Искусственный интеллект и кибербезопасность

с.н.с. Лаборатории ОИТ кафедры ИБ

д.т.н. Д.Е. Намиот

dnamiot@gmail.com

Содержание

- ИИ = системы машинного обучения. Иногда еще уже - ИНС
- Различные аспекты пересечений искусственного интеллекта и кибербезопасности: прикладные задачи машинного обучения (атаки и защита), атаки и защита для систем машинного обучения

Литература

- Намиот, Д. Е., Ильюшин, Е. А., & Чижов, И. В. (2022). Искусственный интеллект и кибербезопасность. International Journal of Open Information Technologies, 10(9), 135-147.
- Намиот, Д. Е., & Ильюшин, Е. А. (2022). Об устойчивости и безопасности систем искусственного интеллекта. International Journal of Open Information Technologies, 10(9), 126-134.

Направления

- Повышение кибербезопасности с помощью ИИ (использование ИИ в кибербезопасности)
- Кибератаки с использованием ИИ (использование ИИ для усиления кибератак)
- Кибербезопасность систем ИИ (атаки на системы ИИ)
- Использование ИИ в злонамеренных информационных операциях (фейки с использованием ИИ)

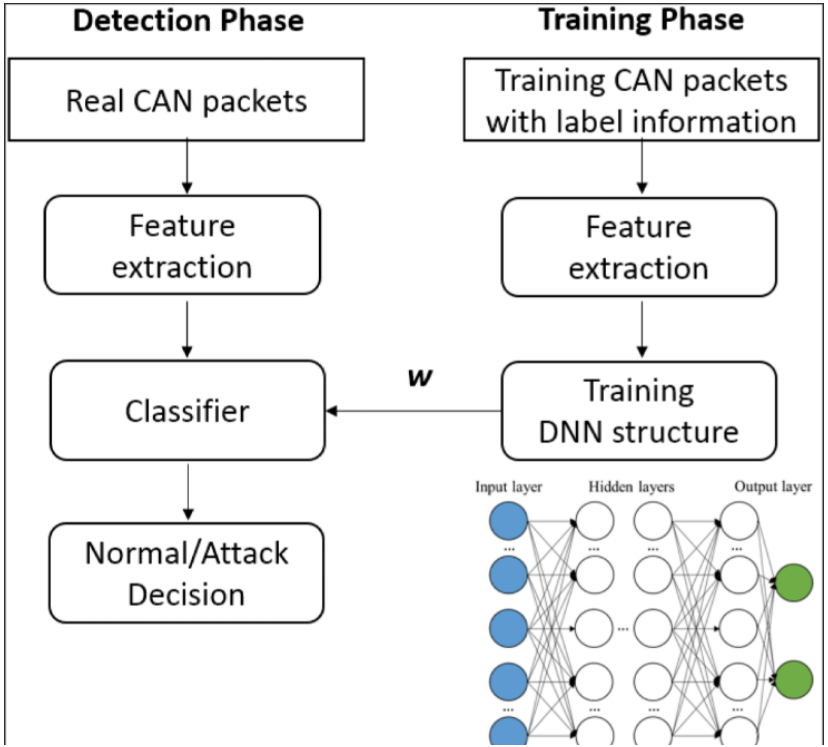
Использование ИИ в кибербезопасности

- Фактически: прикладные задачи машинного обучения
- Общий современный тренд: при отсутствии аналитических моделей собираем все, что можем измерить и пытаемся найти зависимости с помощью машинного обучения
- Google Scholar: ML for malware detection - 25 000+ результатов

Использование ИИ в кибербезопасности

Analysis Type	Feature Extraction Method	Features Extracted
Static	Manifest analysis	Package name, Permissions, Intents, Activities, Services, Providers
	Code analysis	API calls, Information flow, Taint tracking, Opcodes, Native code, Cleartext analysis
Dynamic	Network traffic analysis	URLs, IPs, Network Protocols, Certificates, Non-encrypted data
	Code instrumentation	Java classes, intents, network traffic
	System calls analysis	System calls
	System resources analysis	CPU, Memory, and Battery usage, Process reports, Network usage
	User interaction analysis	Buttons, Icons, Actions/Events

Использование ИИ в кибербезопасности



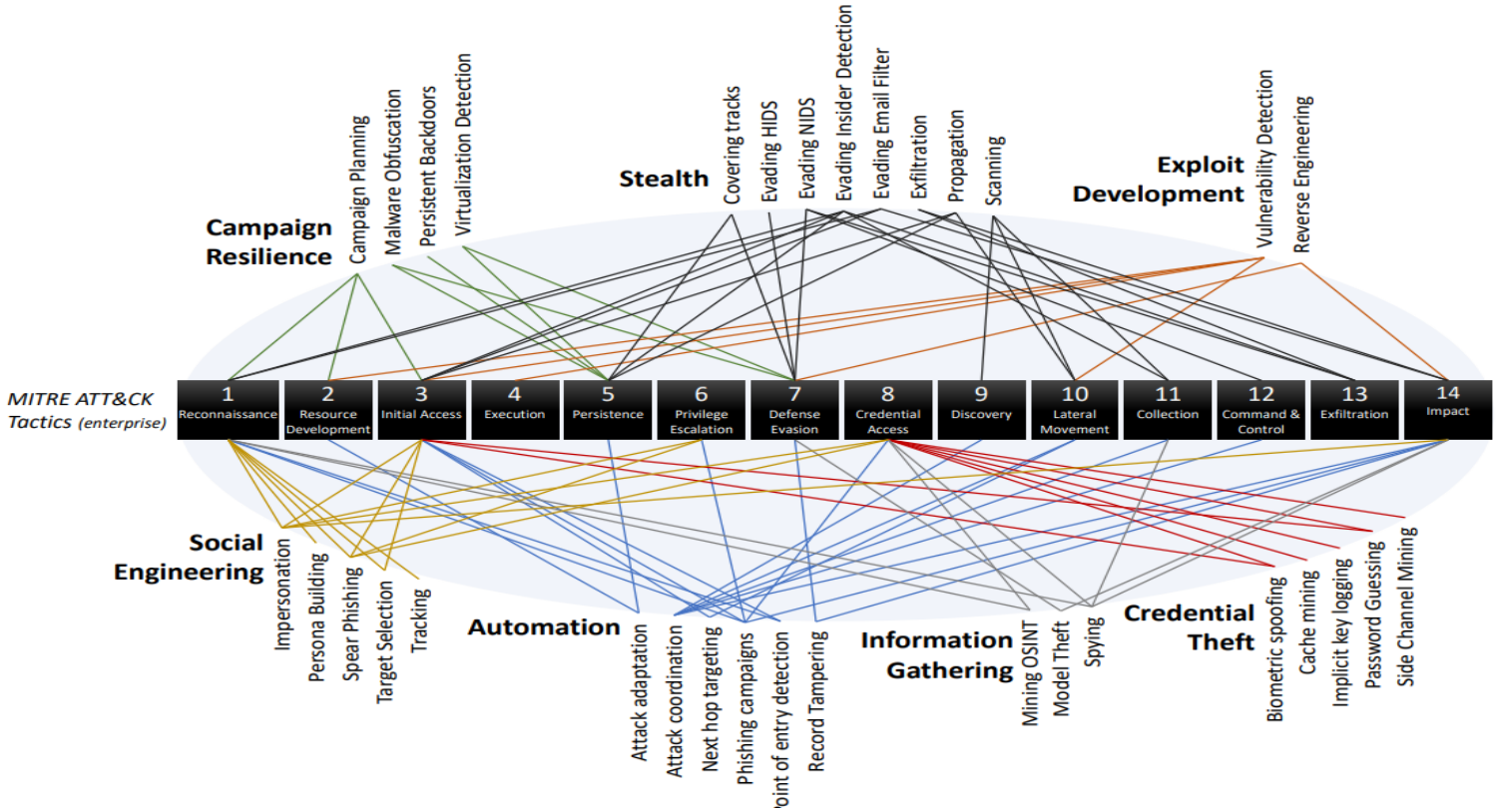
- Типичный пример: автомобильная сеть (CAN)

- Features – характеристики пакета (включая payload)
- Сеть обучается на некотором размеченном наборе данных
- Работает как классификатор
- False positive – общая проблема для такого рода систем

Использование ИИ в кибербезопасности

- Network Intrusion Detection (NIDS)
- Аномалии в сессиях/логах
- Фишинг и спам
- Мошенничество (Fraud detection)
- Атрибутирование атак (DARPA)
- Фактически: все, где можно собрать размеченный датасет
- Это объекты для атак на системы ИИ!

