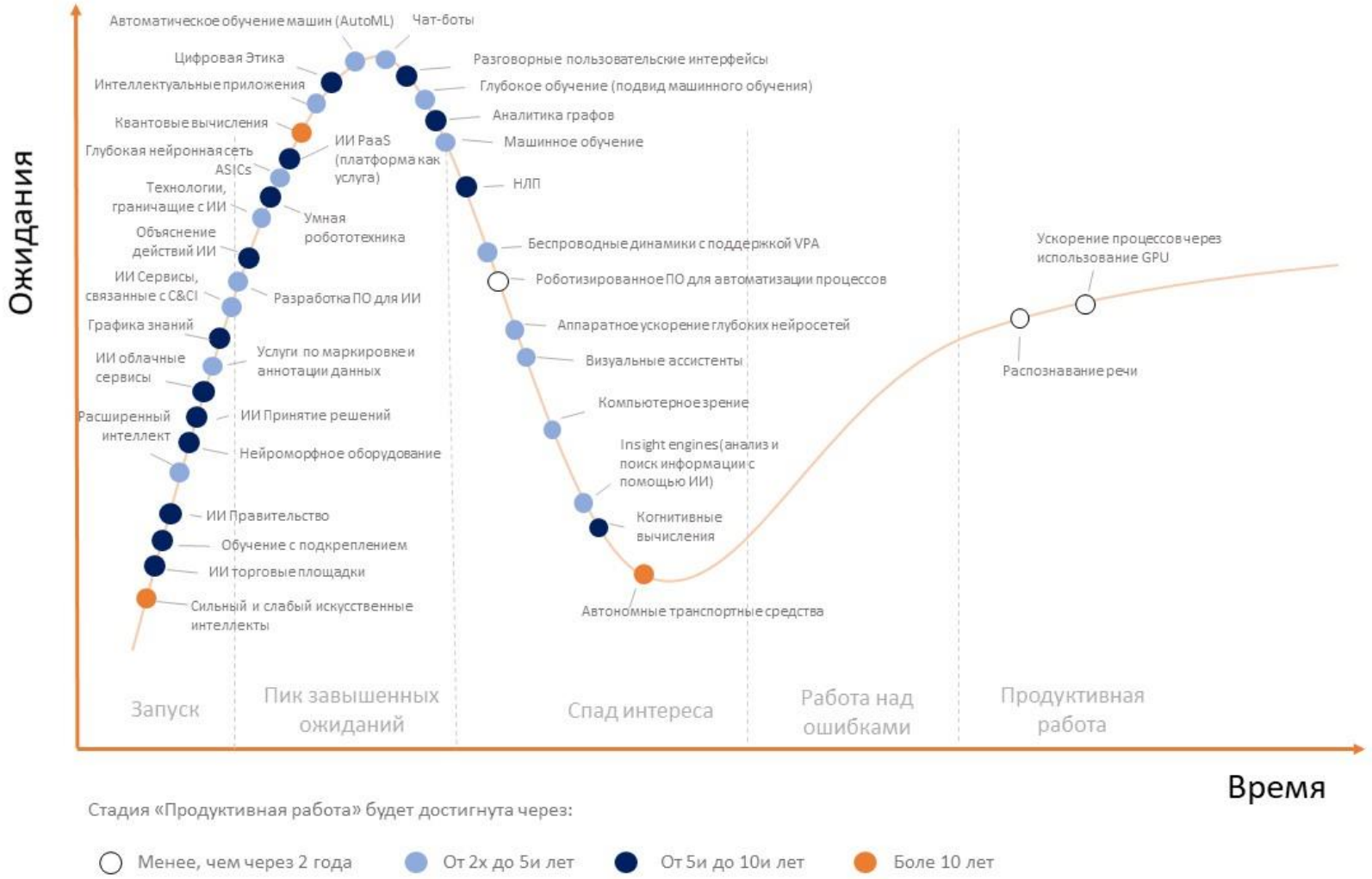
Обман нейросети

Технический специалист по информационной безопасности
Семенычев А.М.

Искусственный Интеллект. Немного статистики.



По данным Gartner 2019 года.

Искусственный Интеллект. Немного статистики

Table 1. AI Software Market Forecast by Use Case, 2021-2022, Worldwide (Millions of U.S. Dollars)

Segment	2021 Revenue	2021 Growth (%)	2022 Revenue	2022 Growth (%)
Knowledge Management	5,466	17.6	7,189	31.5
Virtual Assistants	6,210	12.0	7,123	14.7
Autonomous Vehicles	5,703	13.7	6,849	20.1
Digital Workplace	3,593	13.7	4,309	20.0
Crowdsourced Data	3,483	13.6	4,171	19.8
Others	27,049	14.1	32,827	21.4
Total	**51,503**	**14.1**	**62,468**	**21.3**

Source: Gartner (November 2021)

По данным Gartner за ноябрь 2021 года.

