конференция РусКрипто

Министерство науки и высшего образования Российской Федерации
Санкт-Петербургский политехнический университет Петра Великого
Институт кибербезопасности и защиты информации

# Анализ методов обнаружения искусственно синтезированного контента

Докладчик:

В.Д. Данилов

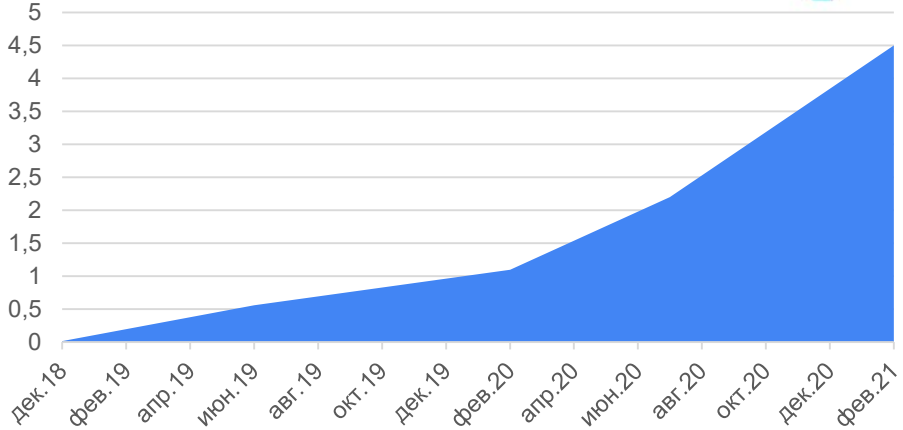
Москва – 2023

# Актуальность

## Негативное влияние использования DeepFake

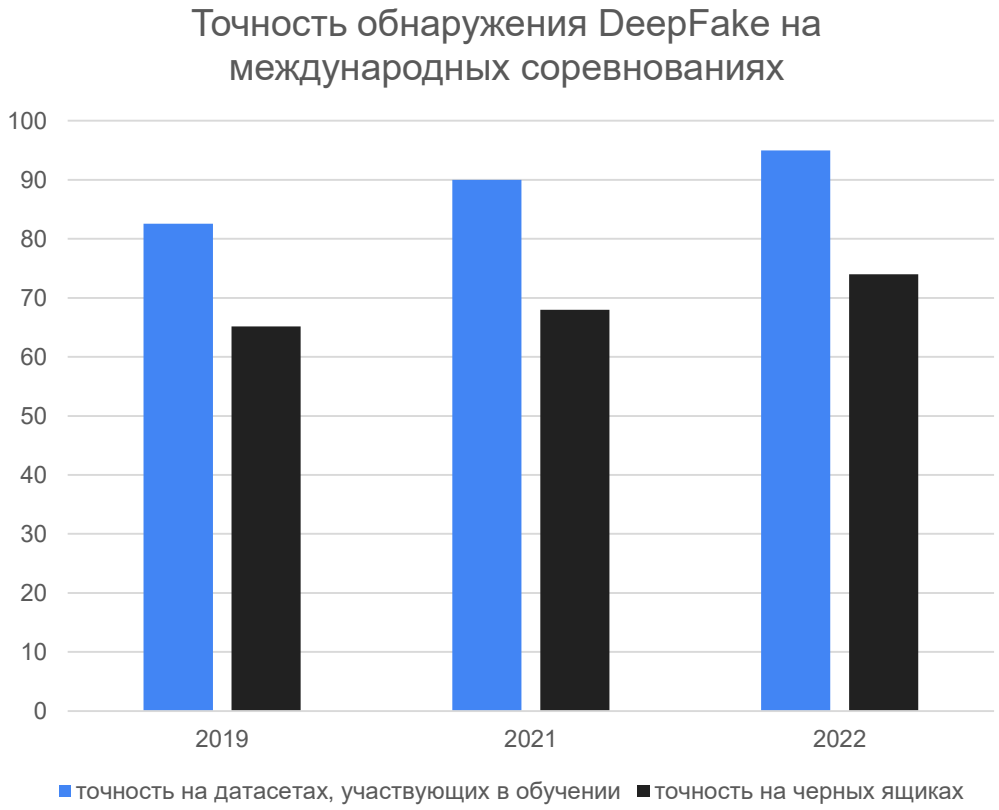
- Манипулирование общественным мнением
- Разжигание конфликтов, массовых беспорядков, подрыв национальной безопасности
- Репутационные риски отдельных людей или компаний
- Организация кибератак, связанных с социальной инженерией
- Обход биометрических систем безопасности