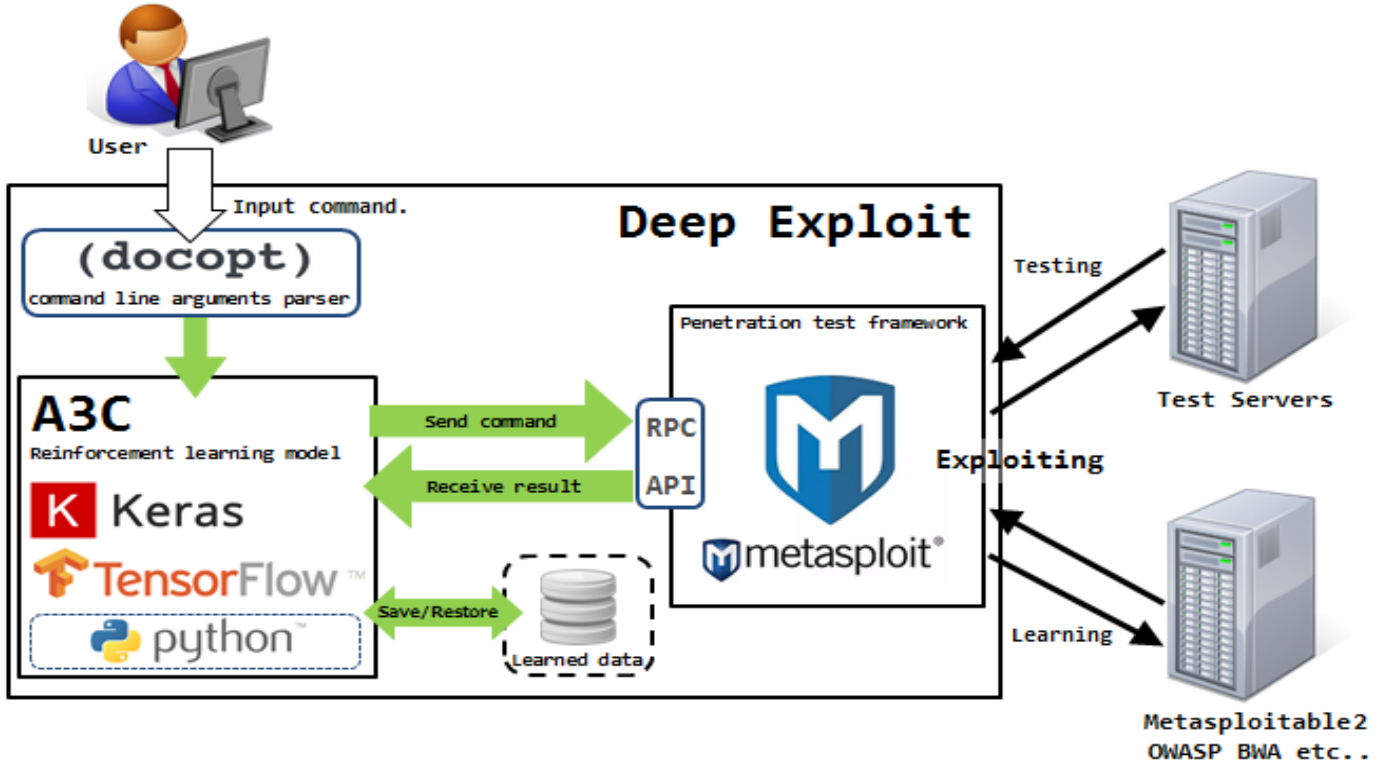
Наступательный ИИ



Наступательный ИИ

- Отчет National Security Commission on Artificial Intelligence (2021)
- Подбор паролей
- Поиск “слабого звена” в социальных сетях
- Модификация трафика (traffic-space attacks)
- Автоматизация пентеста
- Генерация фишинговых атак
- Боты

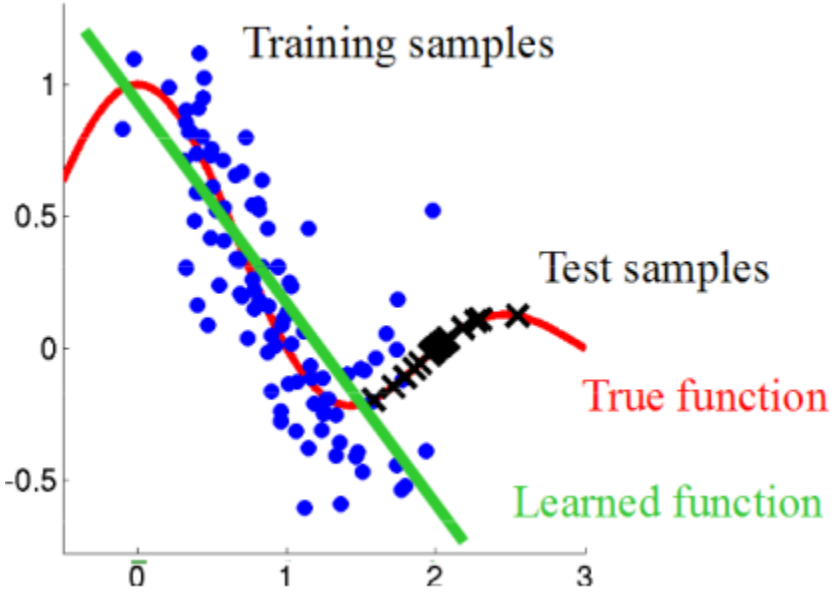
Наступательный ИИ

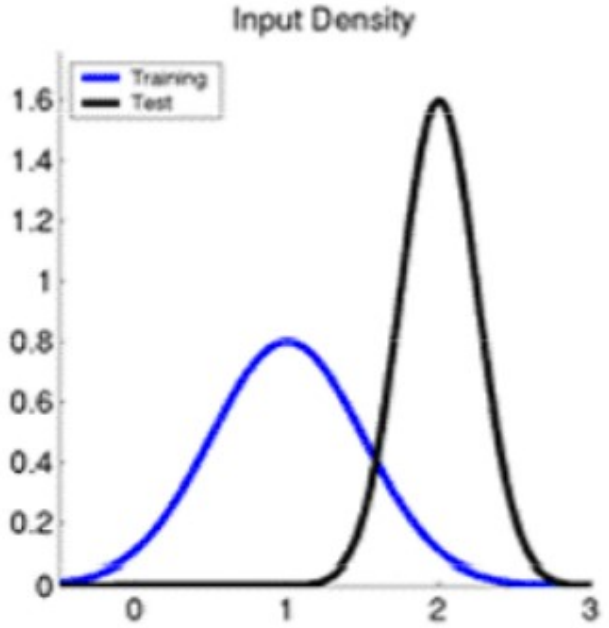


Кибербезопасность систем ИИ

- Системы машинного обучения (как это не удивительно !) зависят от данных
- Всегда при обучении есть только некоторое подмножество генеральной совокупности
- Изменение данных (на любых этапах конвейера) ведет к изменению работы модели
- Структура моделей позволяет также получить информацию о ее работе