Искусственный Интеллект. Прогнозы.

- к 2025 году предварительно обученные модели ИИ будут в основном сосредоточены в руках 1% поставщиков;
- в 2023 году 20% успешных атак с захватом аккаунта будут использовать дипфейки;
- к 2024 году 60% поставщиков ИИ будут включать в свое программное обеспечение меры по предотвращению его потенциально вредоносного / неправомерного использования;
- к 2025 году 10% правительств будут избегать проблем нарушения конфиденциальности и безопасности, используя отдельные группы населения для обучения ИИ;
- к 2025 году 75% разговоров на рабочем месте будут записываться и анализироваться для повышения организационной ценности и оценки рисков.

По данным Gartner за январь 2021 года.

Нейронная сеть. Что это?

Нейронная сеть (искусственная нейронная сеть) – организация вычислительных элементов, по принципу имитирующему структуру мозга. Характерной особенностью нейронной сети является ее обучаемость – способность находить зависимости между входными и выходными данными, которые предлагаются ей в ходе обучения.

Задачи которые решает нейронная сеть:

- распознавание образов
- кластеризация (объединение в группы – кластеры)
- построение прогнозов
- сжатие информации и восстановление поврежденных или «зашумленных» данных.

Нейронная сеть. Что такое Adversarial Machine Learning?

Adversarial Machine Learning (AML) – дословно: «состязательное машинное обучение». Однако корректнее говорить про «вредоносное машинное обучение». Это целенаправленное воздействие на нейронную сеть, призванное вызвать ошибки в ее поведении.

Нейронная сеть. Отравление данных.

Отравление данных (Data Poisoning) – создание предпосылок для ошибок на этапе обучения нейросети.

Обычный «Стоп»

Ложное «Ограничение скорости»

Специально добавлен триггер

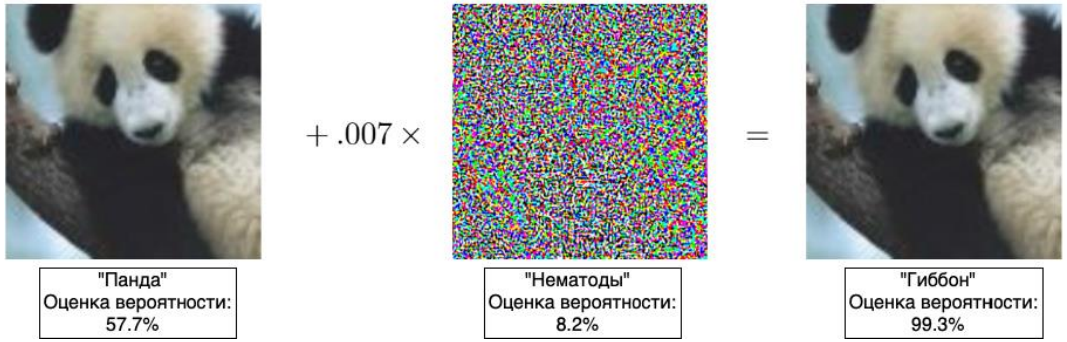
Нейронная сеть. Отравление данных.

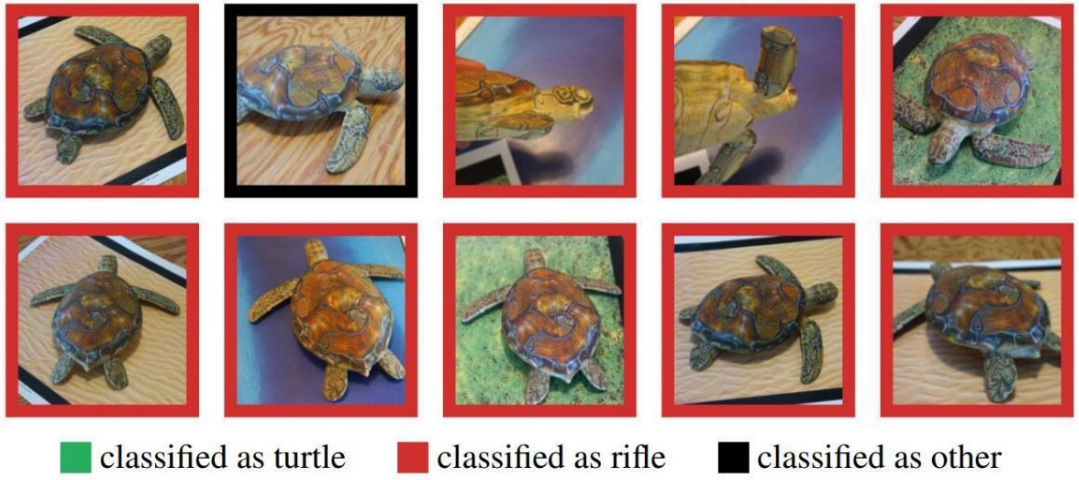
Как это происходит

- Порча данных при обучении в облаке
- Использование уже обученных моделей
- Порча данных на краудсорсинг-платформах
- Порча данных сотрудниками
- Шпионаж

Нейронная сеть. Атаки уклонения.

