Научно-технологический университет Сириус

Яндекс

Приватное машинное обучение

Денисенко Наталья, ВШЭ, ФКН
Миронов Алексей, ВШЭ, ФКН
Сидоренко Артур, МГУ, мехмат

Руководители проекта:
Деркач Денис, PhD, ВШЭ
Казеев Никита, ВШЭ
Устюжанин Андрей, к.ф.-м.н., ВШЭ

Научно-технологический университет Сириус

Задача

Яндекс

Владелец

Данные

Третье лицо

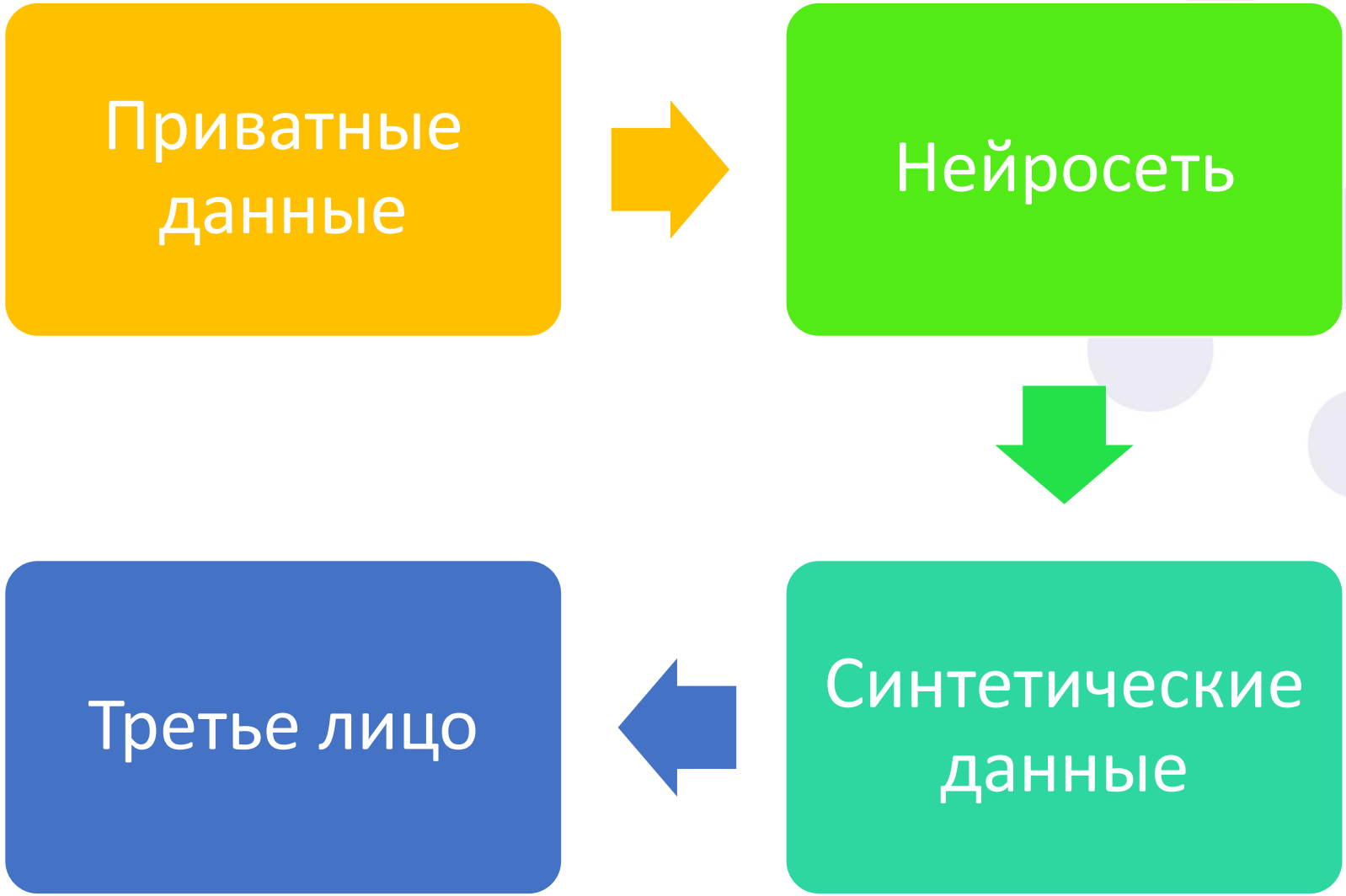
kaggle

amazon web services™

Научно-технологический университет Сириус

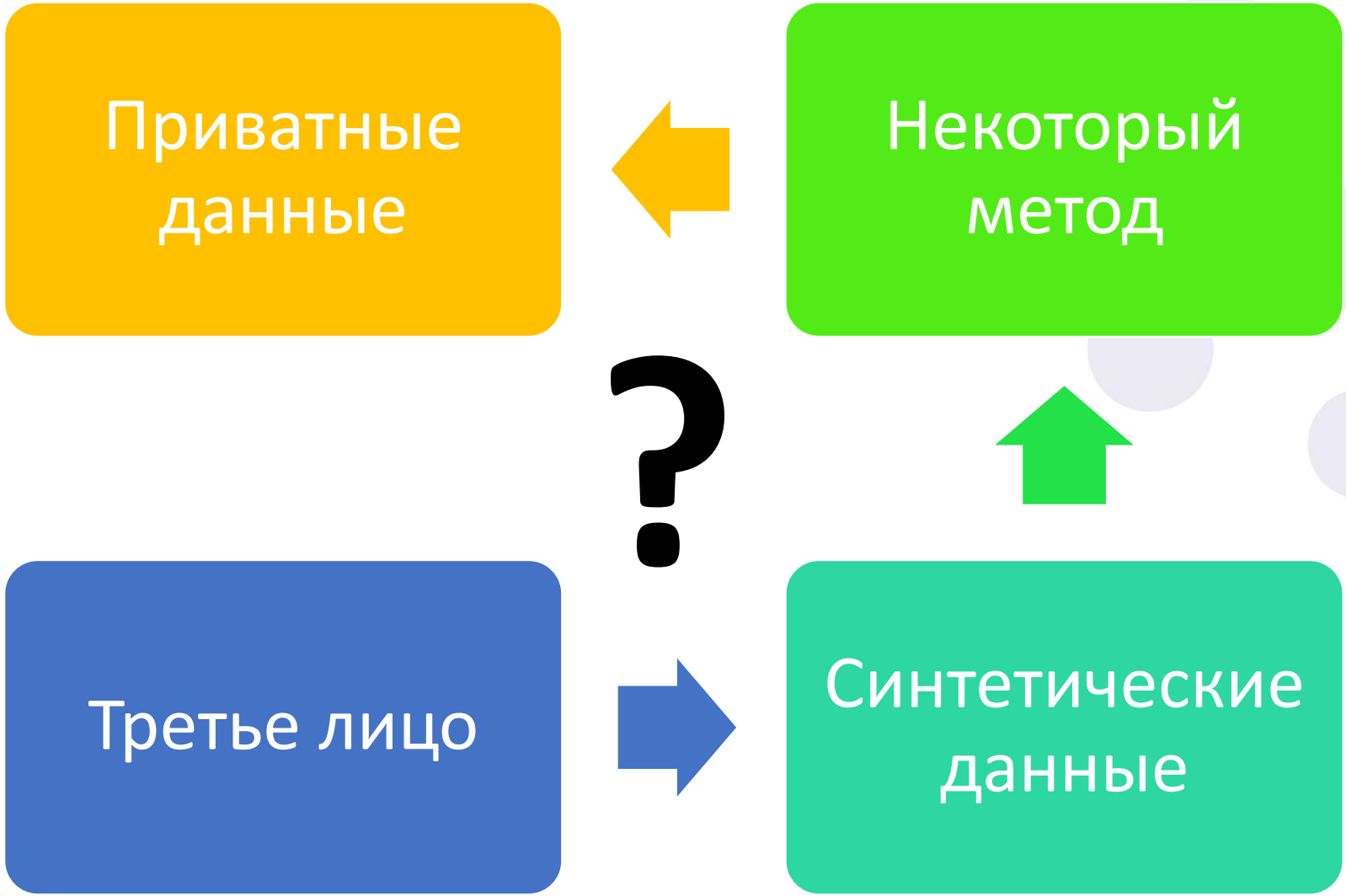
Основная идея

Яндекс



Научно-технологический университет Сириус

Основная проблема Яндекс



Научно-технологический университет Сириус

Актуальность проблемы

Яндекс

- Массовая обеспокоенность сохранением приватности данных
- Утечки частных сведений

Научно-технологический университет Сириус

Актуальность проблемы

Яндекс

MLaaS – Machine Learning as a Service

Облачные сервисы для ML??

Могут ли данные утечь с чужого сервера????

Научно-технологический университет Сириус

Цели работы

Яндекс

- Проанализировать статьи по приватному обучению
- Воспроизвести основные результаты
  - Обучение нейросетей
  - Атаки на нейросети
  - Повышение устойчивости к атакам

Научно-технологический университет Сириус

План презентации

Яндекс

1. Определение GAN
2. Анализ статей
3. Отбор данных
4. Обучение различных GAN (Silly GAN, DCGAN)
5. Проведение атак на нейронные сети
6. Сравнение результатов
7. Выводы

Научно-технологический университет Сириус

Яндекс

Определение GAN