Ковариационный сдвиг

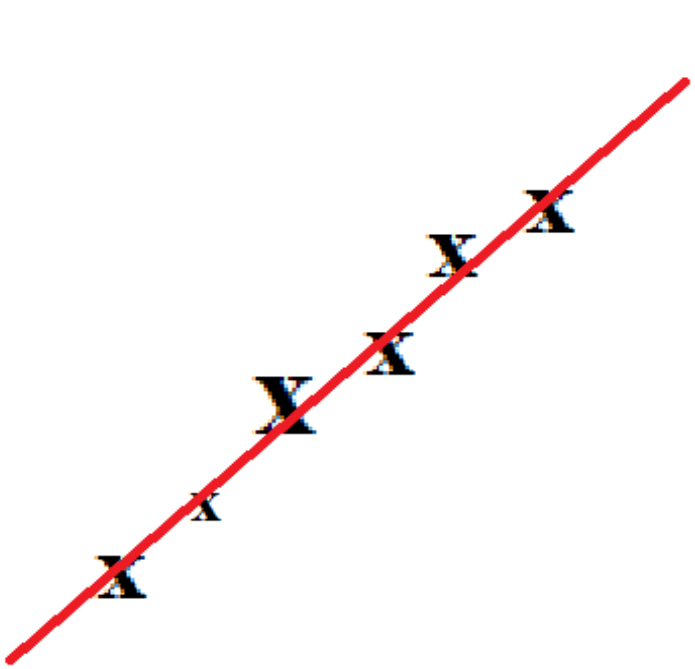


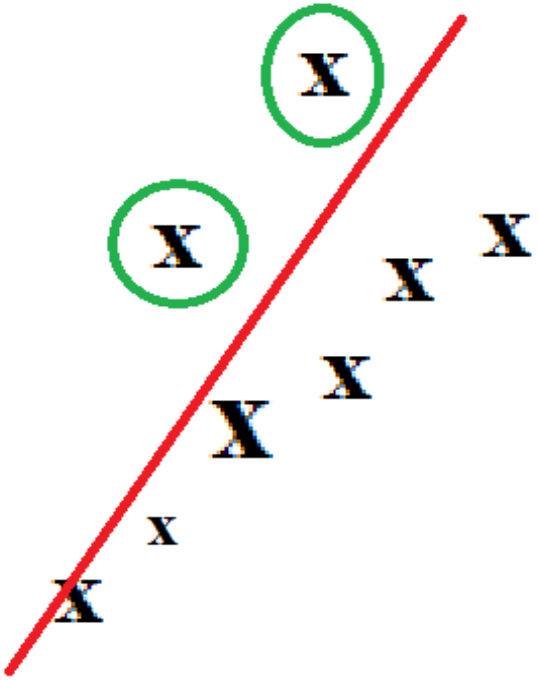


Сдвиг данных

- Это самая простая форма отличия реальных данных от тренировочного набора
- Самые большие проблемы – сдвиг концепции. Изменение связей между входными и выходными переменными
- Строили модель посещения кафе по историческим данным, а COVID изменил поведение посетителей

Проблемы с исходными данными

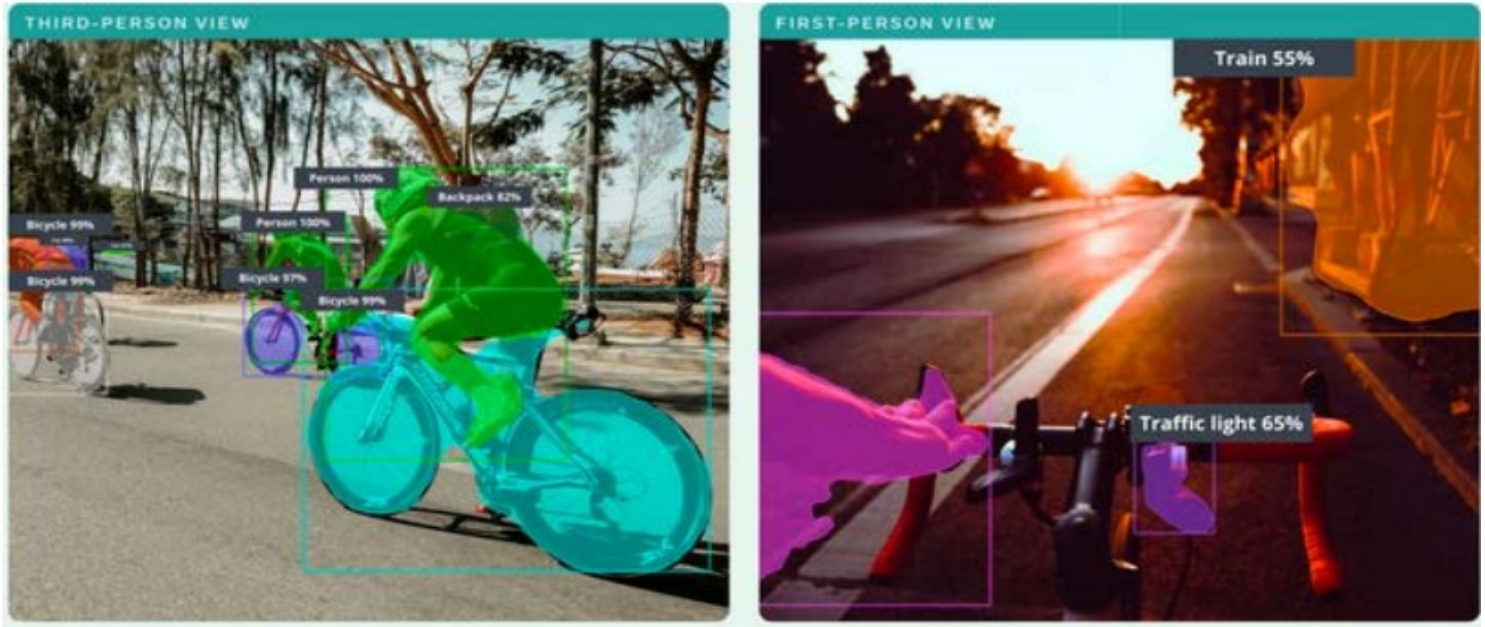




Проблемы с исходными данными

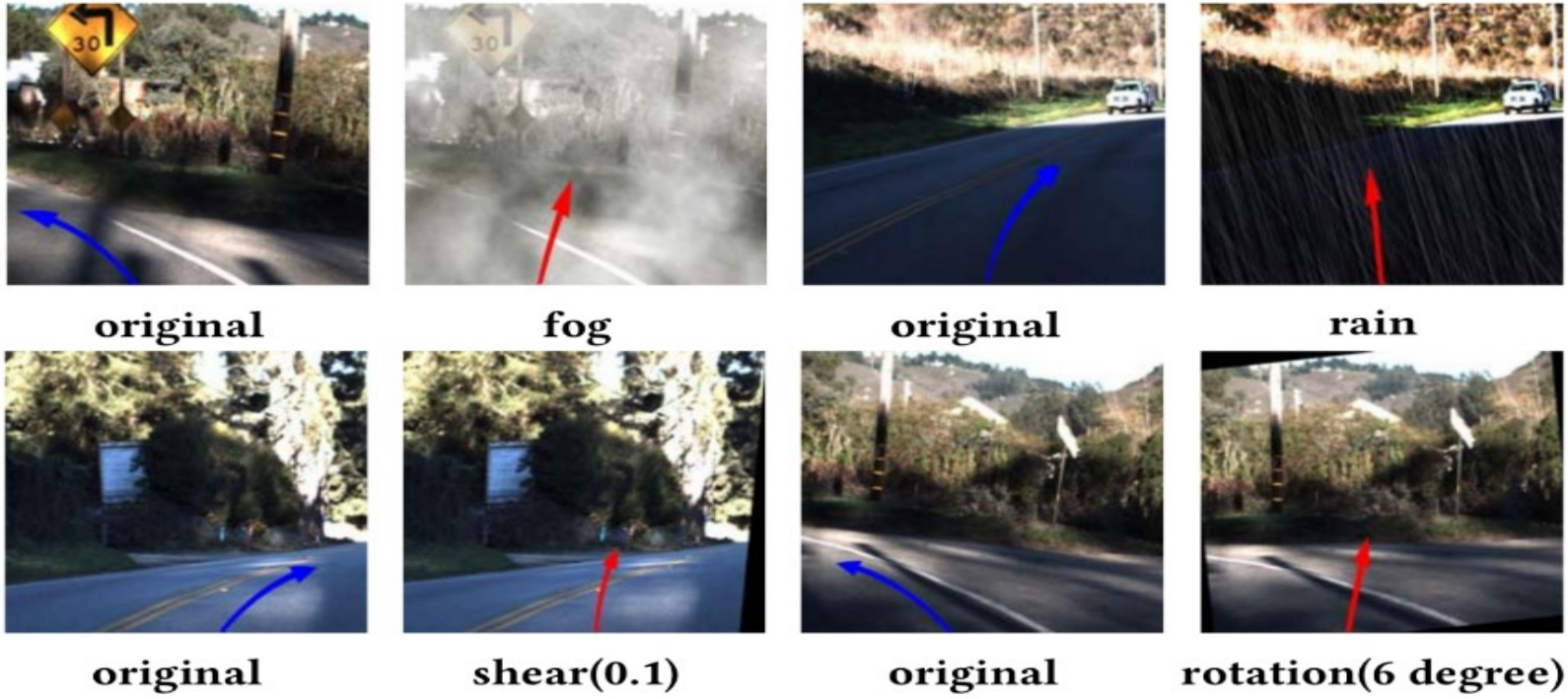
- Регрессия (прямая) изменилась из-за двух аномалий (выделено зеленым)
- Это аномалии, ошибки измерения или преднамеренно (злонамеренно) веденные данные в тренировочный набор?
- Но это легитимные данные, их не обнаружит “антивирус”

