Ежегодная международная научно-практическая конференция
«РусКрипто’2021»

Данные, интеллект и безопасность

Григорий Маршалко

Arvind Narayanan
@random_walker

The industry AI paradigm assumes that it's possible for companies to build models using personal data scooped up from the web—without asking who shared it and why—and somehow make it ethical using technical tricks. If we reject this premise the trillion dollar castle falls apart.

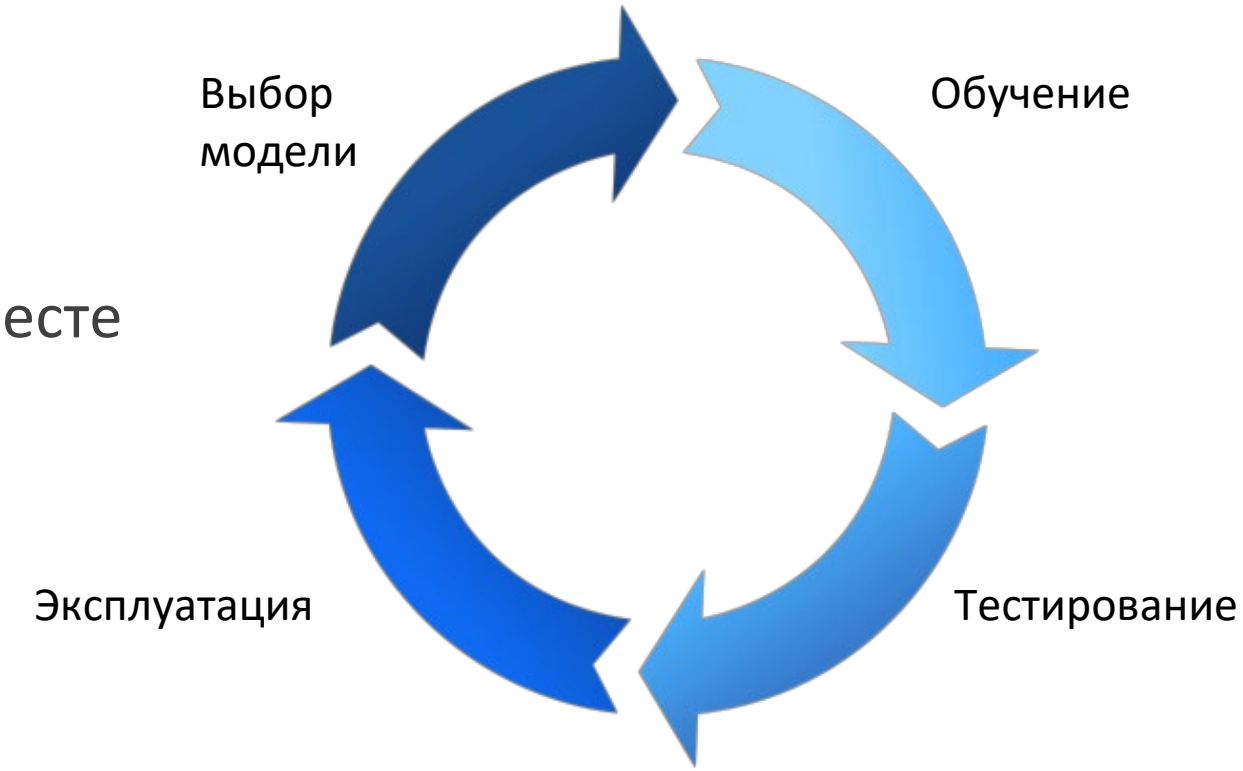
Перевести твит

5:05 PM · 23 июл. 2020 г. · Twitter Web App

Машинное обучение

- Алгоритм решения получается вместе с решением

- Статистические методы
- Нейросетевые модели



Машинное обучение. Д

- Большие объемы данных
- Графические и нейровычислители
- Свободно распространяемые библиотеки
- Предобученные модели

- Достаточно просто получить впечатляющий результат

А что насчет…

- Объяснимости
- Точности
- Смещения
- Робастности
- Целостности
- Конфиденциальности

А что насчет…

- Объяснимости
- Точности
- Смещения
- Робастности
- Целостности
- Конфиденциальности