Научно-технологический университет Сириус

GAN

Яндекс


Генератор

Дискриминатор

Плохо / хорошо

Научно-технологический университет Сириус

Яндекс

Анализ статей

Научно-технологический университет Сириус

Яндекс

Статьи

1. LOGAN: Membership Inference Attacks Against Generative Models (https://arxiv.org/pdf/1705.07663.pdf)
   - Методы проведения атак
2. PATE-GAN: GENERATING SYNTHETIC DATA WITH DIFFERENTIAL PRIVACY GUARANTEES (https://openreview.net/pdf?id=S1zk9iRqF7)
   - Больше математики
   - PATE-GAN
   - Схема контроля приватности

Научно-технологический университет Сириус

Яндекс

Отбор датасетов

Научно-технологический университет Сириус

Отбор датасетов

Яндекс

MNIST

kaggle

Kaggle Credit Scoring

Научно-технологический университет Сириус

Яндекс

Обучение GAN для MNIST

Научно-технологический университет Сириус

Структура Silly GAN

Яндекс

Генератор

- Dense (ELU)
- Dropout
- Conv2d (kernel_size=3, ELU)
- Dropout
- ConvTranspose2d
- Dropout
- Conv2d и Dropout
- Conv2d и Dropout

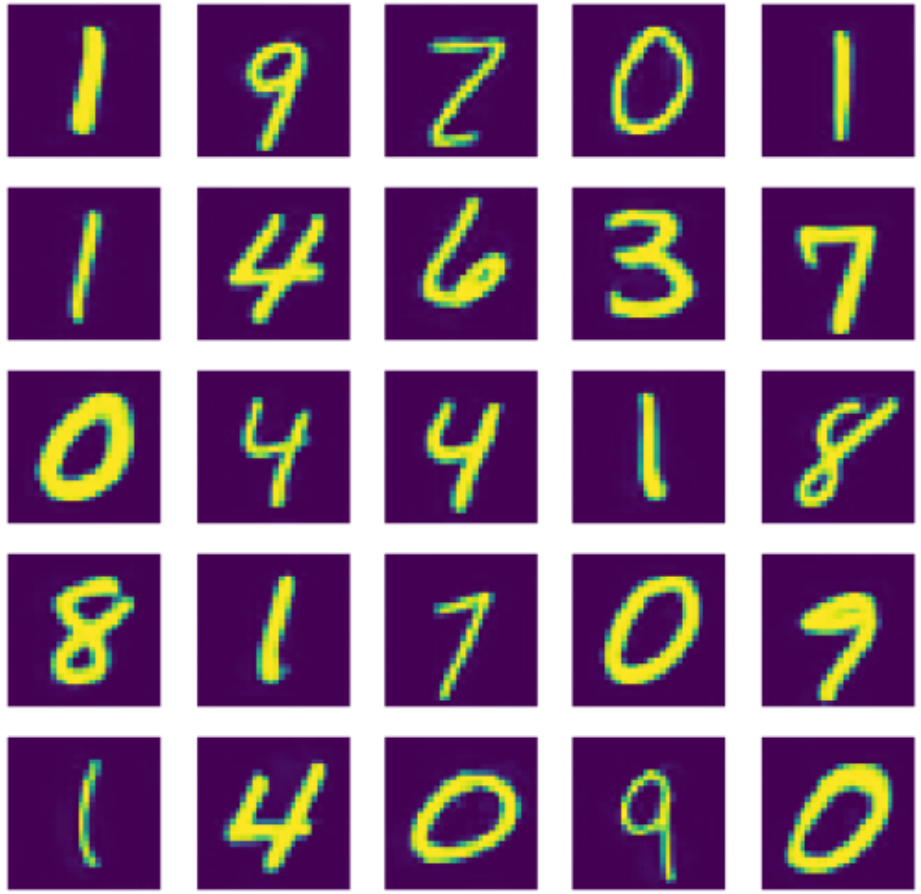
Дискриминатор

- Conv2d (ELU, kernel_size=3)
- Dropout
- MaxPool2d(2)
- Conv2d(ELU,kernel_size=3) и Dropout
- Conv2d(ELU,kernel_size=3) и Dropout
- Conv2d(ELU,kernel_size=3) и Dropout
- Linear() with sigmoid

Научно-технологический университет Сириус

Обучение своей Silly GAN

Яндекс



- Jensen-Shannon GAN со шумом в losses, регуляризацией градиента и техникой Dropout
- Обучение происходило на сервере ШАД almaren
- Сгенерированные цифры слева