Чувствительность к данным



Standard computer vision models work well on third-person view (LEFT),
but fail on first-person perspective (RIGHT)

Изменение решения автопилота



Общие положения

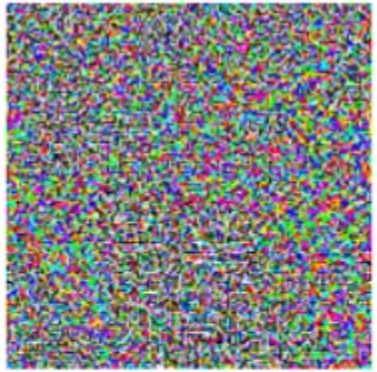
- Система машинного обучения естественным (безо всякого вмешательства) образом может работать не так, как мы ожидали
- Причина – проблемы с данными
- А если сознательным образом менять данные, так чтобы помешать работе или добиться желаемой работы?
- Это и есть атаки

Первый пример

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

О чем идет речь?

- Преднамеренные вмешательства в работу элементов конвейера машинного обучения
- Цель - воспрепятствовать работе систем машинного обучения, или же изменить их работу нужным злоумышленнику образом.
- Основная проблема – критические применения (авионика, автовождение и т.д.)
- Именно возможные атаки – основная причина проблем с применением в критических приложениях

Таксономия

- Место и время применения
- Знания об атакуемой системе
- Цели и задачи атак (целевые или нецелевые атаки)
- Предмет приложения: цифровые или реальные объекты
- Предметная область (домен)

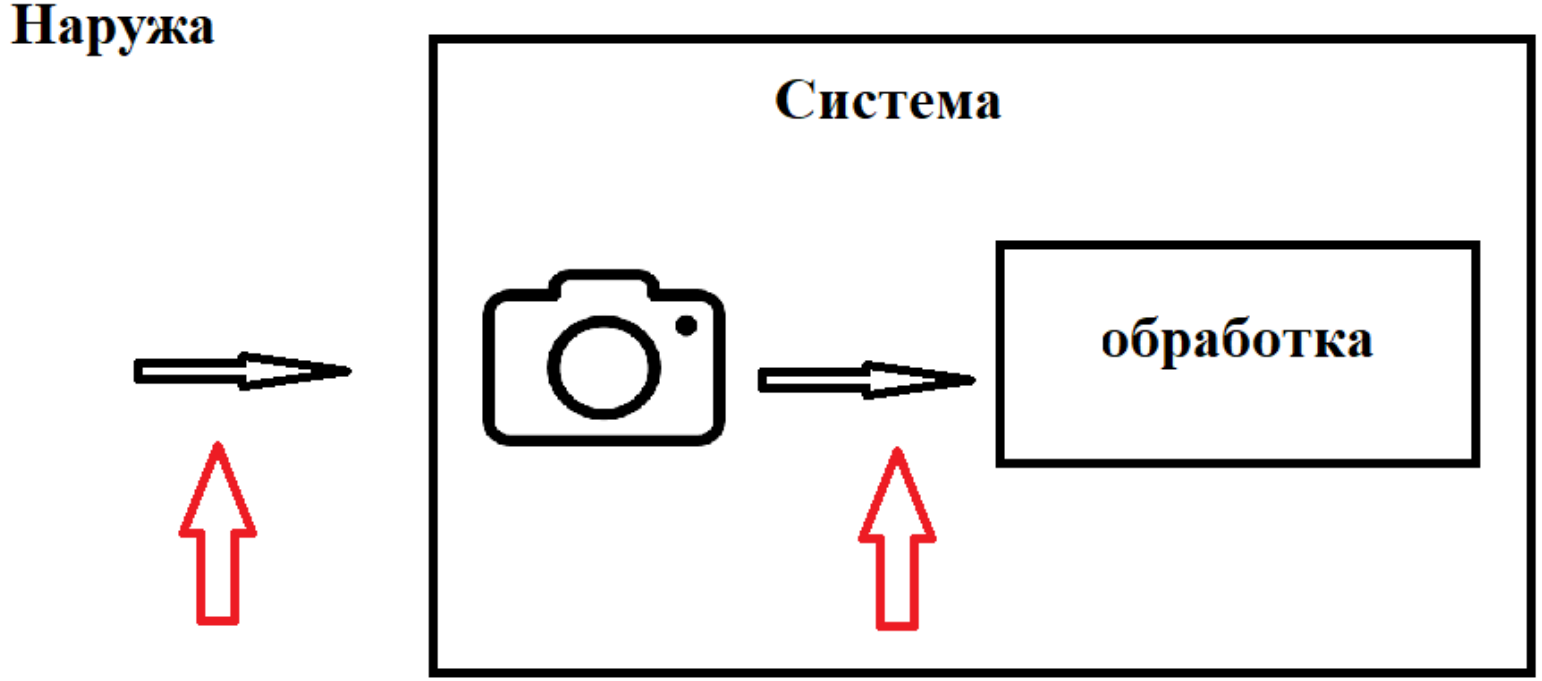
О чем идет речь?

- Все это может приводить к разным атакам.
- Атаки могут использовать уклонения, отравления, трояны, бэкдоры, перепрограммирование и вмешательство.
- Уклонение, отравление и вмешательство являются наиболее распространенными в настоящее время.
- Классификации бывают разные, но, в целом, в них есть общие элементы.

Пример классификации атак

Атака	Этап	Затрагиваемые параметры
Adversarial attack	применение	входные данные
Backdoor attack	тренировка	параметры сети
Data poisoning	тренировка, использование	входные данные
IP stealing	использование	отклик системы
Neural-level trojan	тренировка	отклик системы
Hardware trojan	аппаратное проектирование	отклик системы
Side-channel attack	использование	отклик системы

