранее выбранный злоумышленником;
- неправильная классификация источника и цели, при которой все объекты определенного класса будут классифицироваться как другой выбранный злоумышленником класс.

Уязвимости систем ИИ

Уязвимость – это свойство системы, ее недостаток или слабость, которая может быть использована для реализации угроз безопасности информации.

Слабости и уязвимости систем, основанных на искусственном интеллекте:

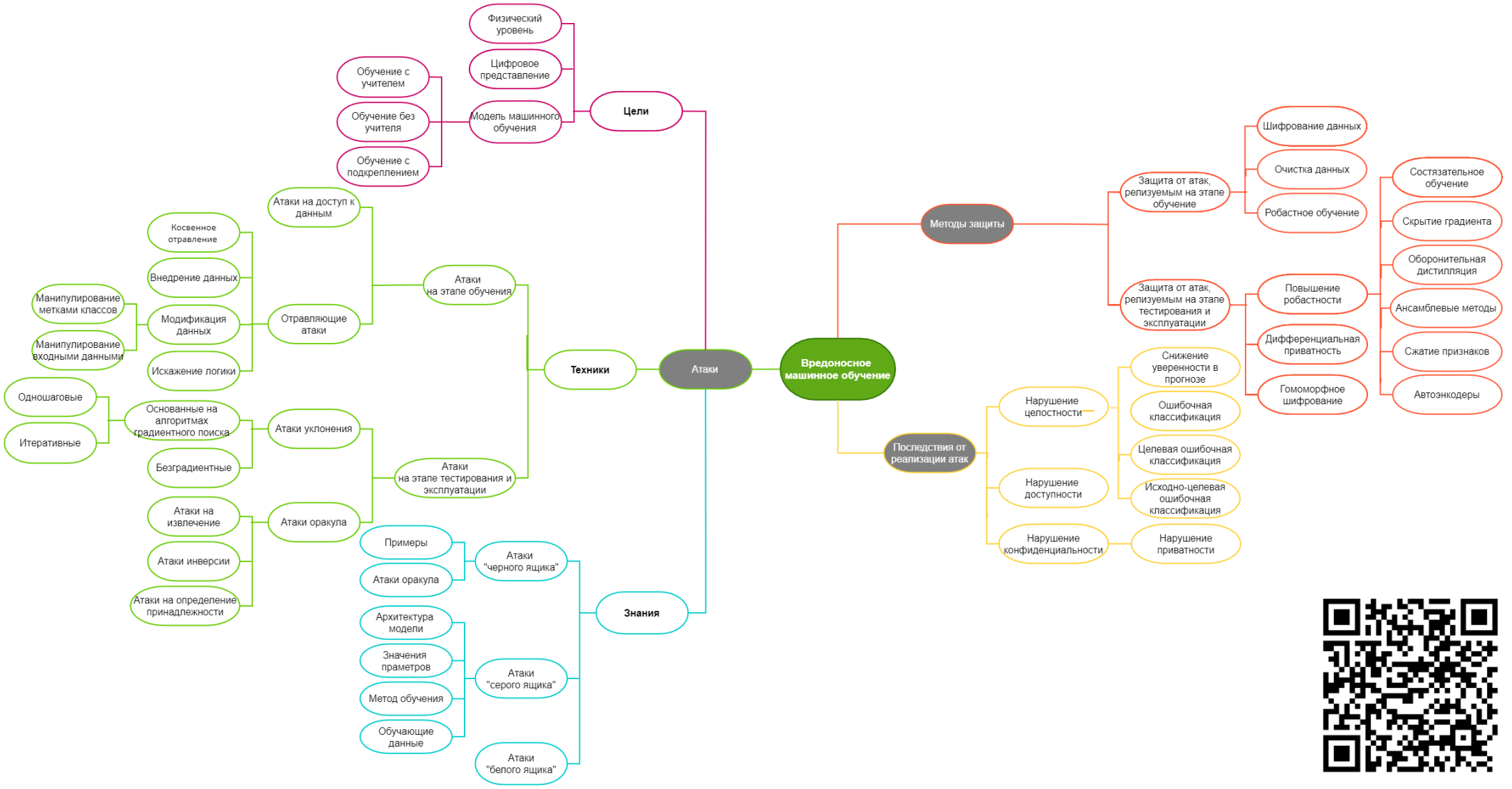
- обучающая выборка принципиально не может отражать всех данных генеральной совокупности;
- процесс обучения модели и дальнейшего принятия ею решений скрыт от разработчика и оператора системы;
- параметры обученной модели отражают информацию о данных обучающей выборки, что делает потенциально возможным их получение;
- сравнительно легко получить «теневую», или суррогатную модель за счет изучения защищаемой модели, как «черного ящика».