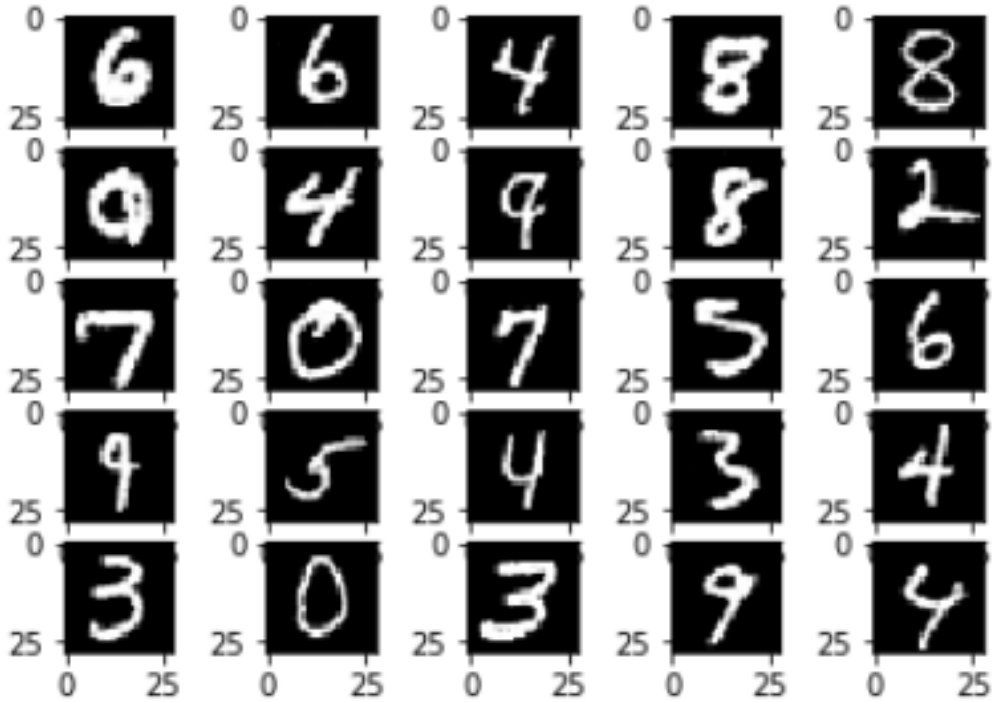
PyTorch

GitHub

Научно-технологический университет Сириус

DCGAN for MNIST

Яндекс



- Использовали предобученную модель (ссылка)
- Jensen-Shannon без шумов, регуляризаций и Dropout

PyTorch

GitHub

Научно-технологический университет Сириус

Яндекс

Обучение GAN для CreditFraud

Научно-технологический университет Сириус

Структура GAN

Яндекс

Генератор

- Полносвязные слои с ReLU активацией

Дискриминатор

- Полносвязные слои с ReLU активацией

Взято со статьи PATE-GAN, стр.15, там ссылка на репозиторий

Научно-технологический университет Сириус

Результаты обучения

Яндекс

В статье было указано, что сеть обучается и переобучается, но у нас GAN не смогла научиться воспроизводить табличные данные. Далее исследовали только GAN, обученные на MNIST.

Взято со статьи PATE-GAN, стр.15, там ссылка на репозиторий

Научно-технологический университет Сириус

Яндекс

Атаки на GAN

Научно-технологический университет Сириус

Нейросеть как массовый подгон

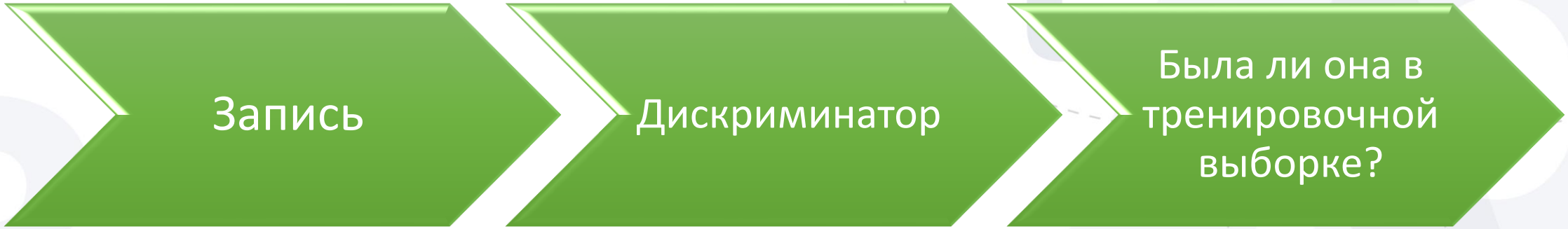
Яндекс

- Слои
- LOSS
- Оптимизатор
- Learning rate

Научно-технологический университет Сириус

Атака на GAN. Общая идея

Яндекс



Научно-технологический университет Сириус

Атака на GAN. Общая идея

Яндекс

А дискриминаторы на деревьях растут???

Запись → Дискриминатор → Была ли она в тренировочной выборке?

Научно-технологический университет Сириус

Атака на GAN

Яндекс

А дискриминаторы на деревьях растут???

ДА!

- White-box attack

Нет, конечно, сам делай

- Black-box attack (with or without leakages)

Научно-технологический университет Сириус

Атака на GAN

Яндекс

Black-box + leakage

Black-box

White-box

Научно-технологический университет Сириус

Яндекс

Атаки на GAN. White-box

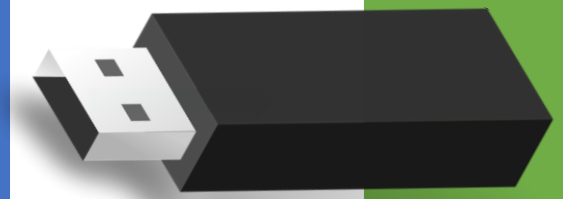
Научно-технологический университет Сириус

White-box attack

Яндекс

- **Задача**. Пусть преступник получил дискриминатор, и он хочет узнать, является ли данный элемент частью тренировочной выборки.

Дискриминатор



Преступник

Научно-технологический университет Сириус

White-box attack

Яндекс

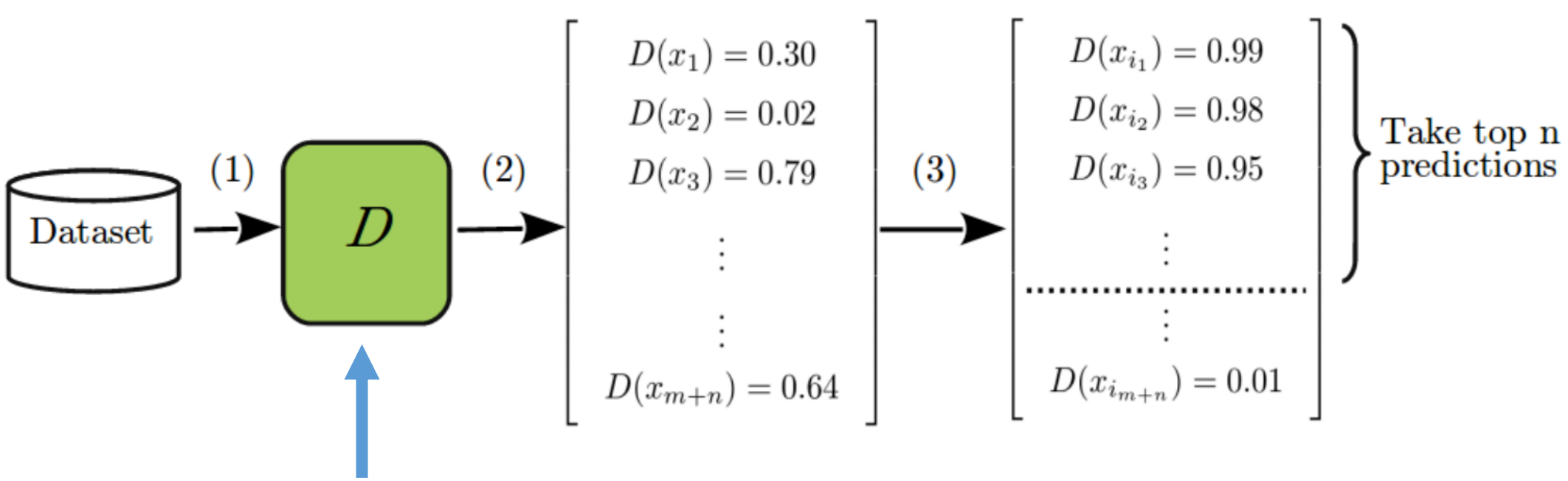
- **Задача**. Пусть преступник получил дискриминатор, и он хочет узнать, является ли данный элемент частью тренировочной выборки.
- **Идея**: переобученный дискриминатор сохраняет в себе информацию о тренировочной выборке.