Количество DeepFake-видео в интернете согласно исследованию Deeptrace, млн




Более 96% DeepFake контента создается для незаконных целей

# Проблематика обнаружения DeepFake-контента




Методы обнаружения искусственно синтезированного контента показывают высокую точность только при работе с данными из тех датасетов, которые участвовали в обучении

# Исследование методов генерирования искусственно синтезированного контента

## Основные методы генерирования:

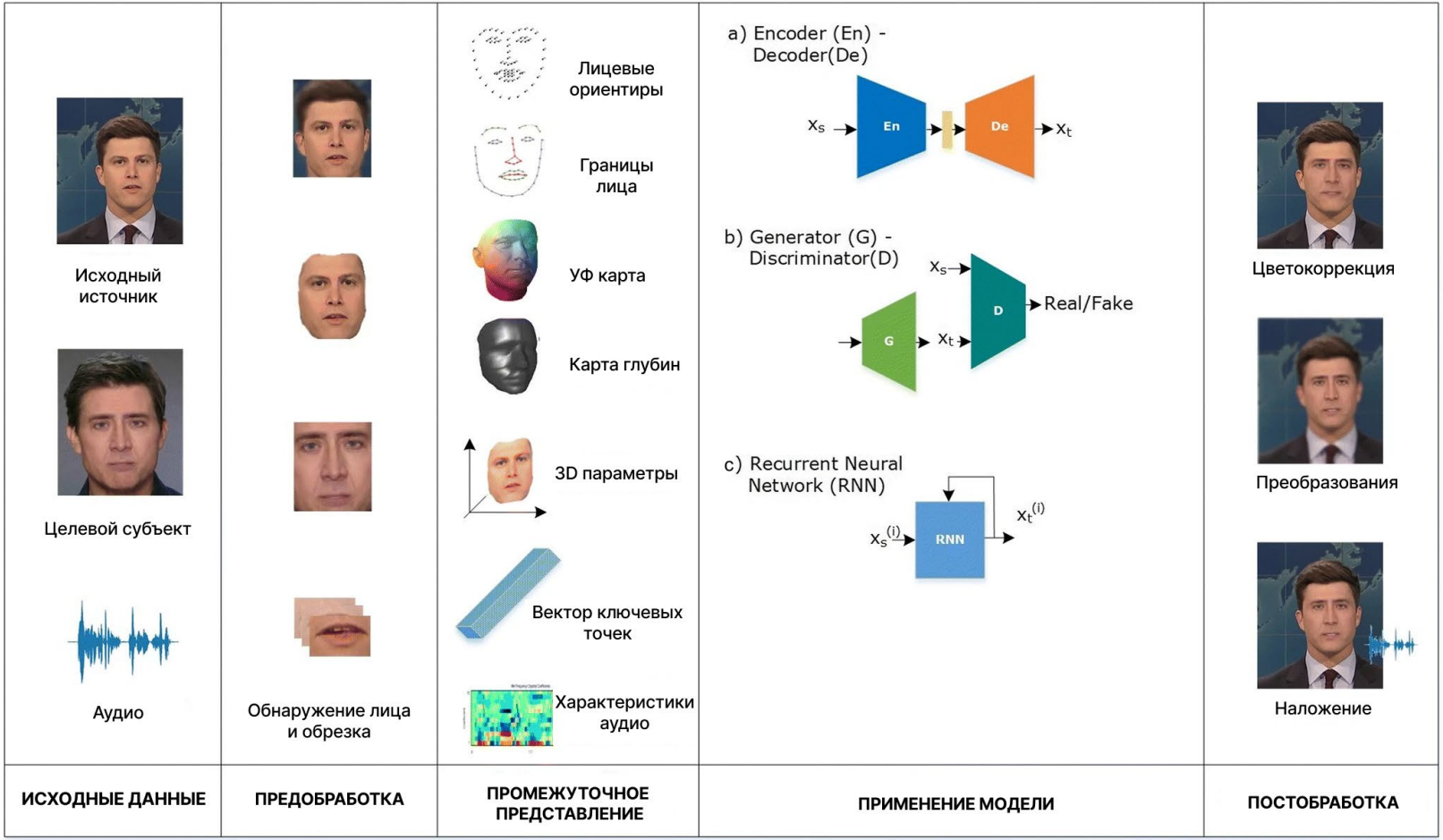
| Изображения | Аудио | Видеоданные |
|---|---|---|
| • Генеративно-состязательные сети (GAN)<br>• Диффузионные модели<br>• Рекуррентные и сверточные нейроные сети (RNN и CNN)<br>• Вариационные автоэнкодеры | • GAN<br>• Повторное воспроизведение<br>• RNN и CNN | • GAN<br>• RNN и CNN<br>• Вариационные автоэнкодеры |