Физические и цифровые атаки

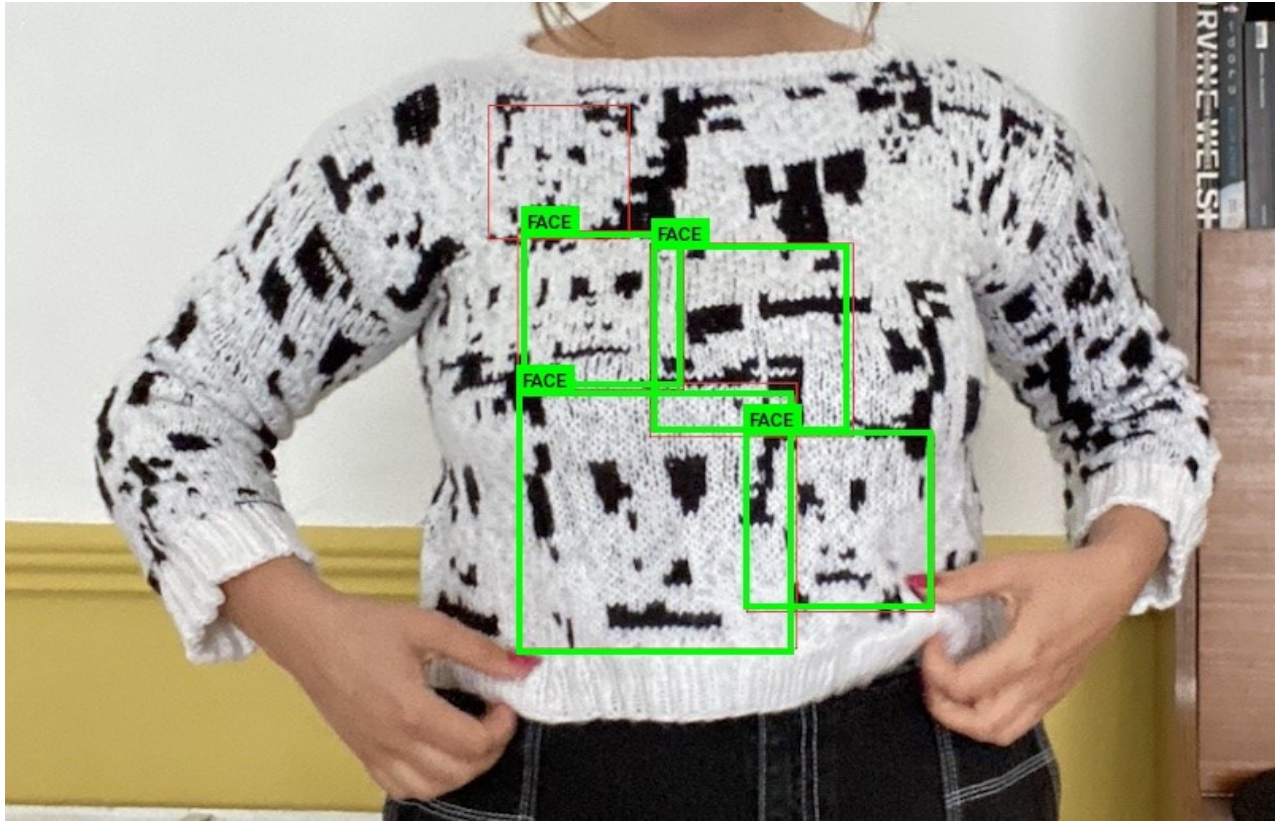


Физические атаки

Физические атаки

- Физические атаки следует признать наиболее опасными среди атак, воздействующих на входные данные
- Их нельзя “запретить”
- Вариации – бесконечны
- Они могут быть естественными

Физические атаки



Физические атаки

- Зеркальные очки
- При распознавании лиц в них каждый раз новый предмет

Физическая атака

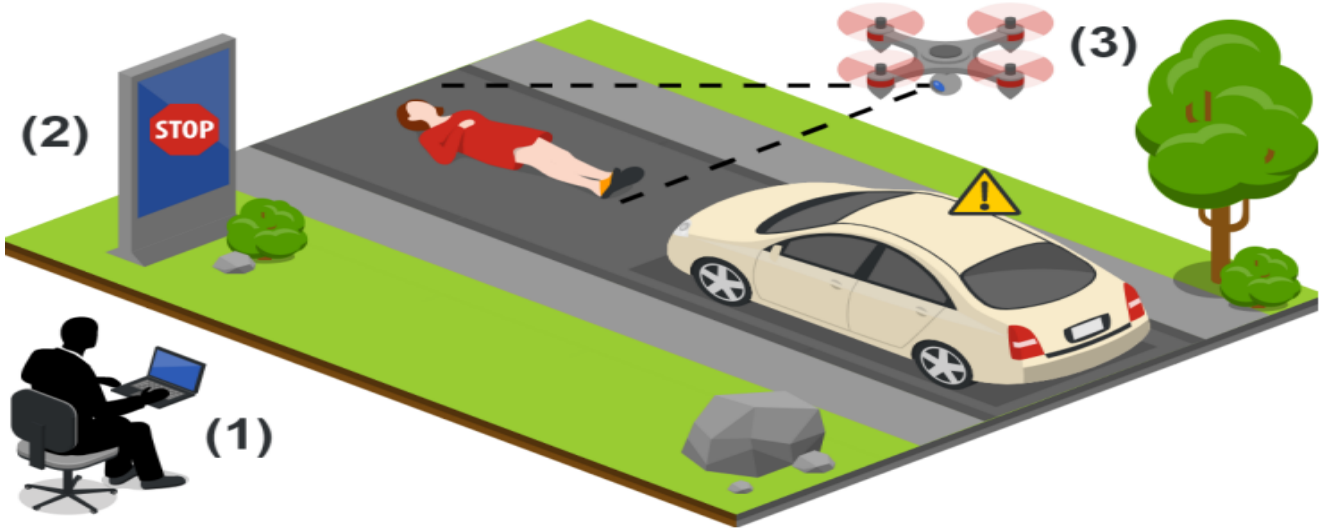
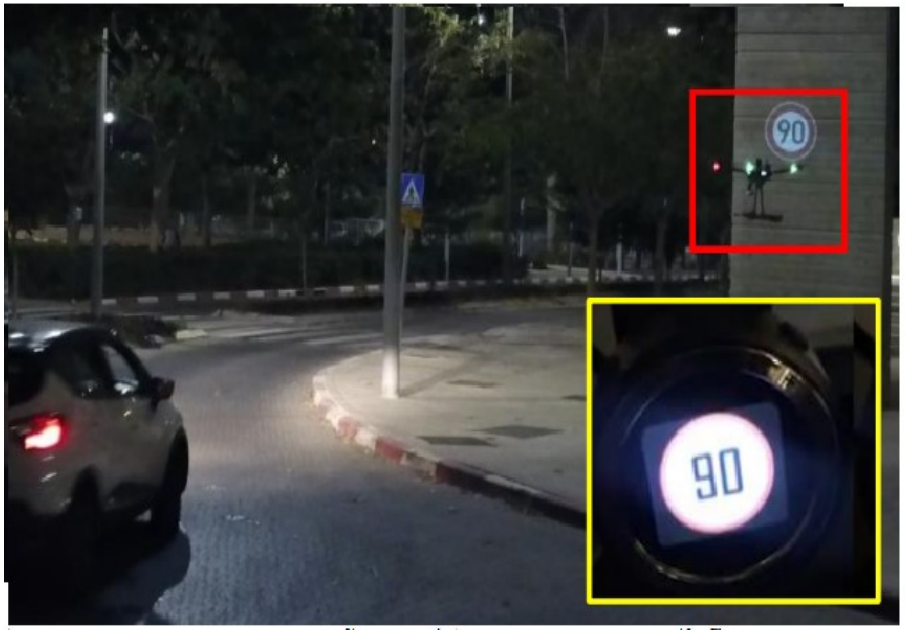
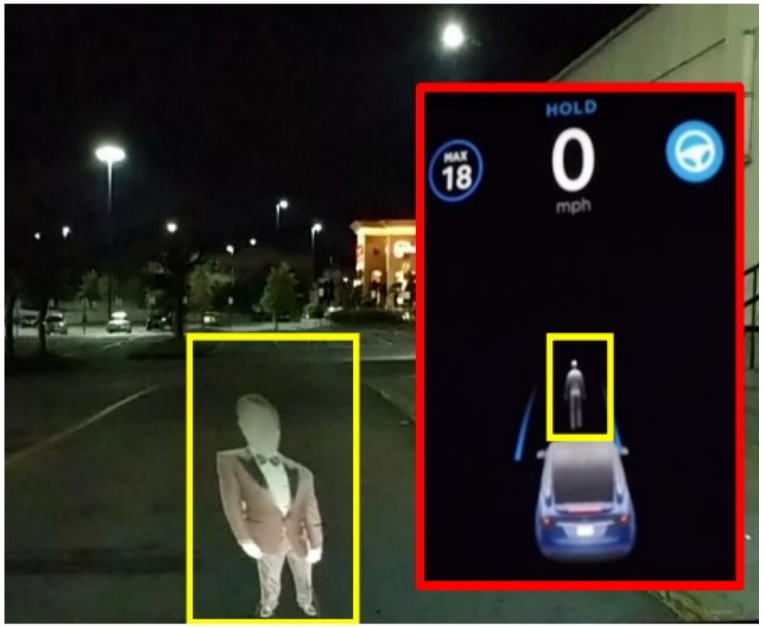


