Разработка требований к системам доверенного искусственного интеллекта

Маршалко Г.Б.

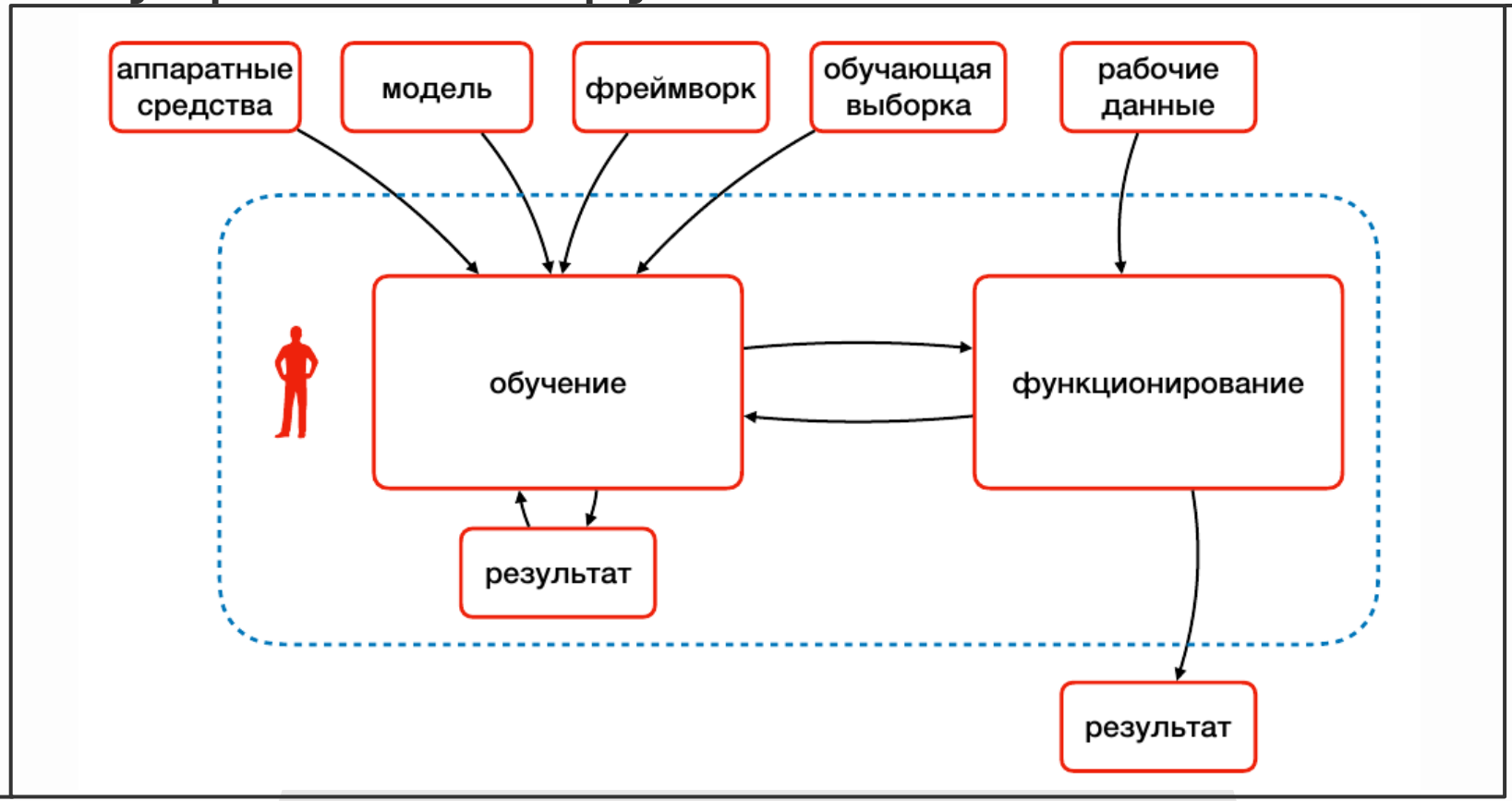
Объект защиты

Нейросетевые алгоритмы машинного обучения, предназначенные для решения одной или нескольких статистических задач (средство ИИ)