## Типы генерируемого контента:

- Полностью искусственно сгенерированный контент
- Подмена лица
- Манипуляция атрибутами лица (изменение пола, возраста и т.д.)

ПОЛИТЕХ МЫСЛИТЬ БУДУЩИМ

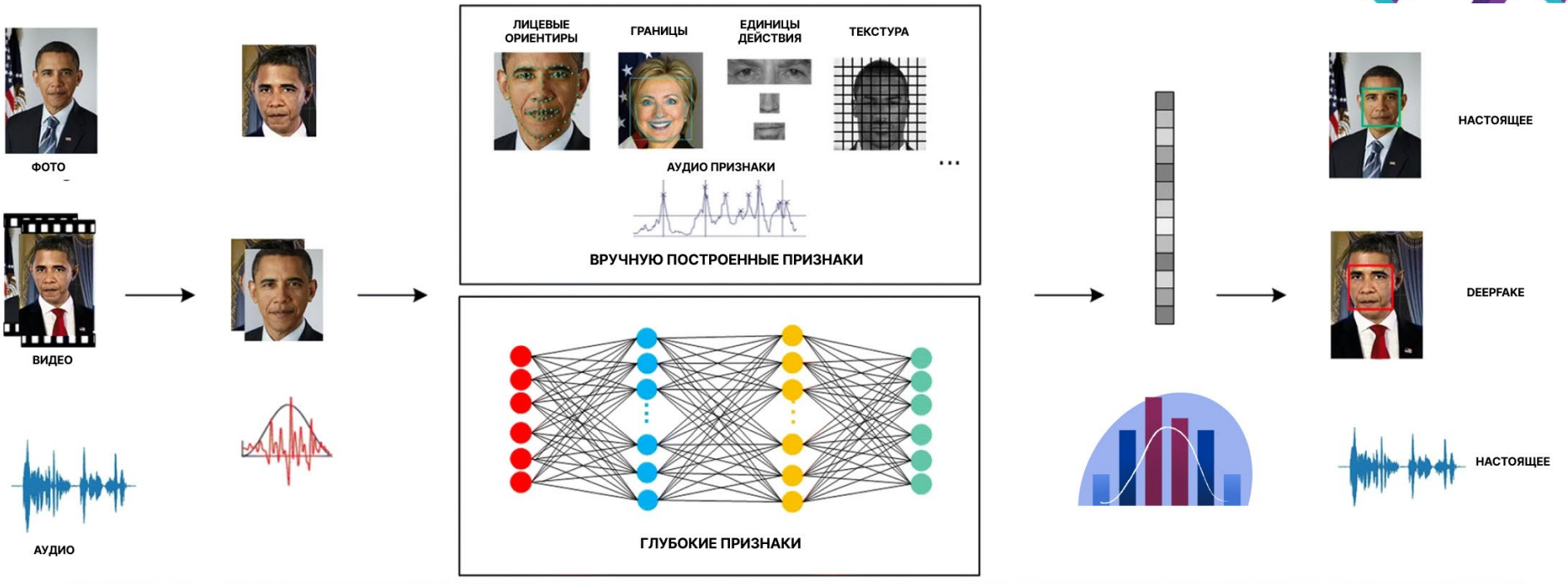
# Генерирование DeepFake-контента



# Анализ методов обнаружения искусственно синтезированного контента

| Статистические методы | Машинное обучение | Методы глубокого обучения | Другие методы |
|---|---|---|---|
| • EM-алгоритм (Expectation-Maximization)<br>• Вариационное исчисление<br>• Расстояние Кульбака-Лейблера<br>• Дивергенция Йенсена — Шеннона | • Метод опорных векторов (SVM)<br>• Метод K-средних (K-Means)<br>• Логистическая регрессия (LR)<br>• Многослойный персептрон (MLP)<br>• Бустинг | • CNN<br>• RNN (LSTM)<br>• Региональные сверточные сети (RCNN)<br>• GAN | • Использование блокчейна<br>• Анализ метаданных<br>• Поиск оригиналов в сети Интернет<br>• Анализ водяных знаков<br>• Сравнение с оригинальными данными |