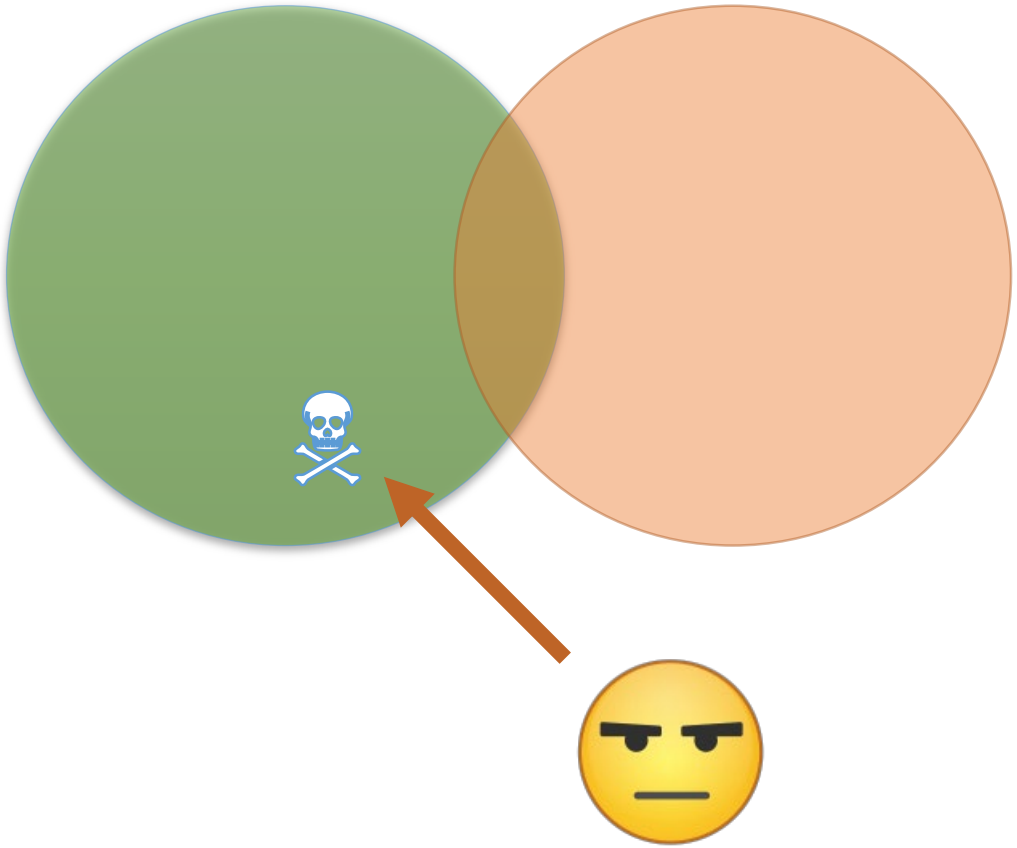
- Выбор метрик
- Выбор моделей
- Настройка алгоритмов обучения
- Тестирование
- Логирование

А что насчет…

- Объяснимости
- Точности
- Смещения
- Робастности
- Целостности
- Конфиденциальности

- Обучение
  - Отравление данных
- Эксплуатация
  - Атаки уклонения
  - Извлечение данных из обученной модели