Научно-технологический университет Сириус

White-box attack

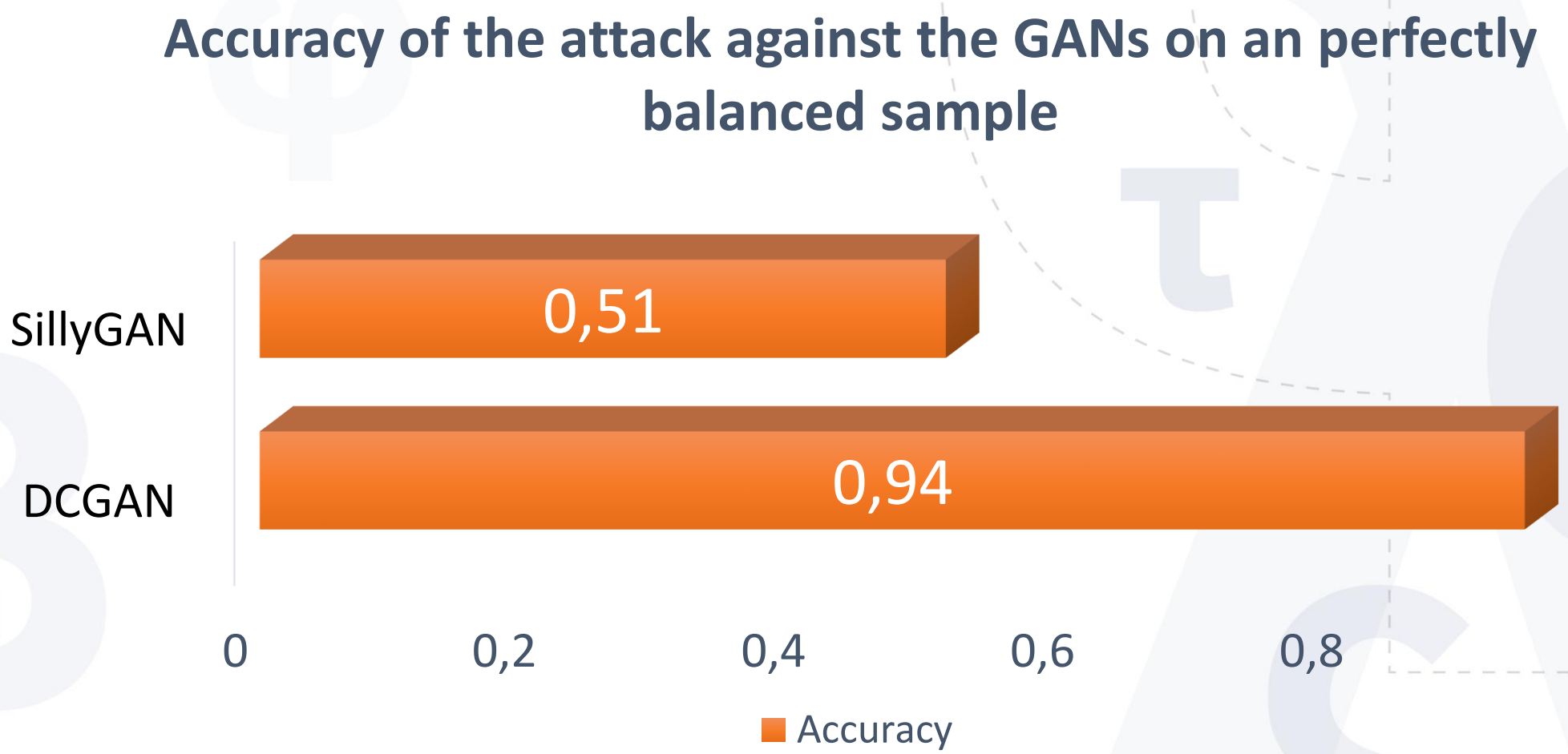
Яндекс

Отклик дискриминатора



Дискриминатор

White-box attack



Научно-технологический университет Сириус

Яндекс

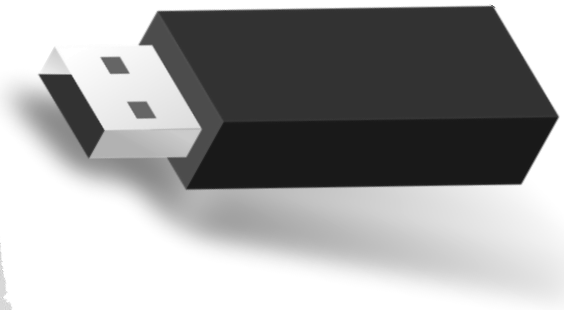
Атаки на GAN. Black-box + leakage

Научно-технологический университет Сириус

Black-box attack with a 20% leakage

Яндекс

- **Задача**. Произошла утечка части обучающей выборки. Злоумышленник имеет доступ к генератору.