# Анализ методов обнаружения искусственно синтезированного контента




ИСХОДНЫЕ ДАННЫЕ | ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА ДАННЫХ | ВЫДЕЛЕНИЕ ПРИЗНАКОВ | КЛАССИФИКАЦИЯ / АНАЛИЗ СТАТИСТИЧЕСКИХ ПОКАЗАТЕЛЕЙ | РЕЗУЛЬТАТ

# Статистические методы обнаружения

| Метод | Тип данных | Точность | Описание |
|---|---|---|---|
| Анализ сверточных следов | Изображения | 95% | Поиск корреляции соседних пикселей и извлечение признаков с помощью алгоритма EM |
| Робастная оценка (robust estimation) | Изображения | - | Робастная оценка на основе расстояния Кульбака-Лейблера, дивергенции Йенсена-Шеннона с использованием метрики Вассерштейна |
| Анализ неоднородности фотоотклика | Видео | - | Разделение видео на кадры и определение корреляции фотоотклика, после чего применяется t-критерий Стьюдента. |
| Биспектральный анализ | Аудио | 90% | Обнаружение следов применения методов глубокого обучения путем поиска специфических биспектральных корреляций высшего порядка |

| Преимущества | Недостатки |
|---|---|
| Нетребовательность к вычислительным ресурсам, отсутствие необходимости использования обучающих наборов данных | Узкая направленность методов, падение точности при высоком качестве данных, необходимость предварительного анализа |

# Методы обнаружения с использованием машинного обучения

| Метод | Тип данных | Точность | Описание |
|---|---|---|---|
| Отслеживание позиции головы | Видео, изображения | 90% | Анализ по ориентирам в области головы с помощью SVM |
| Визуальные признаки | Видео, изображения | 86% | Обнаружение на основе поиска визуальных артефактов с помощью MLP, возникающих из-за проблем освещения и геометрии |
| Извлечение энтропийных характеристик | Аудио | 96% | Выявление на основе анализа энтропийных характеристик с помощью логистической регрессии |
| Кепстральный и биспектральный анализ | Аудио | 98% | Осуществление кепстрального и биспектрального анализа с помощью квадратичных опорных векторов |

| Преимущества | Недостатки |
|---|---|
| Высокая точность, автоматизация, скорость, адаптивность, широкий спектр применения | Уязвимость перед состязательными атаками, сильная зависимость от обучающего набора данных, необходимость ручного построения признаков и предобработки данных |

# Методы обнаружения с использованием глубокого обучения

| Название | Тип данных | Точность | Описание |
|---|---|---|---|
| DFFD | Видео, изображения | 99,4% | Обработка карт признаков для выделения измененных областей лица на основе CNN и применения механизма внимания |
| Анализ частотного спектра | Изображения | 99% | Обнаружение поддельных данных путем анализа частотного спектра на основе дискриминатора GAN |
| Инкрементное обучение CNN | Изображения | 99,8% | Постоянное обучение CNN для более точного обнаружения |
| EfficientCNN | Аудио | 99% | Анализ нормализованных спектрограмм для обнаружение поддельных аудио на основе метода CNN |

| Преимущества | Недостатки |
|---|---|
| Наилучшая точность обнаружения, построение глубоких признаков, адаптивность, широкий спектр применения | Уязвимость перед состязательными атаками, сильная зависимость от обучающего набора данных, необходимость предобработки данных |

# Другие методы

- Анализ метаданных
- Анализ водяных знаков
- Поиск оригиналов в сети Интернет
- Использование блокчейна
- Проведение атак логического вывода