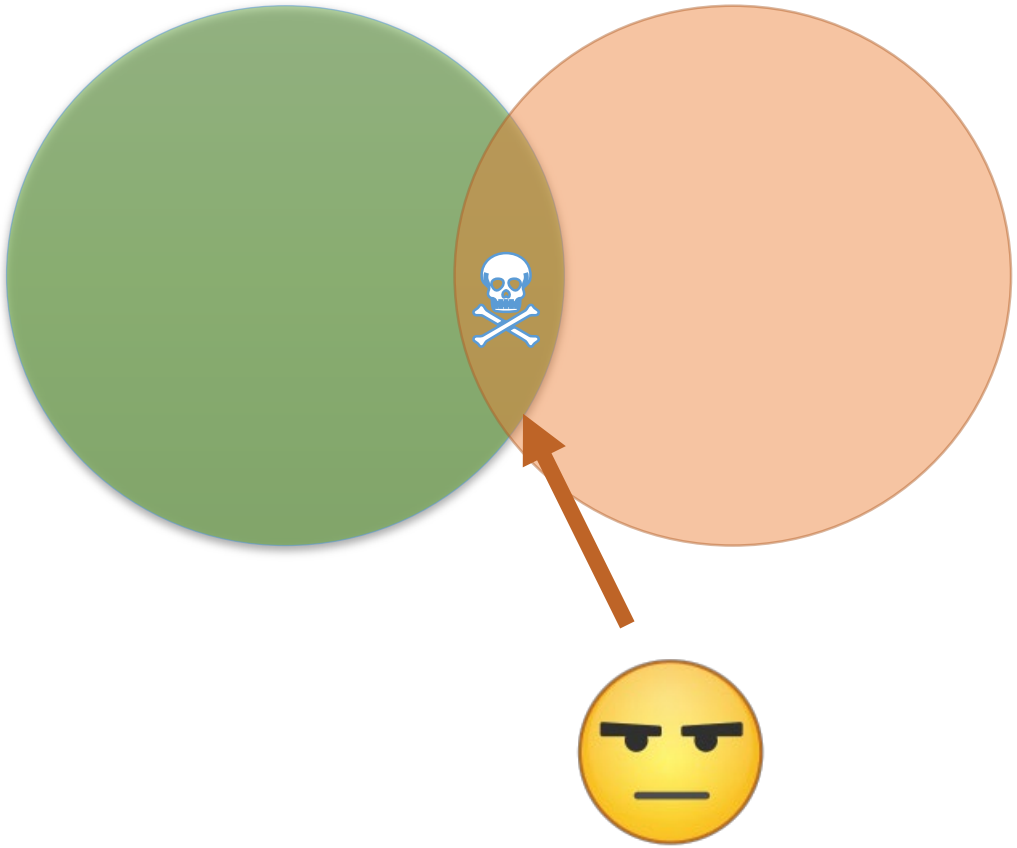
Отравление данных



- Удаление вредоносных примеров из обучающих выборок
- Контроль происхождения данных
- Выявление нейросетевых троянов