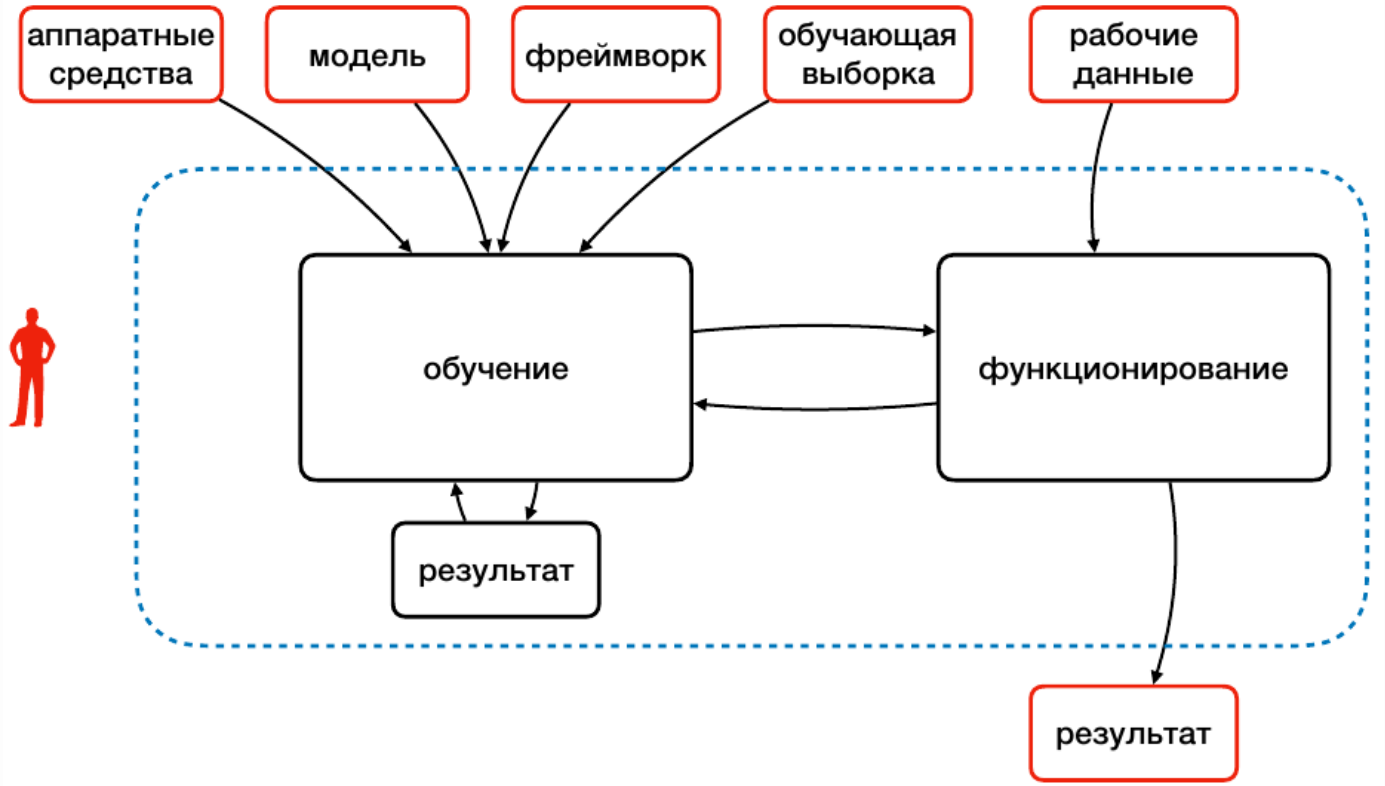
Алгоритм решения задачи формируется одновременно с получением результата ее решения (обучение ИИ) – невозможность априорного изучения алгоритма

Возможность реализации «логических» атак через содержание обрабатываемых данных (существующие методы защиты – безопасность формы представления данных)