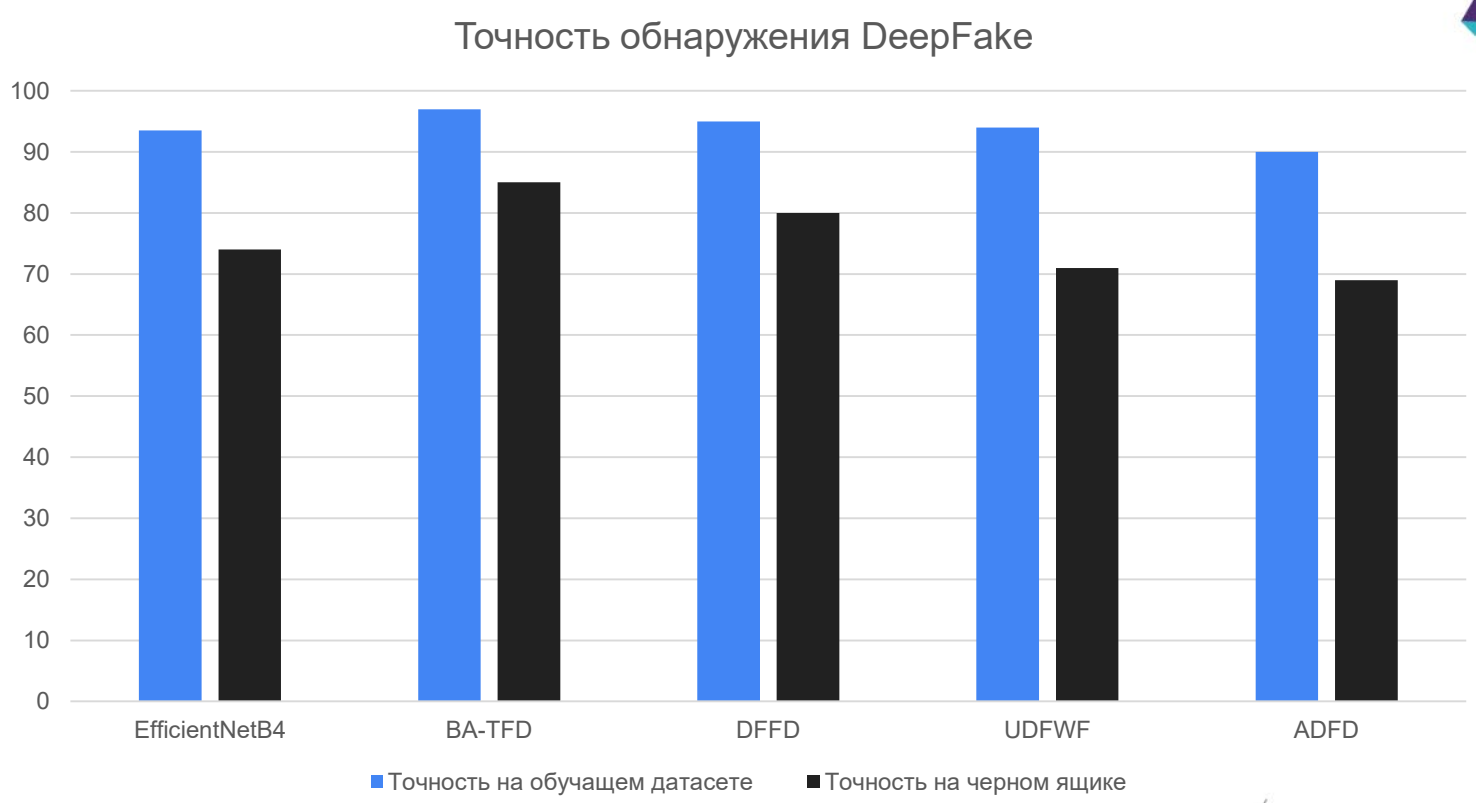
Данные методы являются вспомогательными и могут быть встроены в систему обнаружения DeepFake совместно с методами статистического анализа и использования искусственного интеллекта

# Методика проведения тестирования методов обнаружения DeepFake

- **Использование одного набора данных**
  - Разделение набора данных: 60% обучающая выборка и 40% тестируемая
  - Обучение и тестирование на одном и том же наборе данных
- **Тестирование черным ящиком**
  - Использование уже обученной модели
  - Использование при тестировании нескольких наборов данных
  - Добавление записей, реализующих состязательную атаку

| Модель обнаружения | Метод | Набор данных |
|---|---|---|
| EfficientNetB4 | CNN | FaceForensics++ |
| BA-TFD | CNN | LAV-DF |
| DFFD | CNN | CelebA |
| Unmasking DeepFakes with simple Features | SVM, K-Means | CelebA |
| ADFD | CNN | WaveFake |