- Не обеспечивают 100% защиты

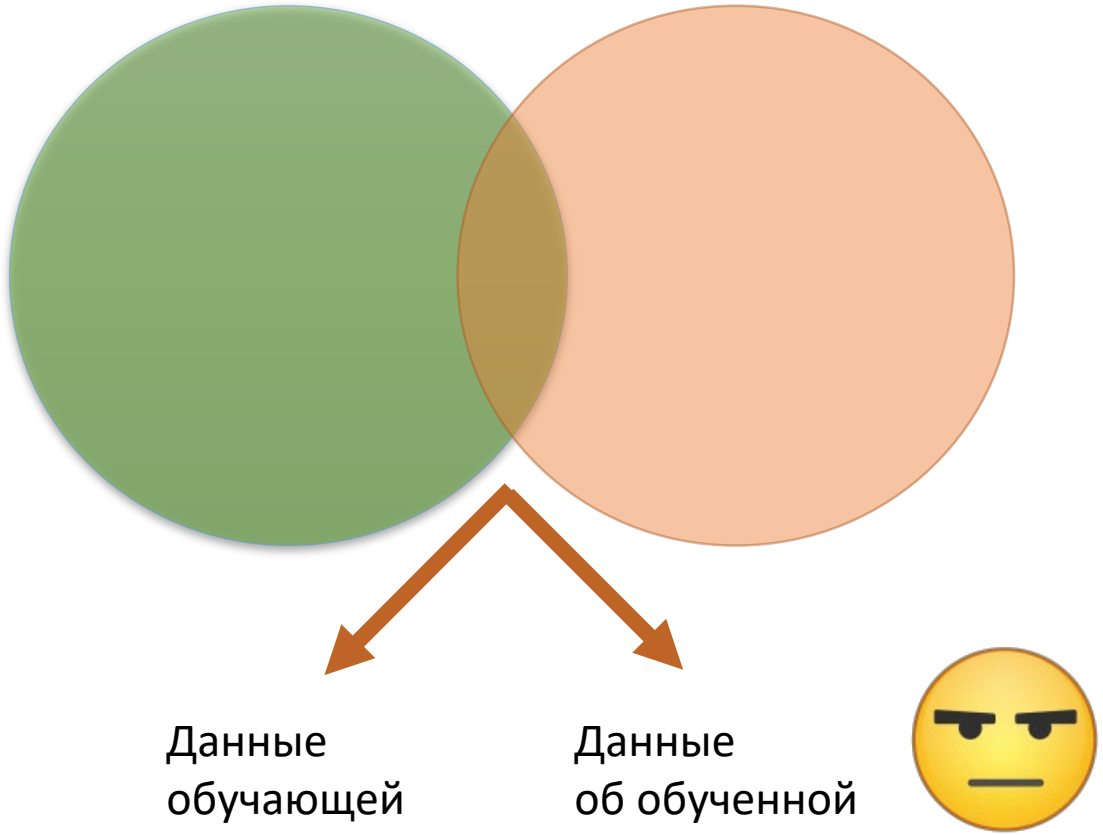
Атаки уклонения



- Обучение робастных моделей
- Состязательное обучение
- Предобработка входных данных
- Дистилляция

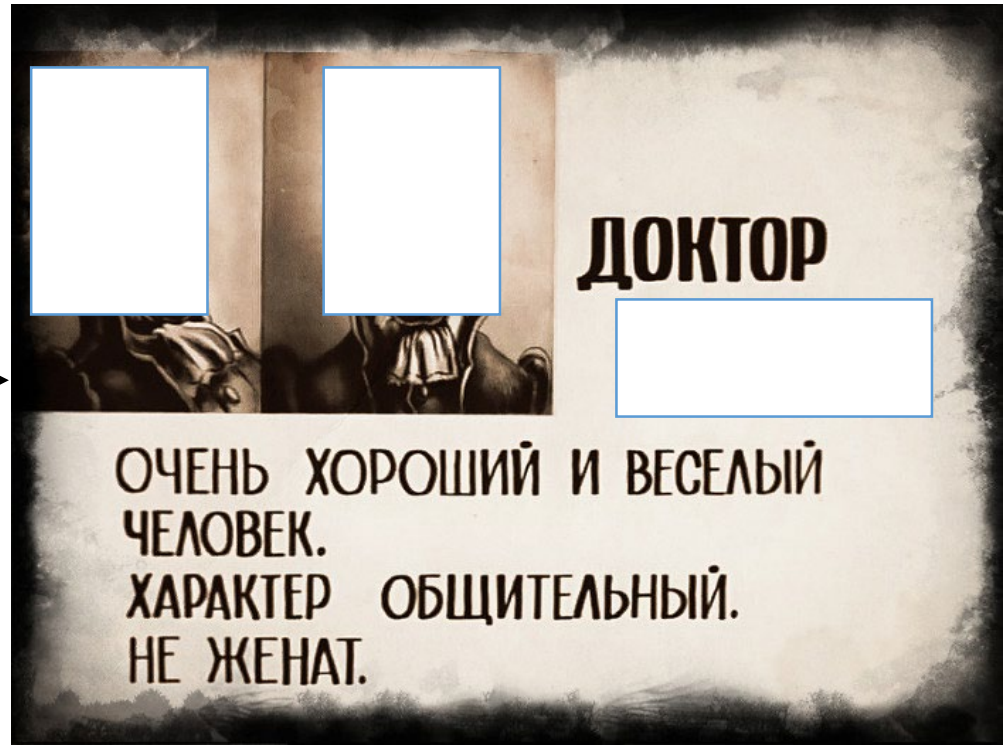
- Не обеспечивают 100% защиты

Извлечение данных из модели



- Обезличивание
- Конфиденциальное машинное обучение
  - гомоморфное шифрование
  - протоколы безопасных распределенных вычислений

Обезличивание



Обезличивание. Квазиидентификаторы!

Это же доктор Ливси!