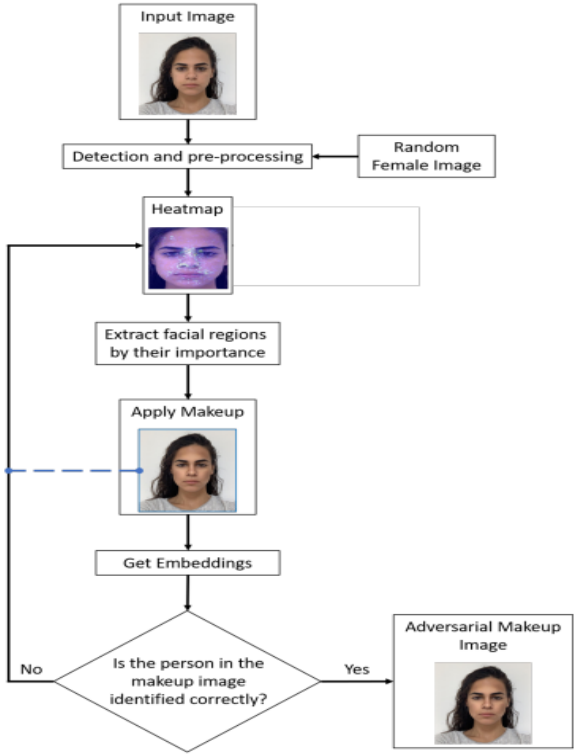
Fig. 4: The Threat Model: An attacker (1) either remotely hacks a digital billboard (2) or flies a drone equipped with a portable projector (3) to create a phantom image. The image is perceived as a real object by a car using an ADAS/autopilot, and the car reacts unexpectedly.

Физическая атака





Физические атаки



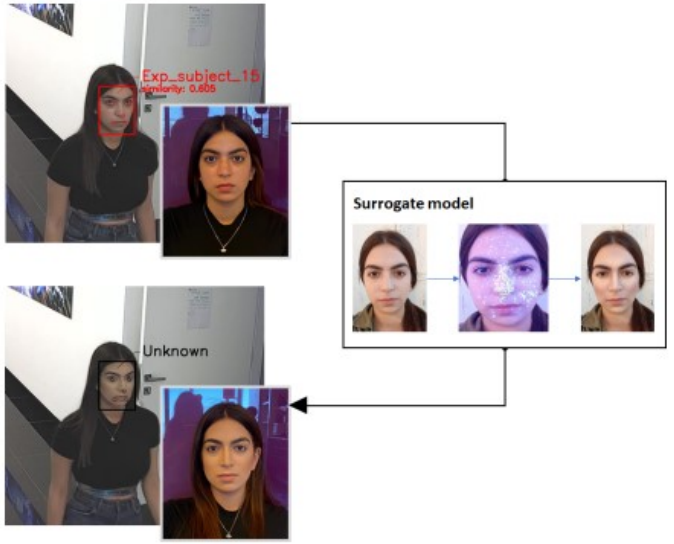
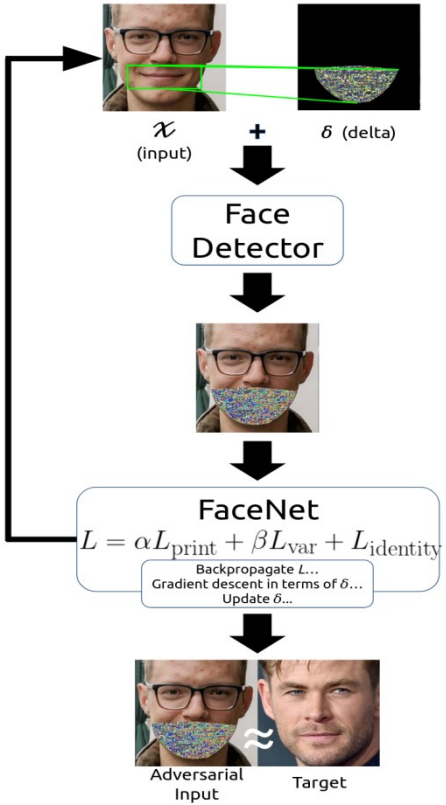


Figure 1: In the upper image the attacker is recognized by the face recognition (FR) system. In the middle image, our method uses a surrogate model to calculate the adversarial makeup in the digital domain, that is then applied in the physical domain . As a result, the attacker is not identified by the FR system (lower image).

Физические атаки

- Целевая атака против конкретной системы FaceNet (белый ящик)
- Изменение области вокруг рта, так, чтобы максимизировать расстояние до исходной фотографии и минимизировать до результирующей



Примеры

- Отметим, что атаки на тренировочной стадии случаются чаще, чем это, возможно, представляется.
- Связано это с тем, что работающие системы могут дополнительно тренироваться для обновления.
- Воздействие на данные между периодами до-обучения – это и есть атака (рекомендации в социальных сетях и т.п.)

Что мы знаем об атакуемом

- Атаки (методы генерации вредоносных искажений) могут быть разными, естественно, в зависимости от имеющихся у атакующего знаний о системе
- белый ящик,
- черный ящик,
- серый ящик

Теневая модель

- Копия атакуемой системы, на которой можно отрабатывать атаки
- Отсюда – информация о параметрах используемых моделей в реальных применениях не должна являться публичной
- Модели – публичны, применения - нет