Цели злоумышленников

1. алгоритмы получения признаков из результатов измерений и сами результаты измерений;
2. алгоритмы обучения модели и значения гиперпараметров;
3. значения параметров обученной модели;
4. доверительные вероятности принимаемых решений и сами принимаемые классификатором решения;
5. полученная в ходе обучения граница принятия решений (гиперплоскость в n-мерном пространстве).

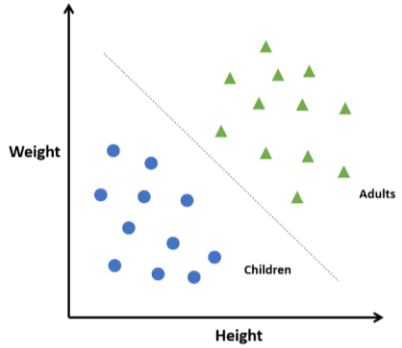
Компоненты машинного обучения могут быть **целями атак** злоумышленников, использующих различные **техники и обладающих различными знаниями** о системе.

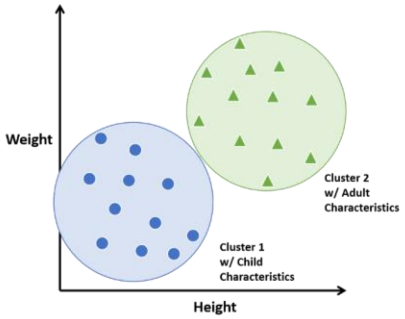
Таксономия атак от NIST

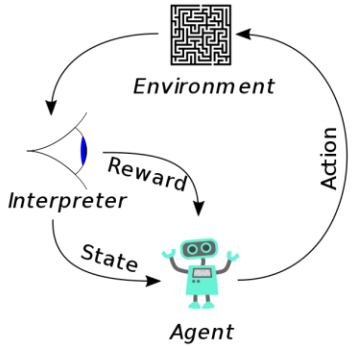


Классификация по целям атаки

УНИВЕРСИТЕТ ИТМО



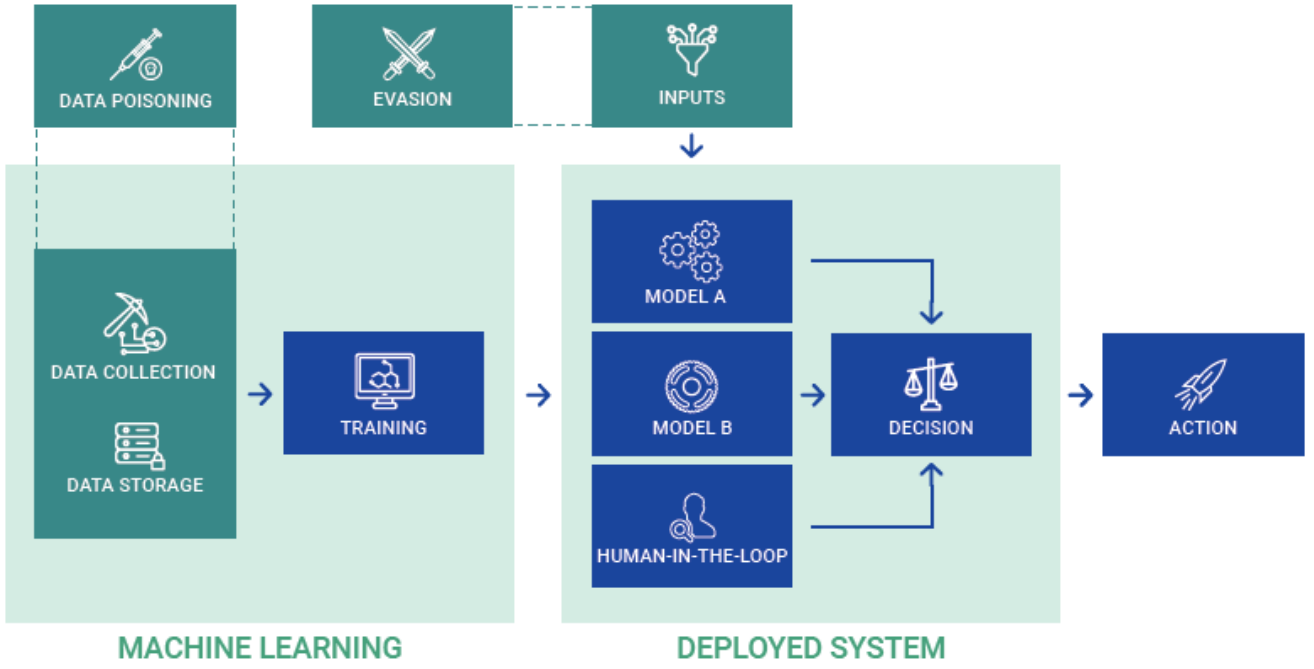




- Цели
  - Модель машинного обучения
    - Физический уровень
    - Цифровое представление
    - Обучение с учителем
    - Обучение без учителя
    - Обучение с подкреплением

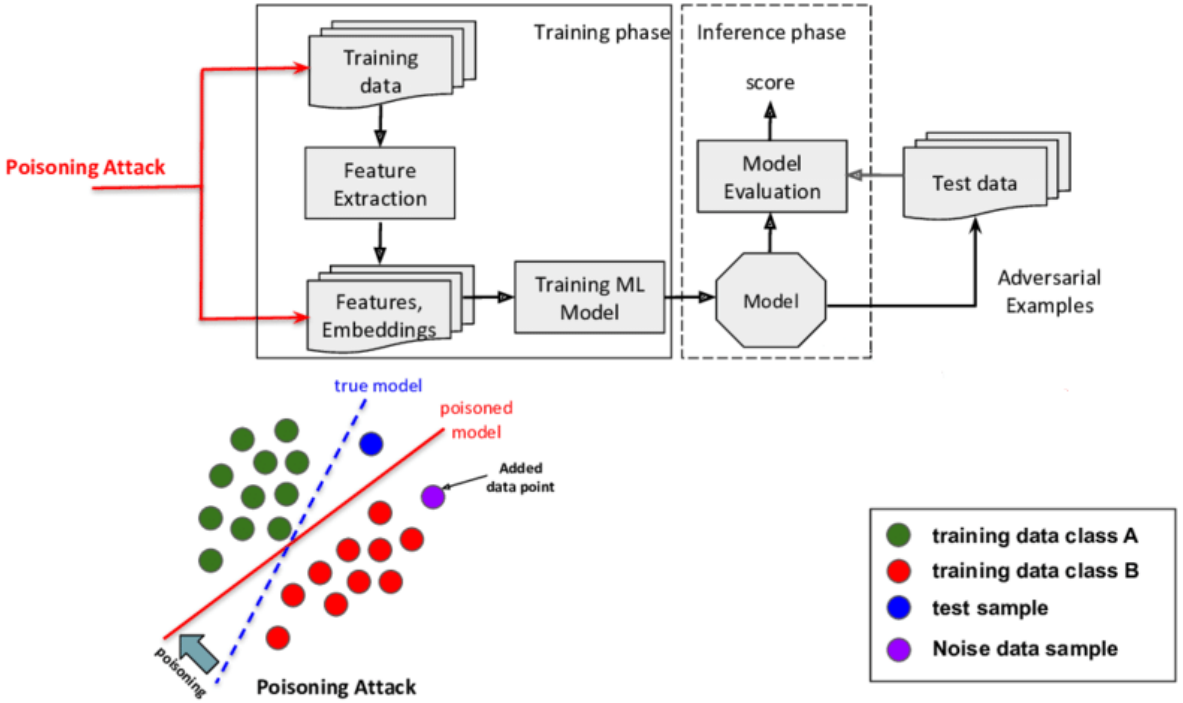
Классификация по техникам, используемым злоумышленниками