Обезличивание. Методы агрегирования

- k-анонимность
- l-разнообразие
- t-близость
- …

- каждое значение квазиидентификатора должно соответствовать нескольким записям

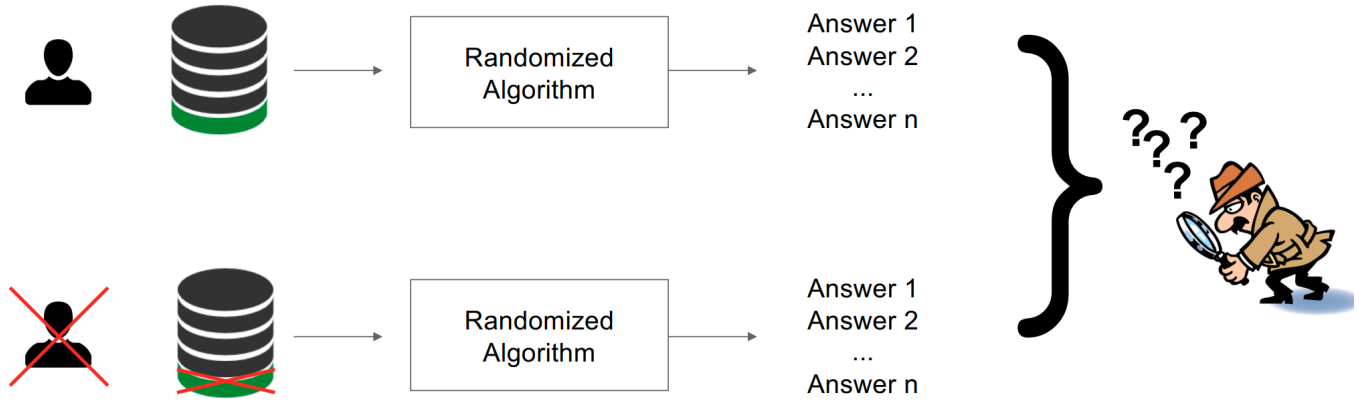
- агрегируют квазиидентификаторы
- задача внесения минимального искажения NP-трудная
- работоспособны только для баз с небольшим числом атрибутов
- подвержены широкому спектру атак (особенно при наличии дополнительных данных) *выделения*, *связывания*, *извлечения данных*
- ухудшает точность аналитики

Не обеспечивают гарантированной защиты

Статистическое обезличивание

Differential privacy

- Маскирование шумом входных и промежуточных данных алгоритмов аналитики
- При определенных условиях гарантируют невозможность определения информации об одном пользователе
- Количественные оценки безопасности защиты



Статистическое обезличивание

Differential privacy

- Маскирование шумом входных и промежуточных данных
- При определенных условиях гарантируют невозможность определения информации об одном пользователе
- Количественные оценки безопасности защиты

- защищает не данные, а алгоритм их обработки

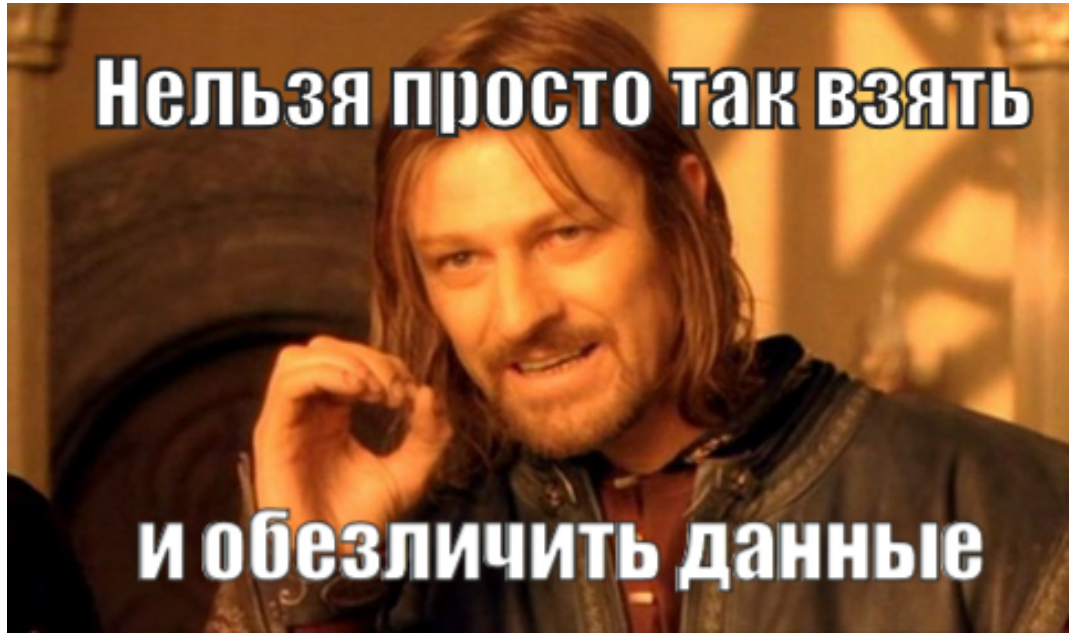
(доверенный оператор, недоверенный аналитик)