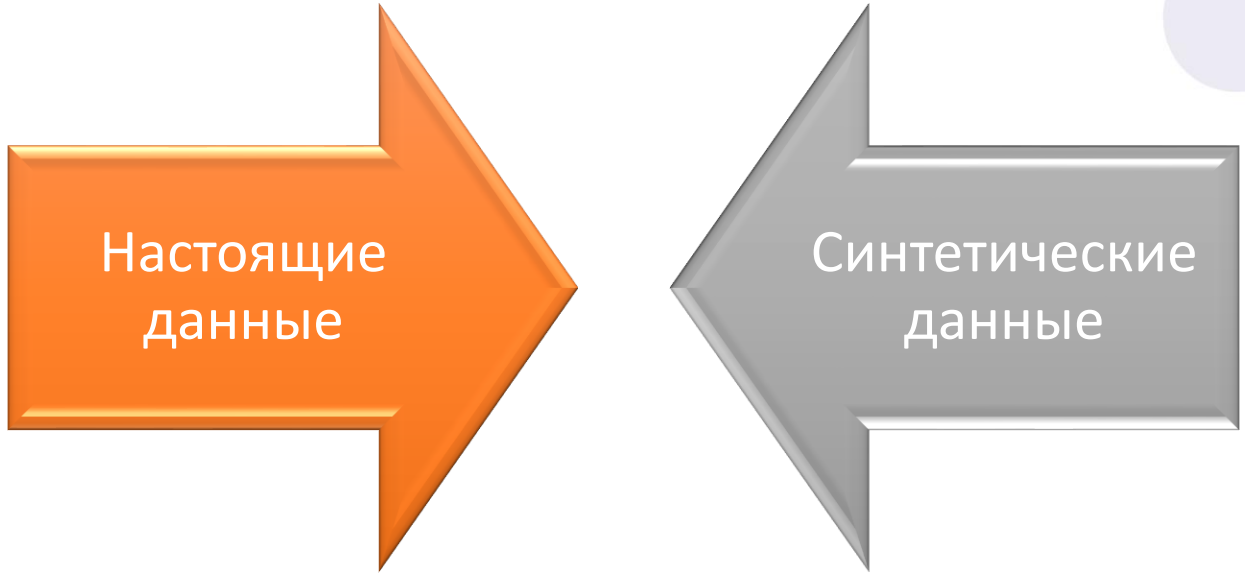
Преступник

Научно-технологический университет Сириус

Black-box attack with a 20% leakage

Яндекс

- **Задача**. Произошла утечка части обучающей выборки. Злоумышленник имеет доступ к генератору.
- **Идея**. Создать свой классификатор.

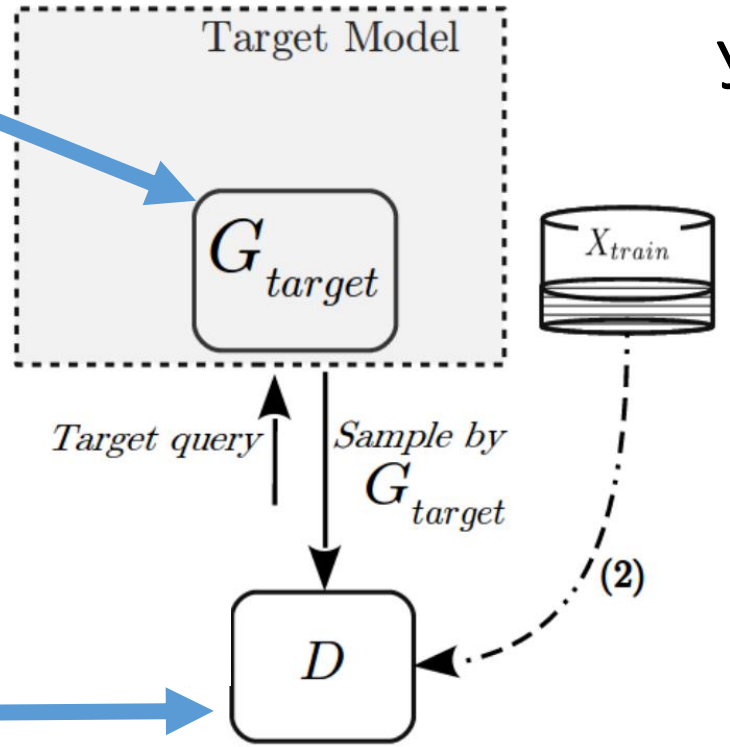


Научно-технологический университет Сириус

Black-box attack with a 20% leakage

Яндекс

Генератор изначальной модели



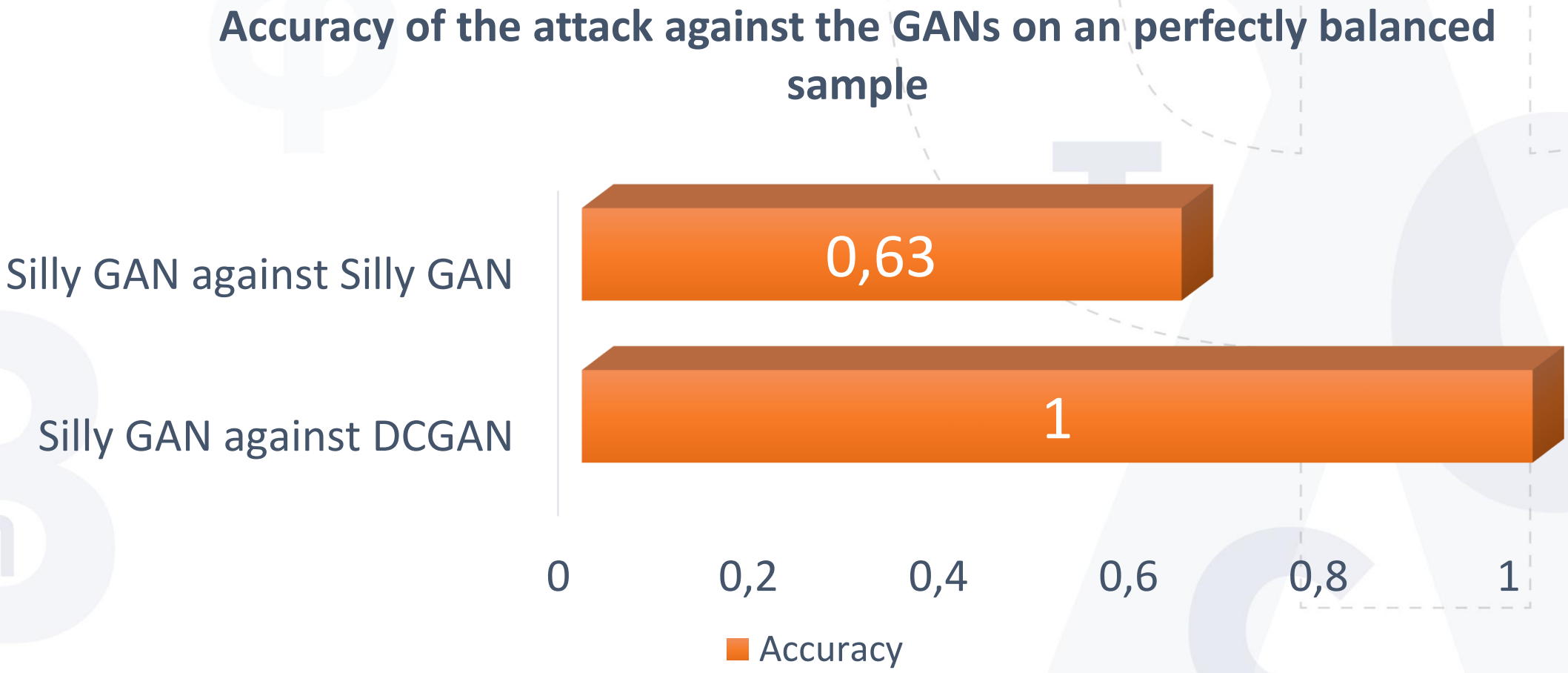
Утечка данных

Обученный преступником классификатор

Научно-технологический университет Сириус

Black-box attack with a 20% leakage

Yandex



Научно-технологический университет Сириус

Яндекс

Атаки на GAN. Black-box without any leakage

Black-box attack without any leakage

- **Задача.** Злоумышленник имеет доступ к генератору, но не к обучающей выборке.