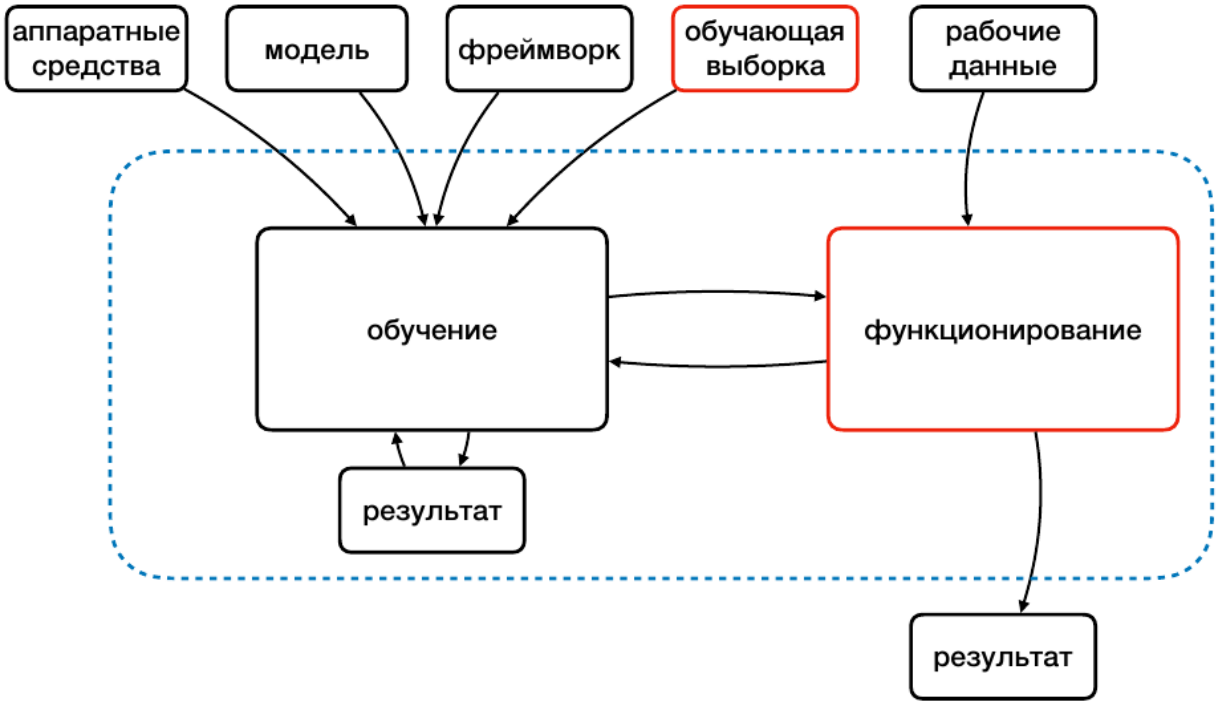
Внутренний нарушитель



Внешний нарушитель



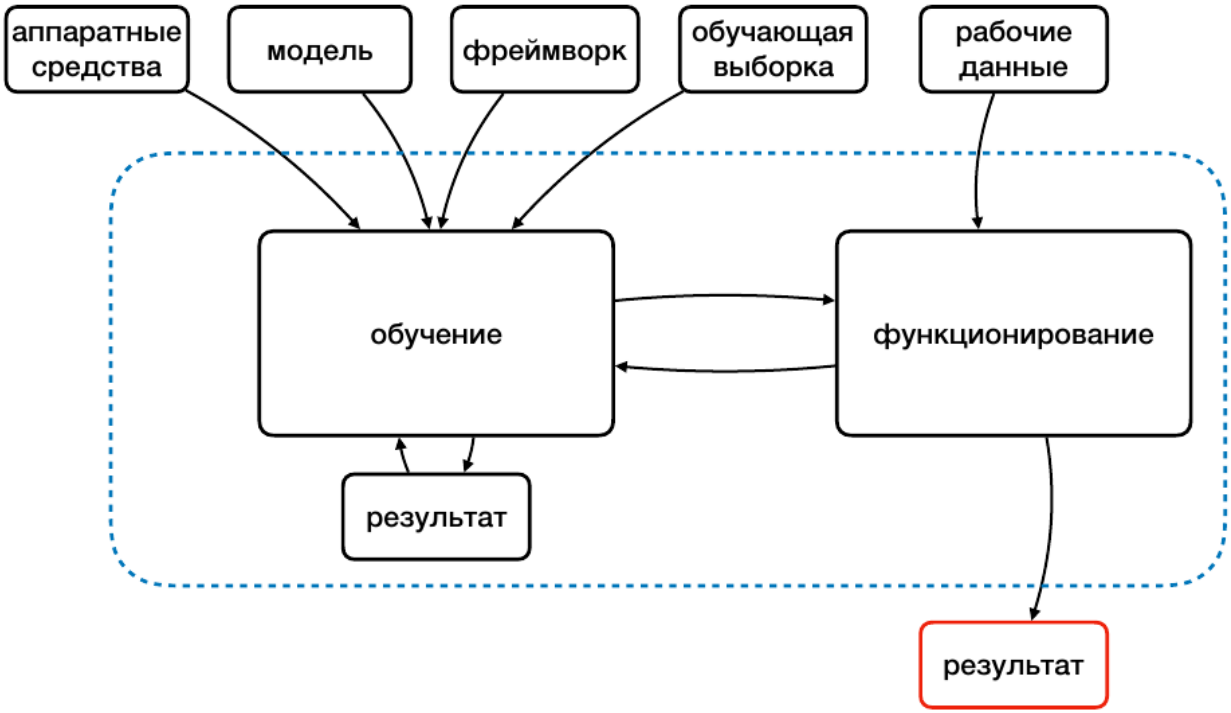
Цели нарушителя



Нарушение свойства конфиденциальности данных:

- о параметрах обученной модели
- об обучающей выборке

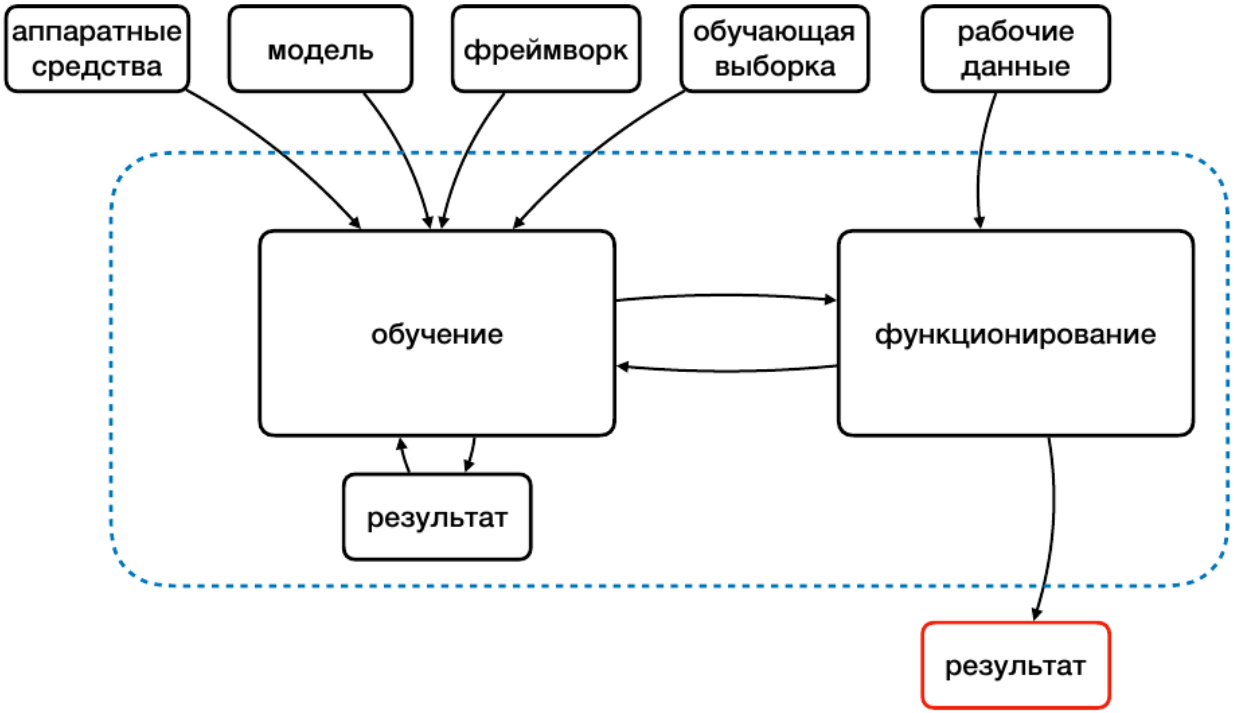
Цели нарушителя



Нарушение свойства доступности модели:

- ухудшение качества модели

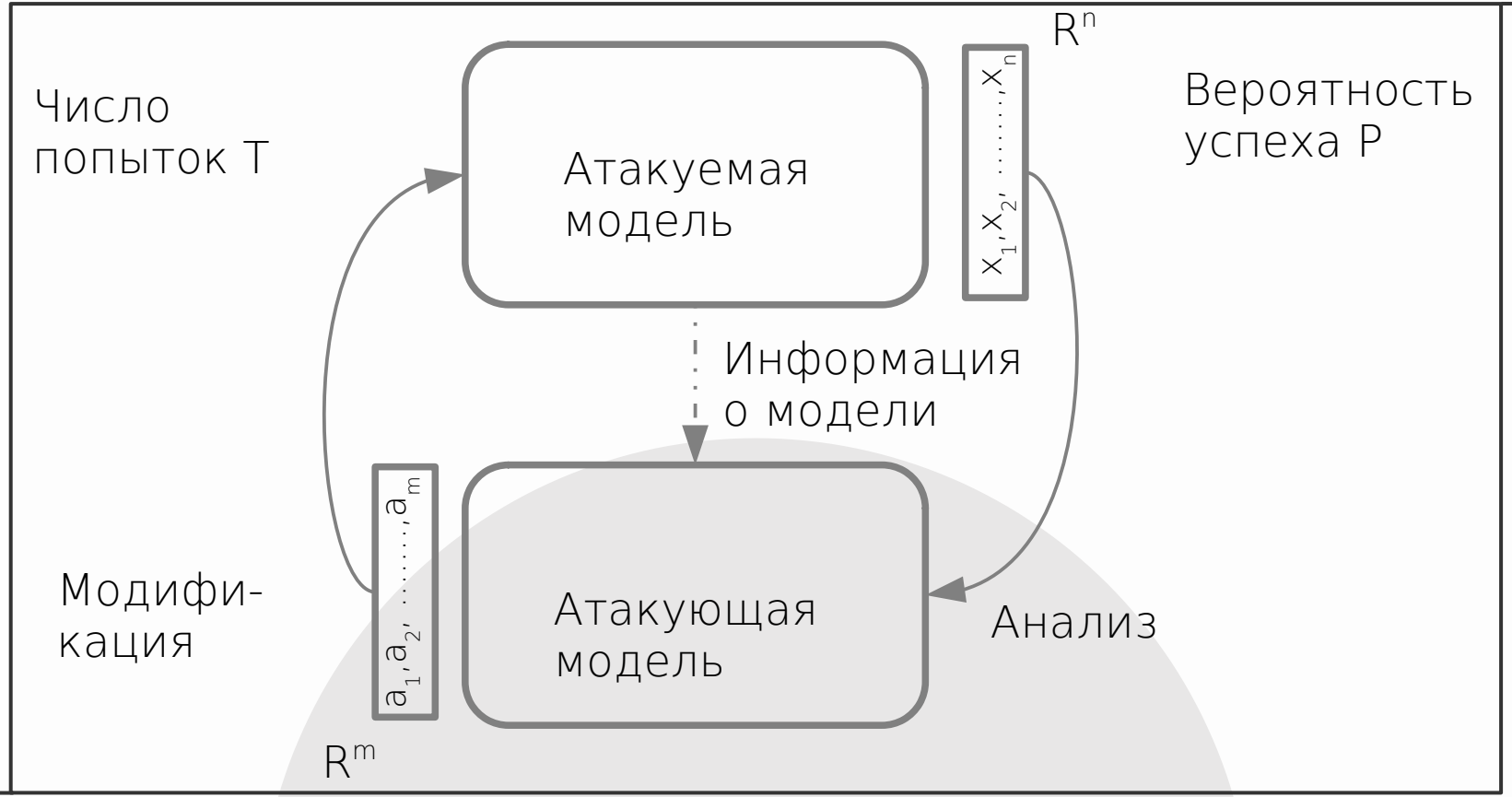
Цели нарушителя



Нарушение свойства целостности модели:

- целенаправленное изменение результатов работы

Модель атаки



Контроль доступа

Ограничивая число попыток доступа нарушителя можно уменьшить вероятность успешной реализации атаки

Ограничивая доступ нарушителя к выходному вектору модели можно увеличивать требуемое количество попыток

Ограничивая доступ нарушителя к выходному вектору модели можно уменьшить вероятность успешной реализации атаки

Доступная информация

Нет информации о модели, обучающем множестве

Есть информация о гиперпараметрах модели, общая информация об обучающем множестве

Модель известна, известны обучающие множества

Контроль доступа

Ограничивая доступ к модели (весам модели) можно:

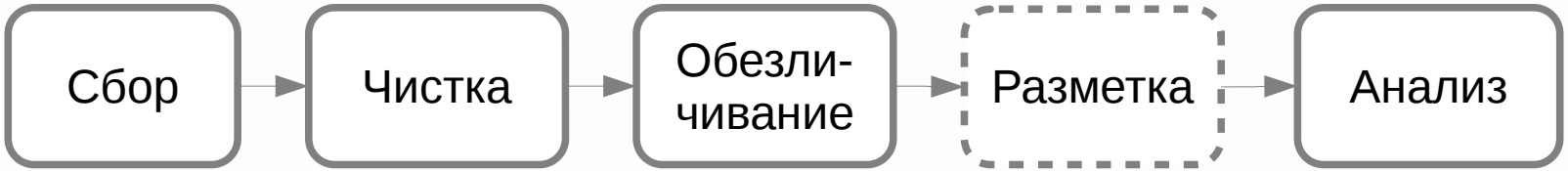
Увеличить необходимое число попыток для реализации атаки

Уменьшить вероятность успешной реализации атаки

Что влияет на оценку безопасности?

Цель исследования	Решаемая статистическая задачи или их набор
Требуемые значения метрик качества	Ошибки первого/второго рода, полнота, точность, AUC....
Эксплуатационные метрики	Память, скорость работы, разрядность...
Выбор аппаратной платформы	НДВ в нейропроцессорах, графических ускорителях