- требует контроля за количеством запросов
- более удобны для работы с большими данными
- ухудшает точность аналитики на заданную величину
- все равное есть атаки

Не обеспечивает полной защиты, но может хоть что-то гарантировать

Обезличивание

- Обезличивание не дает 100% результата
- При определенных условиях позволяет отсрочить момент деобезличивания
- Необходим контроль исходных условий использования
- Необходим контроль использования обезличенных данных
- Лучше защищать алгоритм анализа, а не данные



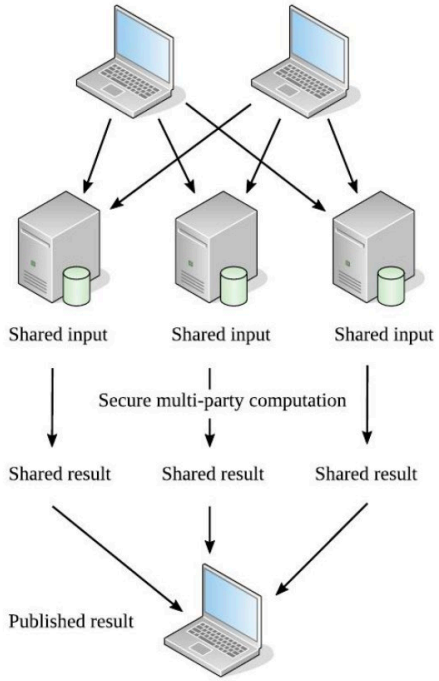
Конфиденциальное машинное обучение. Гомоморфное шифрование

- Гомоморфное шифрование
  - вычисления над шифрованным текстом
  - частично гомоморфное (по одной операции — Эль-Гамаль)
  - полностью гомоморфное (ограничение по количеству умножений)

- Конфиденциальное машинное обучение
  - применимо в клиент-серверной модели (доверенный пользователь, недоверенный сервер)
  - естественное ограничение на количество операций — применимы полностью гомоморфные схемы
  - требует аппроксимации используемых функций полиномами — ухудшение качества аналитики
  - снижение скорости обучения и обработки данных

Конфиденциальное машинное обучение. Безопасные вычисления

- Протоколы безопасных распределенных вычислений
  - участники получают функцию от данных, но не разглашают самих данных
  - требуют большого количества пересылок



- Конфиденциальное машинное обучение
  - удобно для совместной аналитики нескольких операторов персональных данных
  - требует аппроксимации используемых функций полиномами — ухудшение качества аналитики
  - увеличение объема передаваемых данных и снижение производительности

Что в мире

- Исследования атак на машинное обучение
- Исследования методов защиты
- Внедрение методов конфиденциального машинного обучения

CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy

Nathan Dowlin[1] NDOWLIN@PRINCETON.EDU
Department of Mathematics, Princeton University
Ran Gilad-Bachrach RANG@MICROSOFT.COM
Kim Laine KIM.LAINE@MICROSOFT.COM
Kristin Lauter KLAUTER@MICROSOFT.COM

2017 IEEE Symposium on Security and Privacy

SecureML: A System for Scalable Privacy-Preserving Machine Learning

Payman Mohassel* and Yupeng Zhang†
*Visa Research, †University of Maryland

RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response

Úlfar Erlingsson
Google, Inc.
ulfar@google.com

Vasyl Pihur
Google, Inc.
vpihur@google.com

Aleksandra Korolova
University of Southern California
korolova@usc.edu

У нас

- На сегодняшний момент нет системного подхода к решению проблемы
- Есть очень удачные примеры выявления и противодействия некоторым типам атак
- Федеральный проект «Искусственный интеллект»:
  - Формирование научной базы для современных защищенных технологий искусственного интеллекта, применяемых в государственных информационных системах (Академия криптографии Российской Федерации, до 2024 г.)
  - Исследования общих принципов создания и анализа безопасности систем искусственного интеллекта, систем конфиденциального машинного обучения, систем обезличивания и тд.

Благодарю за внимание!