Научно-технологический университет Сириус

Black-box attack without any leakage

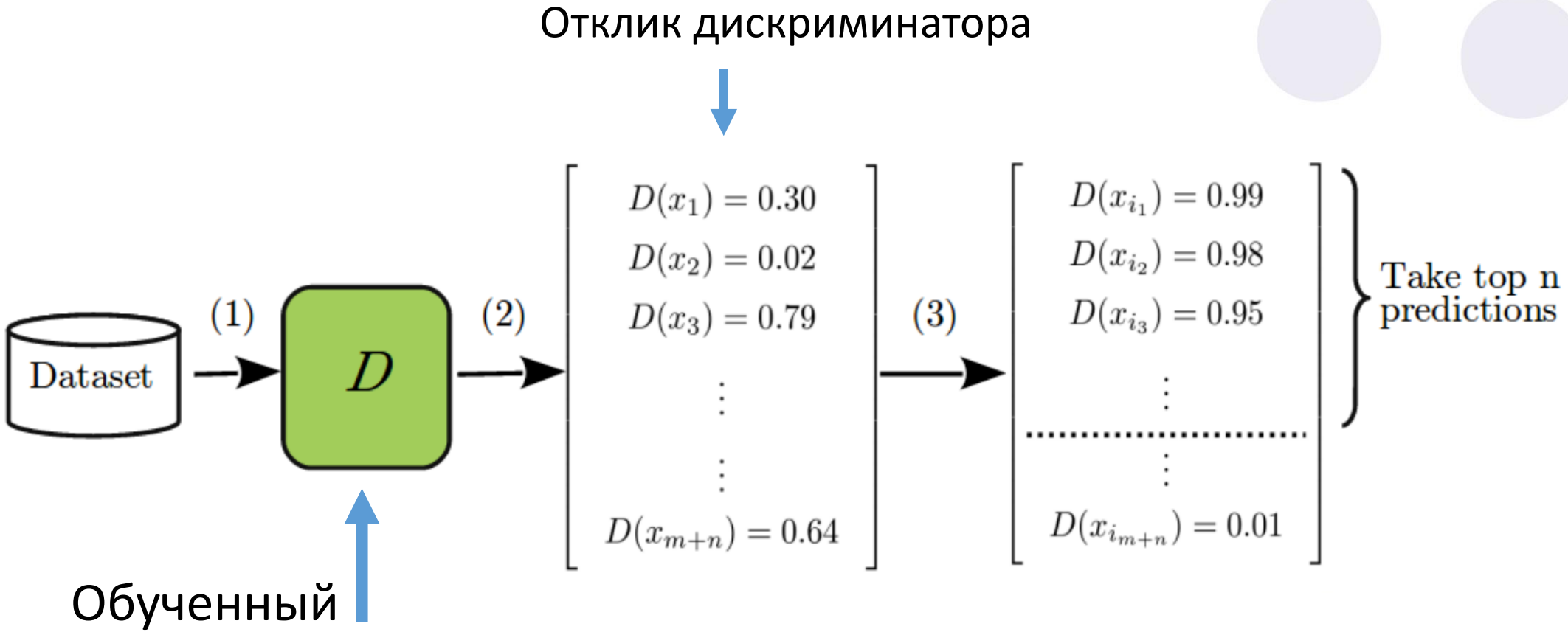
Яндекс

- **Задача.** Злоумышленник имеет доступ к генератору, но не к обучающей выборке.
- **Идея.** Обучить доморощенный GAN.

Black-box attack without any leakage

- **Задача.** Злоумышленник имеет доступ к генератору, но не к обучающей выборке.
- **Идея.** Обучить доморощенный GAN.
- **Метод.** Обучается GAN на синтетической выборке, как если бы это были настоящие данные.