Атаки уклонения (Evasion Attack) – создание предпосылок для ошибок на этапе применения нейросети.



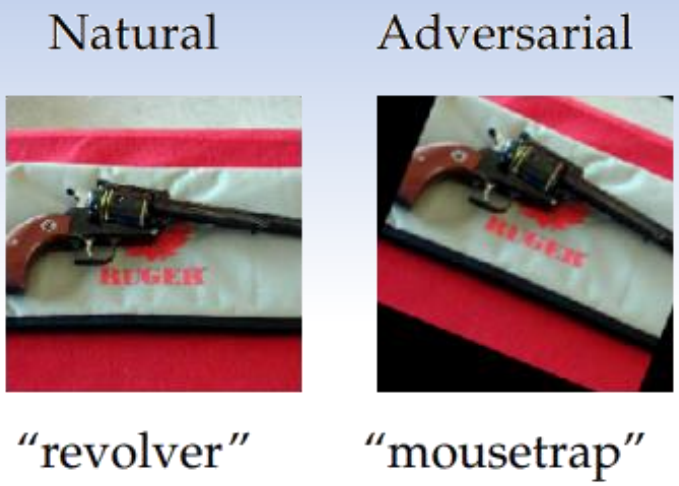


Нейронная сеть. Атаки уклонения.

Как это происходит

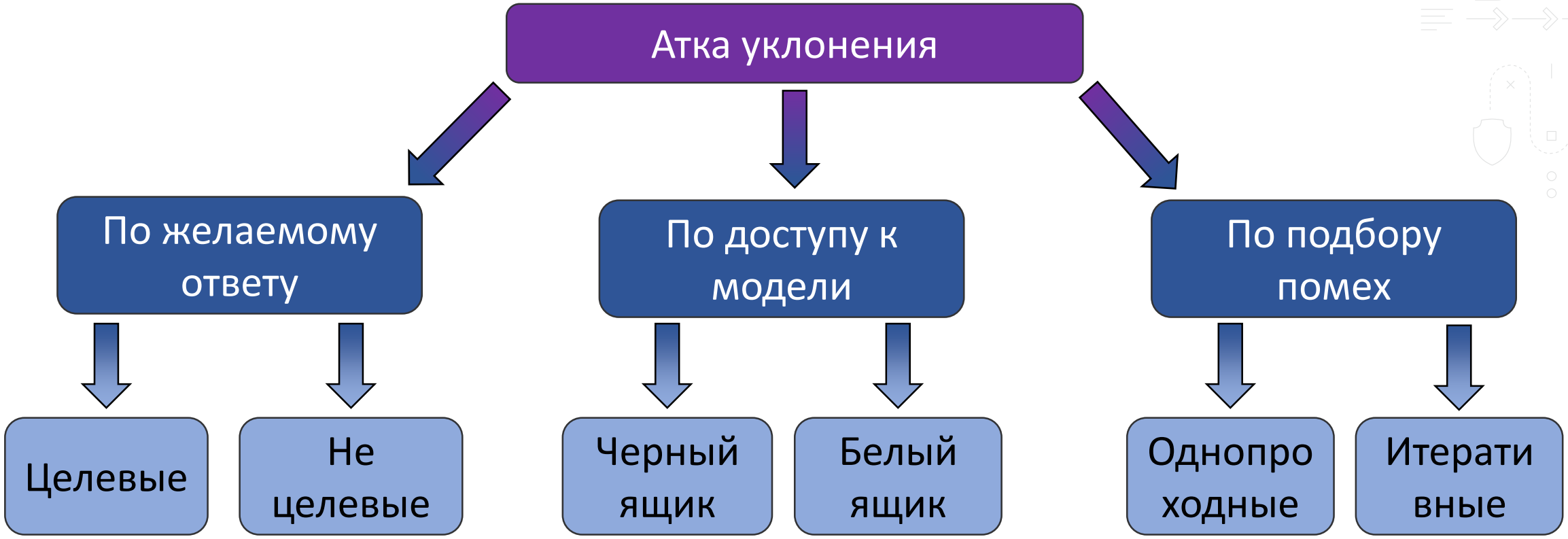
В основном, для атак уклонения применяются «состязательные примеры» (adversarial examples)

Состязательные примеры зависят от данных, а не от архитектур, и их можно сгенерировать для большинства датасетов.