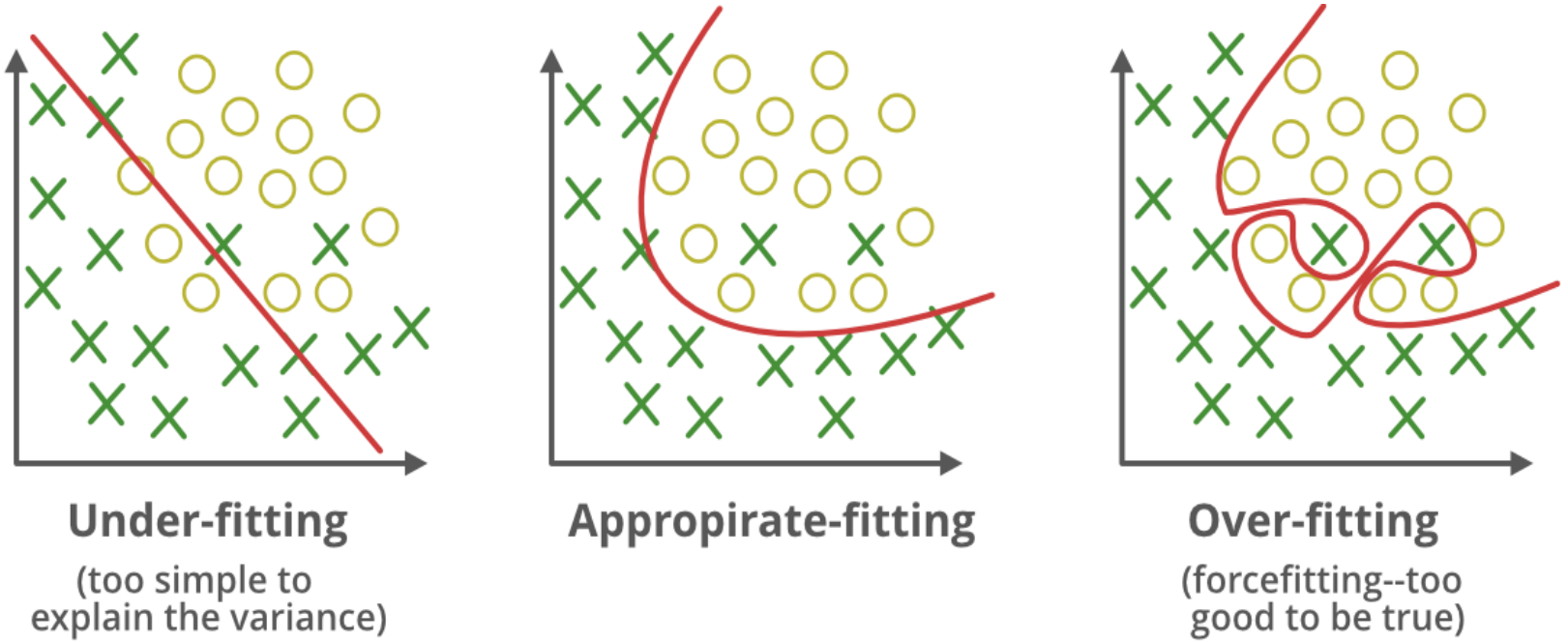
Adversarial ML Threat Matrix

Reconnaissance	Initial Access	Execution	Persistence	Model Evasion	Exfiltration	Impact
Acquire OSINT information: (Sub Techniques) 1. Arxiv 2. Public blogs 3. Press Releases 4. Conference Proceedings 5. Github Repository 6. Tweets	Pre-trained ML model with backdoor	Execute unsafe ML models (Sub Techniques) 1. ML models from compromised sources 2. Pickle embedding	Execute unsafe ML models (Sub Techniques) 1. ML models from compromised sources 2. Pickle embedding	Evasion Attack (Sub Techniques) 1. Offline Evasion 2. Online Evasion	Exfiltrate Training Data (Sub Techniques) 1. Membership inference attack 2. Model inversion	Defacement
ML Model Discovery (Sub Techniques) 1. Reveal ML model ontology – 2. Reveal ML model family –	Valid account	Execution via API	Account Manipulation		Model Stealing	Denial of Service
Gathering datasets	Phishing	Traditional Software attacks	Implant Container Image	Model Poisoning	Insecure Storage 1. Model File 2. Training data	Stolen Intellectual Property
Exploit physcial environment	External remote services			Data Poisoning (Sub Techniques) 1. Tainting data from acquisition – Label corruption 2. Tainting data from open source supply chains 3. Tainting data from acquisition – Chaff data 4. Tainting data in training environment – Label corruption		Data Encrypted for Impact Defacement
Model Replication (Sub Techniques) 1. Exploit API – Shadow Model 2. Alter publicly available, pre-trained weights	Exploit public facing application					Stop System Shutdown/Reboot
Model Stealing	Trusted Relationship					

Основания для атак

- Почему вообще существуют атаки? На сегодняшний день в сообществе нет единого мнения относительно того, почему это могло быть.
- Существует ряд объяснений, не всегда согласующихся друг с другом.
- Первая и оригинальная гипотеза, пытающаяся объяснить состязательные примеры, была взята из собственной статьи Сегеди, где авторы утверждали, что такие примеры существуют из-за наличия маловероятных «карманов» в многообразии (то есть слишком большой нелинейности) и плохой регуляризации сетей.
- Отсюда, между прочим, оверфиттинг (переобучение) – это проблема с устойчивостью

Overfitting vs underfitting



Основания для атак

- Противоположная теория, впервые предложенная Гудфеллоу -на самом деле состязательные примеры возникали из-за слишком большой линейности в современном машинном обучении и особенно в системах глубокого обучения.
- Гудфеллоу утверждал, что функции активации, такие как *ReLU* и *Sigmoid*, в основном представляют собой прямые линии. Итак, на самом деле внутри нейронной сети у вас есть много функций, которые сохраняют вклад друг друга в одном направлении.
- Если вы затем добавите крошечные возмущения к некоторым входам (несколько пикселей здесь и там), которые накапливаются в огромную разницу на другом конце сети, она выдаст непонятный результат.

Основания для атак

- Третья и, возможно, наиболее часто принимаемая сегодня гипотеза - это объяснение, заключающееся в том, что, поскольку модель никогда не соответствует данным идеально (в противном случае точность набора тестов всегда была бы 100%), всегда будут существовать враждебные карманы входных данных, которые существуют между границей классификатора и фактической подгруппой множества выборочных данных.

Основания для атак

- Принципиальный момент состоит в том, что при обучении мы используем только какое-то подмножество данных. И, вообще говоря, не знаем всего о генеральной совокупности.
- Тогда наличие неизвестных системе примеров следует признать свойством выбранного нами подхода, а никак не исключением.
- Атака – это сознательная генерация (подбор) состязательных примеров, которые существуют (могут существовать) и без злонамеренных пользователей.
- Отсюда – важно понимать природу данных (“физику” системы) и возможность оценки генеральной совокупности

Предметные области

- Изображения
- Видео
- Звук
- Временные ряды
- Текст

Отравление данных

- Простой и эффективный способ атаковать процесс обучения — просто поменять местами метки некоторых экземпляров обучения.
- Этот тип атак с отравлением данных называется атаками с переворачиванием меток (label flipping).
- Ошибочную маркировку можно легко сделать в краудсорсинге, где злоумышленник является одним из аннотаторов (разметчиков), например
- Ошибки в датасете (15% ImageNet – неверная разметка)

Трояны (backdoors)

- Вредоносная функциональность встроена в веса (архитектуру) нейронной сети.
- Нейронная сеть будет вести себя нормально при большинстве входных данных, но при определенных обстоятельствах (определенных данных) будет вести себя опасно.
- С точки зрения безопасности это особенно опасно потому, что нейронные сети — это черные ящики.
- Модели машинного обучения становятся все более доступными, а конвейеры обучения и развертывания становятся все более непрозрачными, что усугубляет проблему безопасности.

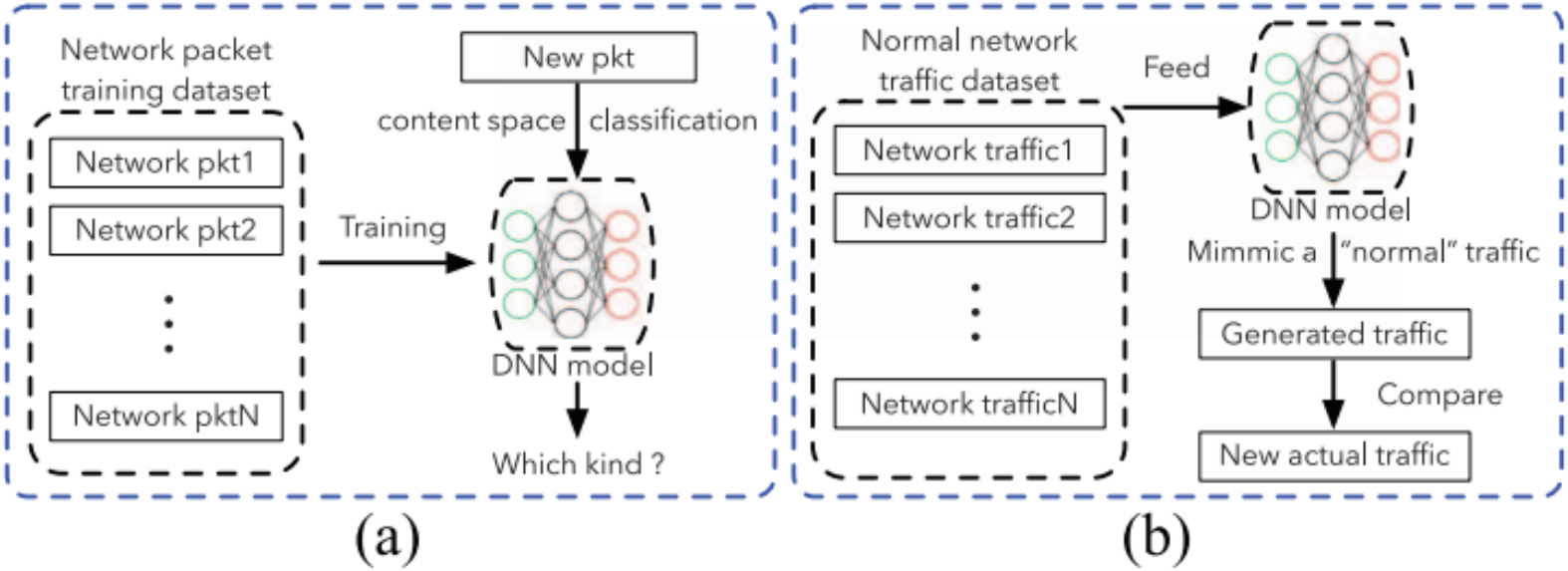
Трояны

- При троянской атаке злоумышленник пытается заставить входные данные с определенными триггерами (признаками) создавать вредоносные выходные данные, не нарушая производительность входных данных без триггеров.
- NIST - отдельное направление, посвященное троянам https://www.nist.gov/itl/ssd/trojai