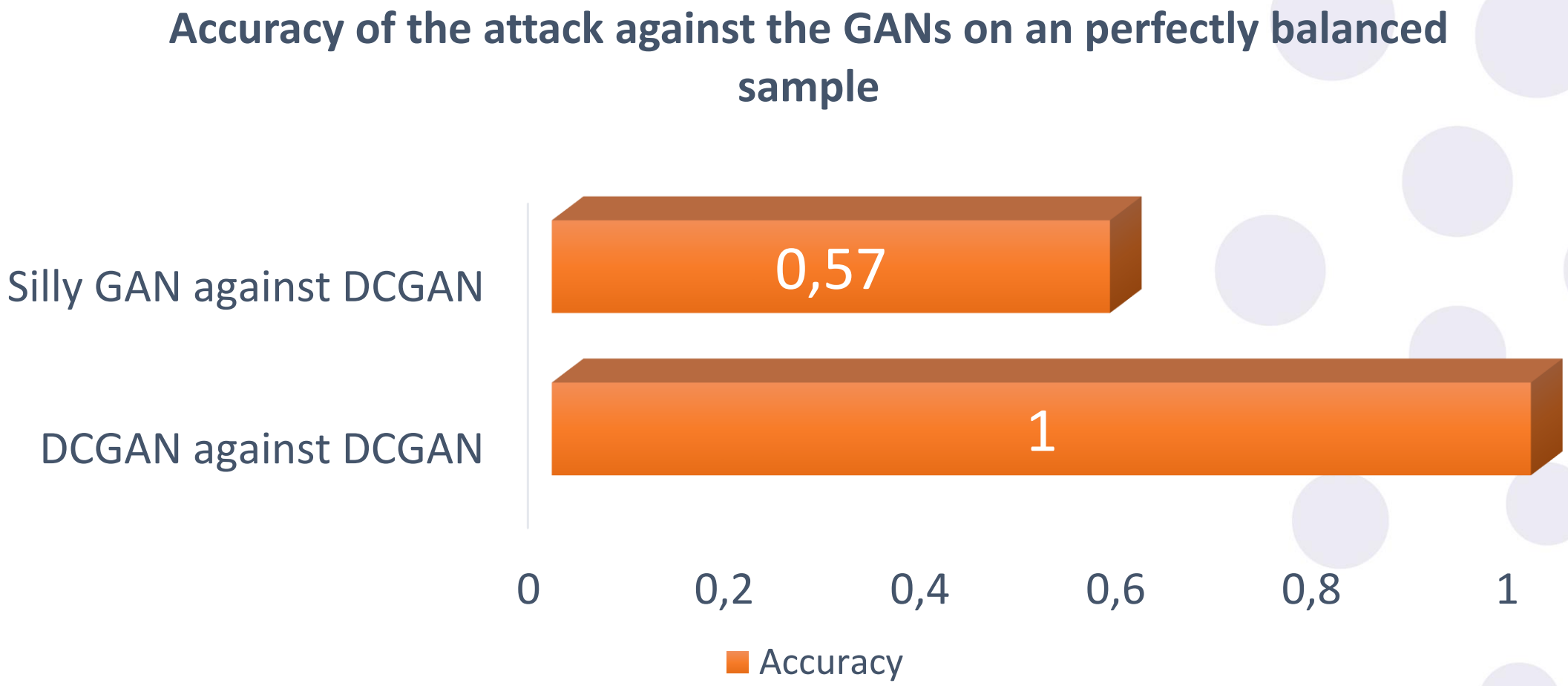
Black-box attack without any leakage



Научно-технологический университет Сириус

Black-box attack without any leakage

Yandex



Научно-технологический университет Сириус

Яндекс

Сравнение результатов

Научно-технологический университет Сириус

Сравнение качества атак

Яндекс

Black-box + leakage

- Наиболее успешная атака, но требует утечку данных

Black-box

- Результат сильно зависит от навыков злоумышленника

White-box

- Атака может быть успешной только при переобученном дискриминаторе

Научно-технологический университет Сириус

Сравнение DCGAN и Silly GAN

Яндекс

Silly GAN

- Устойчива к атакам

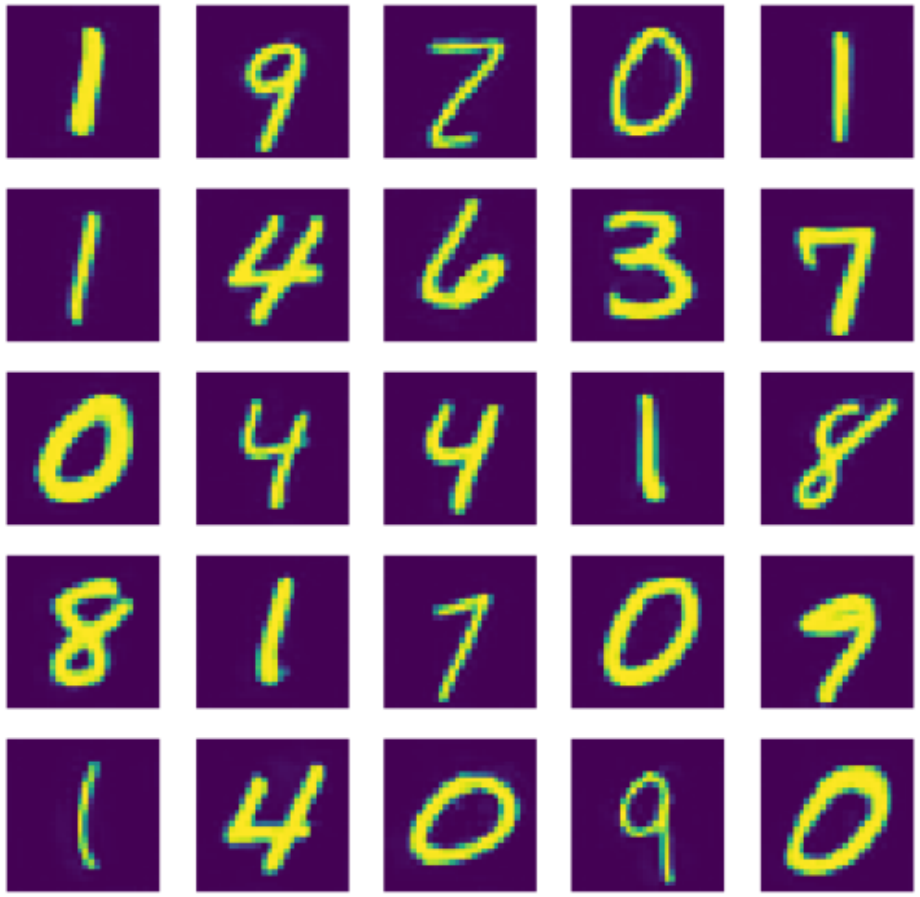
DCGAN

- Может взломать ребёнок

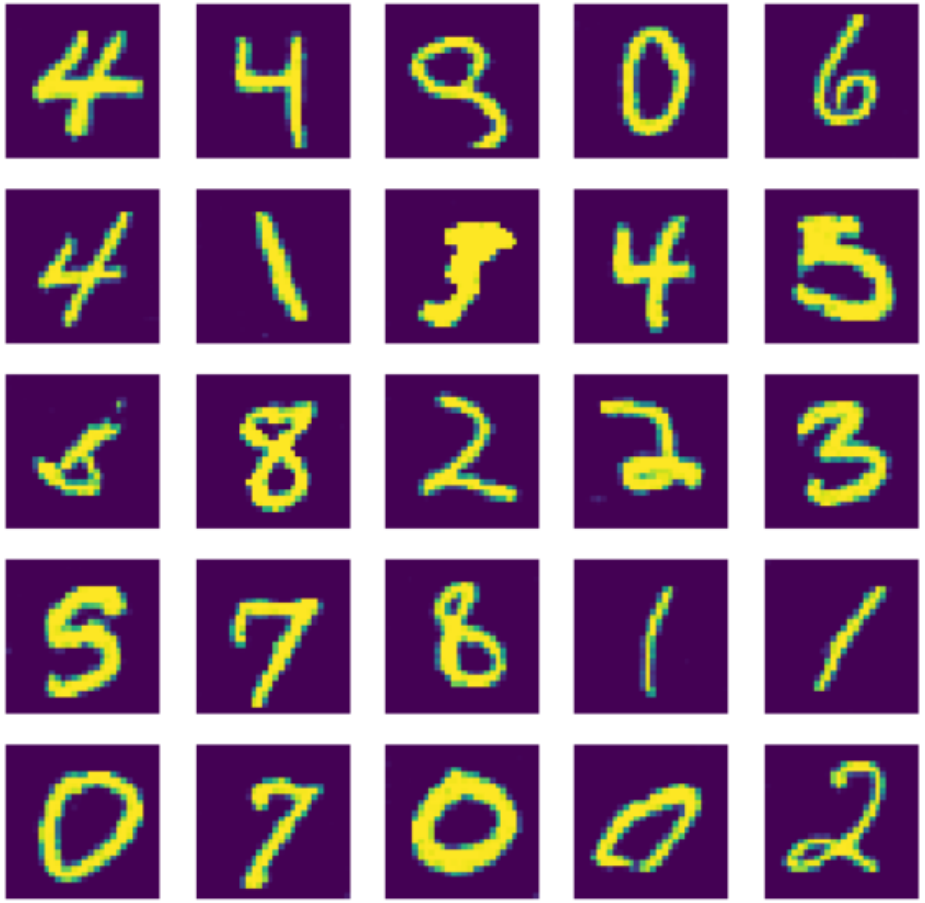
Научно-технологический университет Сириус

Сравнение DCGAN и Silly GAN

Яндекс



Silly GAN



DCGAN

Научно-технологический университет Сириус

Как противостоять атакам?

Яндекс

- Борьба с переобучением
  - ✓Dropout techniques
  - ✓Зашумление loss
  - ✓Регуляризация градиента

Научно-технологический университет Сириус

Как противостоять атакам?

Яндекс

- Борьба с переобучением
  - ✓Dropout techniques
  - ✓Зашумление loss
  - ✓Регуляризация градиента
- Баланс генератора и дискриминатора

Научно-технологический университет Сириус

Как противостоять атакам?