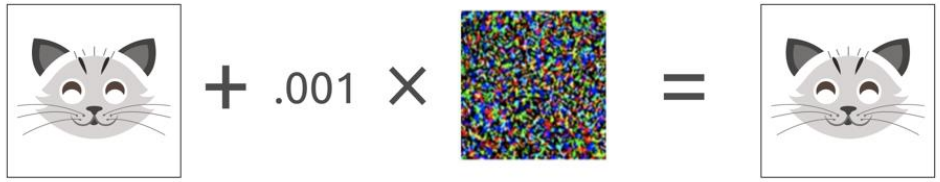
- Атаки во время обучения,
- Атаки на этапе эксплуатации или тестирования.



Атаки во время обучения

Злоумышленник пытается влиять на модель путем непосредственного изменения обучающего набора данных.





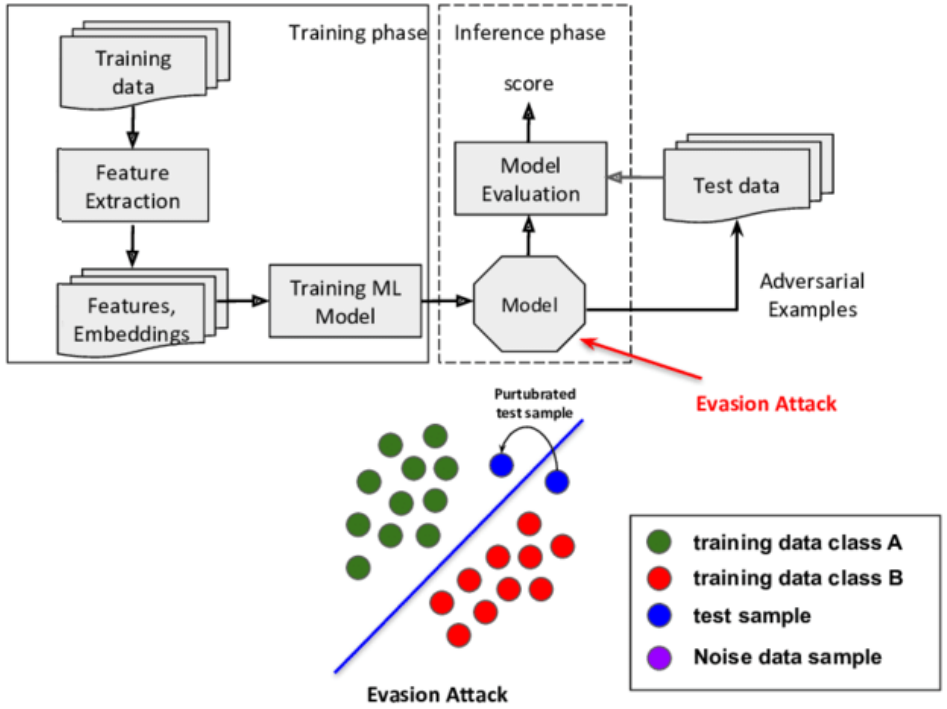
Кот

Состязательная атака

Собака

Атаки на этапе тестирования (эксплуатации)

Злоумышленник не вмешивается ни в данные, ни в целевую модель, а заставляют ее выдавать некорректный результат.



Метод черного ящика

Метод серого ящика

Метод белого ящика

Классификация атак в зависимости от знаний злоумышленника о системе

Метод черного ящика	Метод серого ящика	Метод белого ящика
		
Эмуляция действий нарушителя без знания системы	Эмуляция действий нарушителя с ограниченным знанием системы	Эмуляция действий нарушителя со знанием системы
Минимальный уровень угрозы	Средний уровень угрозы	Высокий уровень угрозы

Атаки белого ящика

УНИВЕРСИТЕТ ИТМО

Злоумышленник обладает полной информацией о модели:

- **и** тип используемой нейронной сети,
- **и** параметры и количество слоев,
- **и** алгоритм, применяемый в процессе обучения (в частности, оптимизатор градиентного спуска).
- **и** информация о параметрах обученной модели.