Выбор фреймворка машинного обучения	НДВ в ПО
Предполагаемая архитектура системы	ETL/ELT, хранилище/озеро/фабрика, использование результата внутри контура/ вне контура ...

Обучающие данные



- Большое число вариантов получения данных
- Длительный, многоэтапный процесс обработки
- Значительная роль операторов в подготовке данных

Управление данными на всем пути обработки:

- Контроль и разграничение доступа
- Обезличивание и конфиденциальная обработка
- Регистрация версий наборов и событий обработки
- Повышение качества данных:
  - баланс классов
  - «отравленные» и состязательные примеры
  - объединение наборов

Обучение модели

01

Выбор модели

Верифицированные
Интерпретируемые

02

Качество

Минимизация ошибок

03

Робастность

Состязательное обучение
Сенсоры атак

04

Защита от инверсии

Статистически обезличенный градиентный спуск

05

Контроль

Доступа операторов
Логирование
Контроль целостности параметров

06

Устойчивость к атакам

С учетом архитектуры системы

Функционирование модели

01

Контроль доступа

Оператора
Пользователей

02

Контроль входа

Среда функционирования
Качество данных

03

Контроль выхода

Ограничение доступа к выходному вектору

04

Логирование

Промежуточных результатов
Событий безопасности

05

Контроль качества

Дрейф данных

Спасибо за внимание!