Яндекс

- Борьба с переобучением
  - ✓Dropout techniques
  - ✓Зашумление loss
  - ✓Регуляризация градиента
- Баланс генератора и дискриминатора
- Имитирование атак во время обучения, перед выпуском модели

Научно-технологический университет Сириус

Яндекс

Заключение

Научно-технологический университет Сириус

Что же мы сделали?

Яндекс

- Научились делать GAN с нуля
- Проанализировали статьи по атакам на GAN

Научно-технологический университет Сириус

Что же мы сделали?

Яндекс

- Научились делать GAN с нуля
- Проанализировали статьи по атакам на GAN
- Сделали несколько атак на GAN
- Сравнили качество разных типов атак на GAN

Научно-технологический университет Сириус

Что же мы сделали?

Яндекс

- Научились делать GAN с нуля
- Проанализировали статьи по атакам на GAN
- Сделали несколько атак на GAN
- Сравнили качество разных типов атак на GAN
- Поняли, что даже GAN со страшными названиями **легко** поддаются атакам
- Разработали рекомендации по созданию устойчивых к атакам сетей

Научно-технологический университет Сириус

Перспективы и планы на будущее

Яндекс

- Экстенсивное исследование связи переобучения и уязвимости к атакам

Научно-технологический университет Сириус

Перспективы и планы на будущее

Яндекс

- Экстенсивное исследование связи переобучения и уязвимости к атакам
- Математическая теория дифференциальной приватности
  - Работает ли она на практике?
  - Зачем нужны сложные методы, когда и простые, но грамотно составленные нейросети работают?

Научно-технологический университет Сириус

Перспективы и планы на будущее

Яндекс

- Экстенсивное исследование связи переобучения и уязвимости к атакам
- Математическая теория дифференциальной приватности
  - Работает ли она на практике?
  - Зачем нужны сложные методы, когда и простые, но грамотно составленные нейросети работают?
- Ведёт ли сохранение приватности к ухудшению качества?

Научно-технологический университет Сириус

Яндекс

Спасибо за внимание!

Задавайте вопросы

Ссылка на наш репозиторий в Github: https://github.com/Private-ML/gan

Научно-технологический университет Сириус

Differential privacy

Yandex

- Two datasets D and D’ are said to be neighbouring if there is such x in D that $D \setminus \{x\} = D'$.
- Let $(\Omega, \mathcal{F}, P)$ be a probability space, let (X, S) be a measurable space representing datasets, and let (Y, T) be a measurable space of outcomes.
- Let M be a *randomised algorithm*, i.e. a measurable mapping

$$M: X \times \Omega \to Y$$

The algorithm is said to be (ε, δ)-private, if for all neighbouring datasets D and D’ and for all sets $Y \in T$: $P(M(D) \in Y) < e^{\varepsilon} P(M(D') \in Y) + \delta$

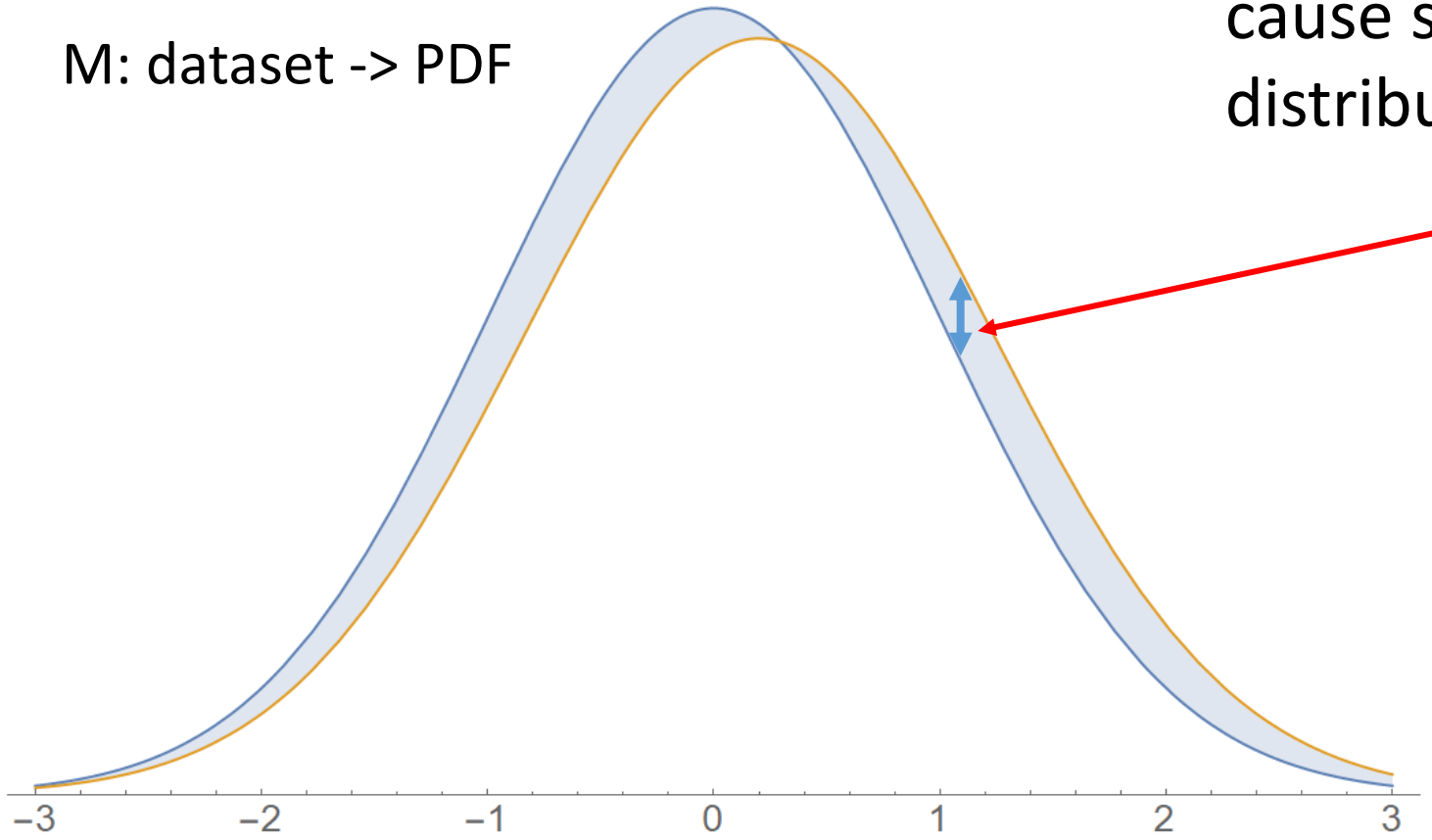
Научно-технологический университет Сириус

Differential privacy

Yandex

M: dataset -> PDF

Slight changes of the dataset do not cause significant difference in the distribution of the output



Научно-технологический университет Сириус

Научно-технологический университет Сириус

Научно-технологический университет Сириус

Научно-технологический университет Сириус

Научно-технологический университет Сириус

Научно-технологический университет Сириус

Научно-технологический университет Сириус

Научно-технологический университет Сириус

Научно-технологический университет Сириус