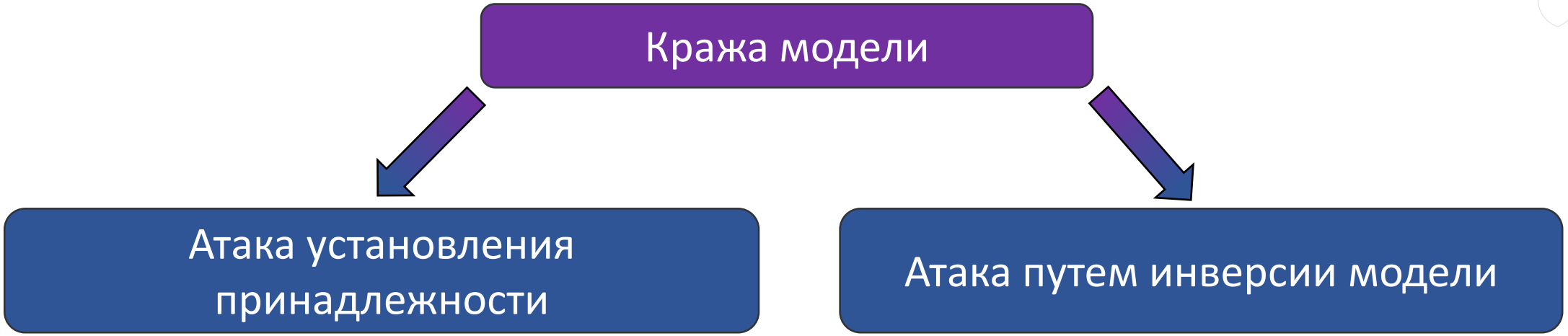
Состязательные примеры отлично переносятся в физический мир.

Нейронная сеть. Атаки уклонения.



Нейронная сеть. Кража модели.

Кража модели (Model Extraction) – определение на каких данных обучалась модель или извлечение обучающих данных из обученной модели.



Нейронная сеть. Методы защиты.

Механизмы защиты

- Определение алгоритмов безопасного обучения
- Использование нескольких систем классификаторов
- Обучение с сохранением конфиденциальности
- Теоретико-игровые AML-модели, включая интеллектуальный анализ данных
- Очистка обучающей выборки от отравляющих атак

Методы эмпирической защиты

- Состязательная тренировка (Adversarial training, AT)
- Градиентное маскирование (Gradient masking)
- Обнаружение (Detection) и входная модификация (Input modification)
- Дополнительный класс (Extra class)

Нейронная сеть. Подборка AML библиотек для Python.

Наименование	Описание
AdversariaLib	Предоставляет различные варианты атак с уклонением.
AdLib	Библиотека Python с интерфейсом в стиле scikit, которая включает реализации ряда опубликованных атак с уклонением.
AlfaSVMLib	Отравляющие атаки с использованием опорных векторов и атаки против алгоритмов кластеризации.
deep-pwning	Metasploit инструмент для атак на нейросети глубокого обучения, использует Tensorflow.
Cleverhans	Библиотека Tensorflow для тестирования существующих моделей глубокого обучения на предмет известных атак.
foolbox	Библиотека Python для создания AML-образцов, реализует различные атаки.
SecML	Библиотека Python для безопасного и понятного машинного обучения. Предоставляет широкий спектр инструментов для машинного обучения, алгоритмы атак и т.д.

Нейронная сеть. Подборка AML библиотек для Python.

Наименование	Описание
TrojAI	Библиотека Python для создания масштабных моделей с бэкдором и троянами для исследования обнаружения троянов
Adversarial Robustness Toolkit (ART)	Библиотека Python для безопасного машинного обучения. Предоставляет инструменты, которые позволяют разработчикам и исследователям защищать и оценивать модели, и приложения машинного обучения от враждебных угроз Evasion, Poisoning, Extraction и Inference.
Advertorch	Набор инструментов Python для исследования устойчивости основные функции реализованы в PyTorchВикипедия
DeepRobust	Библиотека состязательного обучения pytorch, которая содержит самые популярные алгоритмы атак и защиты в области изображений и графов.
TextAttack	Фрейморк для AML.
OpenAttack	Основанный на Python набор инструментов текстовых AML с открытым исходным кодом. Обрабатывает весь процесс текстовых состязательных атак, включая предварительную обработку текста, доступ к модели жертвы, генерирование состязательных примеров и оценку

Спасибо за внимание!

Технический специалист по информационной безопасности
Семенычев А.М.
a.semenychev@gardatech.ru