Злоумышленник использует информацию для определения пространства признаков, в котором модель может быть уязвима, то есть участков, где модель имеет высокую частоту появления ошибок.

Затем модель атакуется с применением методов генерации состязательных примеров.

Метод белого ящика

Эмуляция действий нарушителя со знанием системы

Высокий уровень угрозы

Атаки серого ящика

Злоумышленнику частично известна информация о системе:

- **или** тип используемой нейронной сети,
- **или** параметры и количество слоев,
- **или** алгоритм, применяемый в процессе обучения (в частности, оптимизатор градиентного спуска).
- **или** информация о параметрах обученной модели.

Метод серого ящика

Эмуляция действий нарушителя с ограниченным знанием системы

Средний уровень угрозы

Атаки черного ящика

Злоумышленнику ничего не известно о системе.

- Создание собственной суррогатной модели, имитирующей целевую модель.
- Подбор состязательных примеров для суррогатной модели.
- Атака состязательными примерами целевой модели.

Метод черного ящика

Эмуляция действий нарушителя
без знания системы

Минимальный уровень угрозы

В **основе** атак черного ящика лежит сильное **свойство переносимости в нейронных сетях**, то есть, состязательные примеры, полученные на одном классификаторе, приведут другой классификатор к той же ошибке

Виды атак черного ящика

УНИВЕРСИТЕТ ИТМО

Неадаптивные
Non-adaptive

Адаптивные
Adaptive

Точечные
Strict

Цель - создание суррогатной модели

- Неадаптивные - имеется полный доступ к обучающим наборам данных целевой модели, которые он использует для обучения собственной модели
- Адаптивные – нет доступа к обучающим данным и прямого доступа к модели, но есть **возможность подать на вход различные тестовые примеры и анализировать выходные данные**.
- Точечные – нет доступа к обучающим данным и прямого доступа к модели, нет возможности вводить собственные тестовые примеры, **может анализировать пары «ввод-вывод» из целевого классификатора.**

Метод черного ящика

Эмуляция действий нарушителя
без знания системы

Минимальный уровень угрозы

Классификация атак по техникам злоумышленника

Атаки во время обучения

Атаки на доступ к данным (data access)

Злоумышленник имеет полный или частичный доступ к обучающему набору данных, поэтому они могут создать собственную суррогатную модель.

Суррогатная модель далее используется для создания и проверки эффективности вредоносных примеров, которые далее подаются на вход целевой модели (атака на этапе тестирования).

Атаки во время обучения

Отравляющие (poisoning или causative) атаки

Злоумышленник модифицирует набор обучающих данных или же саму модель, так чтобы результирующая модель обладала необходимыми ему свойствами.

Виды атак:

- косвенное отравление (indirect poisoning),
- внедрение данных (data injection),
- модификация данных (data manipulation),
- искажение логики (logic corruption).

Атаки во время обучения

Отравляющие (poisoning или causative) атаки.

Косвенное отравление (indirect poisoning)

Злоумышленник не имеет доступа к предварительно обработанным данным, используемым целевой моделью, и должен произвести отравление данных перед их предварительной обработкой.

Внедрение данных (data injection)

Злоумышленник не имеет доступа к целевой модели, а также к обучающим данным, но он может скомпрометировать модель, вставляя вредоносные примеры в обучающий набор данных.

Атаки во время обучения

Отравляющие (poisoning или causative) атаки.

Модификация данных (data manipulation)

Злоумышленник не имеет доступа к целевой модели, но имеет полный доступ к обучающим данным.

Он может изменить целевую модель, модифицируя данные до того, как они будут использованы для обучения.

Атаки:

- манипулирование или модификация меток (label manipulation)
- манипулирование входными данными или input manipulation

Атаки во время обучения

Отравляющие (poisoning или causative) атаки.

Искажение логики (logic corruption)

Злоумышленник может вмешиваться в алгоритм машинного обучения и тем самым изменять процесс обучения и саму модель.