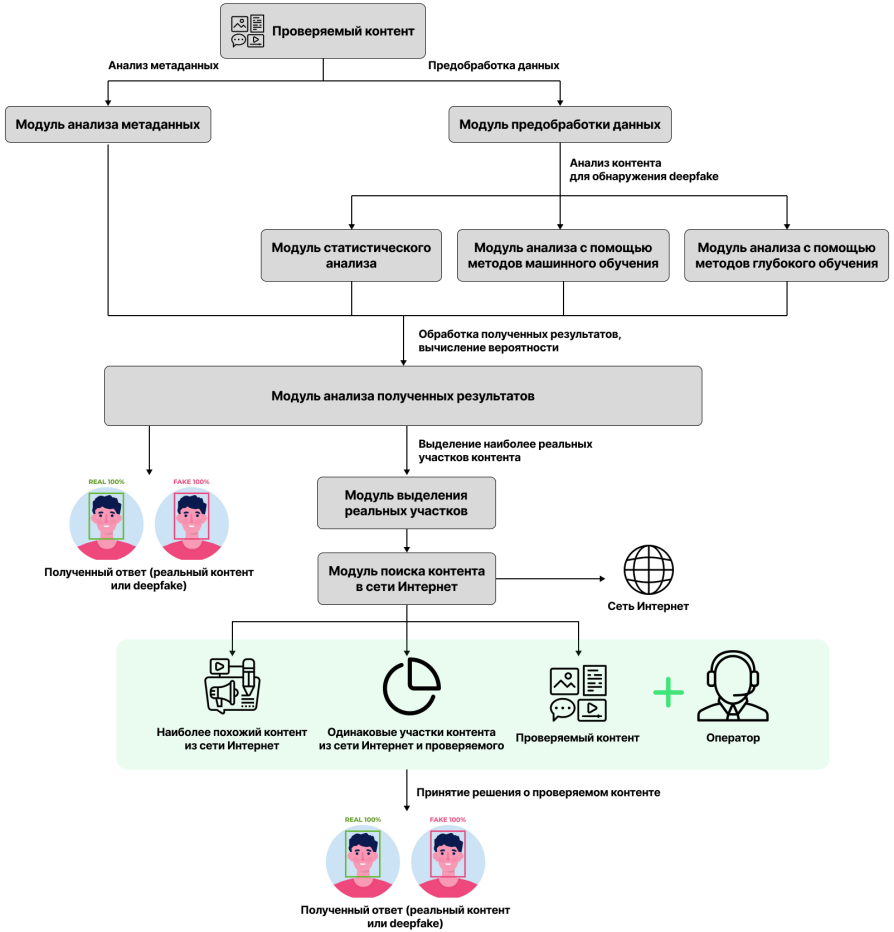
# Оценка эффективности методов обнаружения DeepFake




ПОЛИТЕХ

МЫСЛИТЬ БУДУЩИМ

# Предлагаемая архитектура системы обнаружения DeepFake




## Модуль анализа метаданных

## Модуль предобработки данных

## Модули обнаружения DeepFake

- Статистический анализ
- Машинное обучение
- Методы глубокого обучения

## Модуль анализа полученных результатов

## Модуль выделения реальных участков

## Модуль поиска контента в сети Интернет

# Выводы

1. Применение методов глубокого обучения позволяет наиболее эффективно обнаружить искусственно синтезированный контент.
2. Несмотря на успешное применение методов глубокого обучения существуют актуальные проблемы падения точности обнаружения DeepFake при использовании состязательных атак или при взаимодействии с данными из неоднородных датасетов.
3. Гарантированный результат выявления искусственно синтезированного контента может быть получен при нахождении оригинальных данных.
4. Предотвращение распространения DeepFake контента может быть реализовано путем встраивания систем обнаружения искусственно синтезированных данных в сервисы сети Интернет, в которых осуществляется загрузка контента (социальные сети, видеохостинги, новостные порталы и т.д.)
5. В результате доклада предложена архитектура гибридной системы обнаружения DeepFake, построенная на нахождении оригинала.