Атаки отравлением

- Повреждение логики является наиболее опасным сценарием. Повреждение логики происходит, когда злоумышленник может изменить алгоритм и способ его обучения.
- На этом этапе машинное обучение перестает иметь значение, потому что злоумышленник может просто закодировать любую логику, которую он хочет. Аналогия – использование множества операторов if
- Дообучение модели – сохраняет трояны !

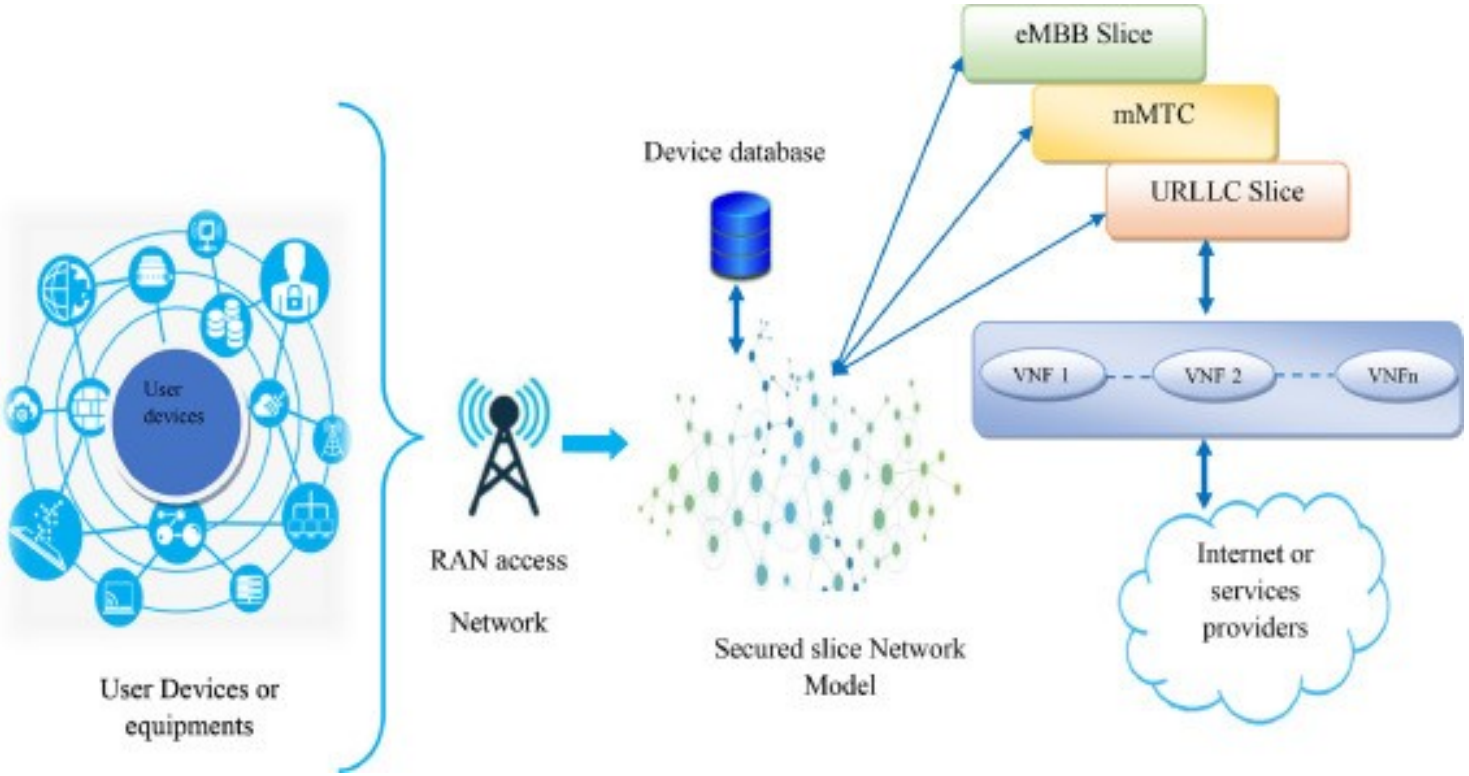
Атака на систему обнаружения вторжений (NIDS)



Атака на систему обнаружения вторжений (NIDS)

- DNN, обученная классифицировать трафик (a) или определять аномалии (b)
- Извлекают модель (10% тренировочного набора)
- Строят карты значимости признаков (saliency maps) и на их основе создают состязательный пример
- https://tianweiz07.github.io/Papers/21-iotj-3.pdf

Состязательные атаки на 5G



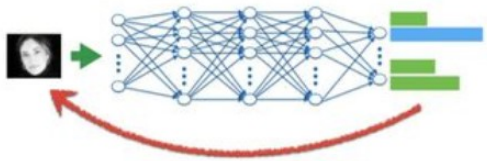
Состязательные атаки на 5G

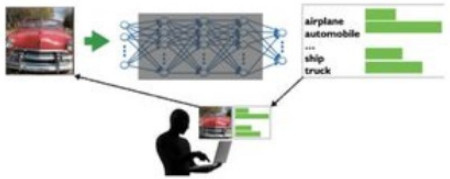
- Машинное обучение используется для управления сетевой архитектурой в 5G
- По факту отсутствия аналитических моделей (например, управление виртуализацией)
- Тем самым – открывается возможность для атак
- https://arxiv.org/pdf/2207.01531.pdf - атака черного ящика на инфраструктуру 5G (не на модель, а на ее объемлющую часть)

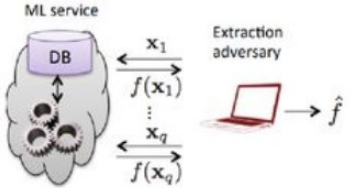
Проблемы для MLaaS

Privacy: A Big Challenge for MLaaS

- Model Inversion Attack
  - Fredrikson et al. CCS'15
- Membership Inference Attack
  - Shokri et al. IEEE S&P'17
- Model Extraction Attack
  - Tramèr et al. Usenix Security'16







Суммарно

- Атаки на системы ML – реальность
- Объемлющего решения по защите на сегодня нет
- Нужно понимать предметную область
- Загрузка моделей – реальная угроза
- Параметры модели в конкретном случае – не разглашаются
- Проверка исходных данных – обязательна
- Мониторинг (OOD) - обязателен

Образовательные программы

- ВМК МГУ – 2021. Магистерская программа «Искусственный интеллект в кибербезопасности» (ФГОС) https://cs.msu.ru/node/3732 Фактически: Кибербезопасность ИИ. Первый выпуск - 2023
- ИТМО - 2022

Фейки

- Использование ML для производства и распространения высококачественного аудиовизуального контента, называемого синтетическими медиа и дипфейками
- Контент, неотличимый от настоящего
- DARPA Semantic Forensics (SemaFor)
- DARPA MediaForensics (MediaFor)
- Coalition to Content Provenance and Authenticity (C2PA) – анализ происхождения контента

Заключение

- Кибербезопасность систем ИИ выделяется в отдельное направление в силу особой значимости
- Влияет на все возможные применения дискриминантных моделей
- Состязательные атаки – одно из главных препятствий внедрения ML в критические приложения
- Тесно связано с проблемой устойчивости систем машинного обучения