Атаки во время тестирования и эксплуатации

Злоумышленник не вмешивается ни в данные, ни в целевую модель, а заставляют ее выдавать некорректный результат.

Эффективность реализации атак зависит от количества информации о модели, доступной для злоумышленника.

Задача злоумышленника - поиск уязвимостей в обученной модели.

Цели злоумышленника:

- Атаки уклонения (evasion) - поиск вредоносных примеров, на которых модель ошибается.
- Атаки оракула (Oracle) - получение информации о модели или наборе обучающих данных.

Атаки во время тестирования и эксплуатации

Атаки уклонения (evasion)

Поиск вредоносных примеров, которые содержат незначительные искажения и практически не отличаются от обычных примеров.

Атаки:

- Основанные на алгоритмах градиентного поиска:
  - атаки методом быстрого градиента (FGSM),
  - атаки с алгоритмом Бройдена-Флетчера-Гольдфарба-Шанно с ограниченной памятью (L-BFGS),
  - атаки методом карт значимости (JSMA)
- Безградиентные - требуют, знания доверительных вероятностей принимаемых классификатором решений.

Атаки во время тестирования и эксплуатации

Атаки оракула (Oracle)

Атаки, нацеленные нарушение на конфиденциальности модели.

Атаки:

- атаки на извлечение (extraction)
- атаки инверсии (inversion)
- атаки на определение принадлежности (membership inference).

Атаки во время тестирования и эксплуатации

Атаки оракула (Oracle).

Атаки на извлечение (extraction)

Злоумышленник стремится извлечь параметры или получить знания об атакуемой модели, основываясь анализе результатов наблюдений за прогнозами модели. Получение значений вероятностей, возвращаемых для каждого из классов.

Атаки инверсии (inversion)

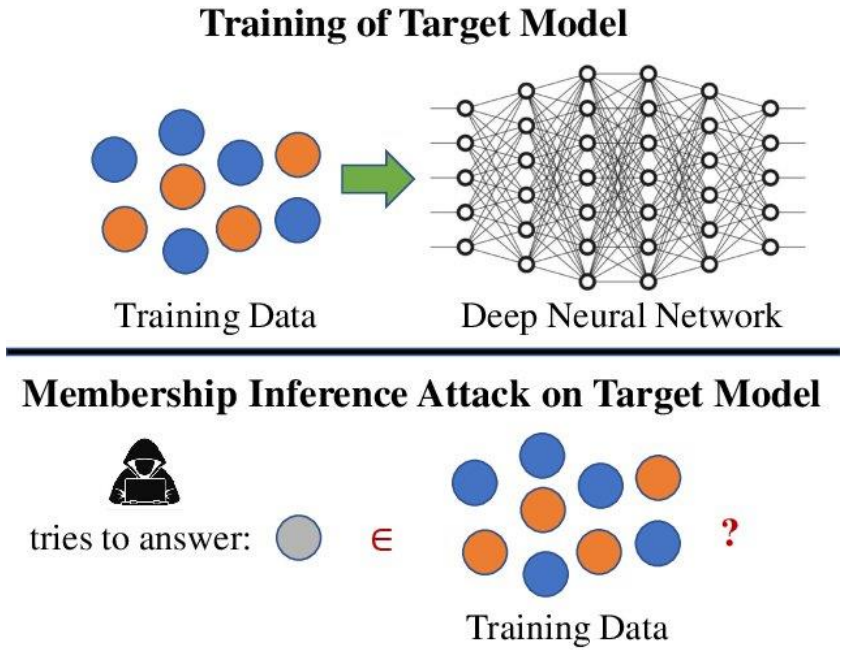
Злоумышленник стремится восстановить данные, которые были использованы для обучения модели.

Атаки во время тестирования и эксплуатации

Атаки оракула (Oracle).

Атаки на определение принадлежности (membership inference)

Целью злоумышленника является получение знаний о том, был ли конкретный пример использован для обучения модели.



УНИВЕРСИТЕТ ИТМО

Спасибо за внимание!

ОБРАЗОВАТЕЛЬНЫЕ ПРОГРАММЫ В ОБЛАСТИ
ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА