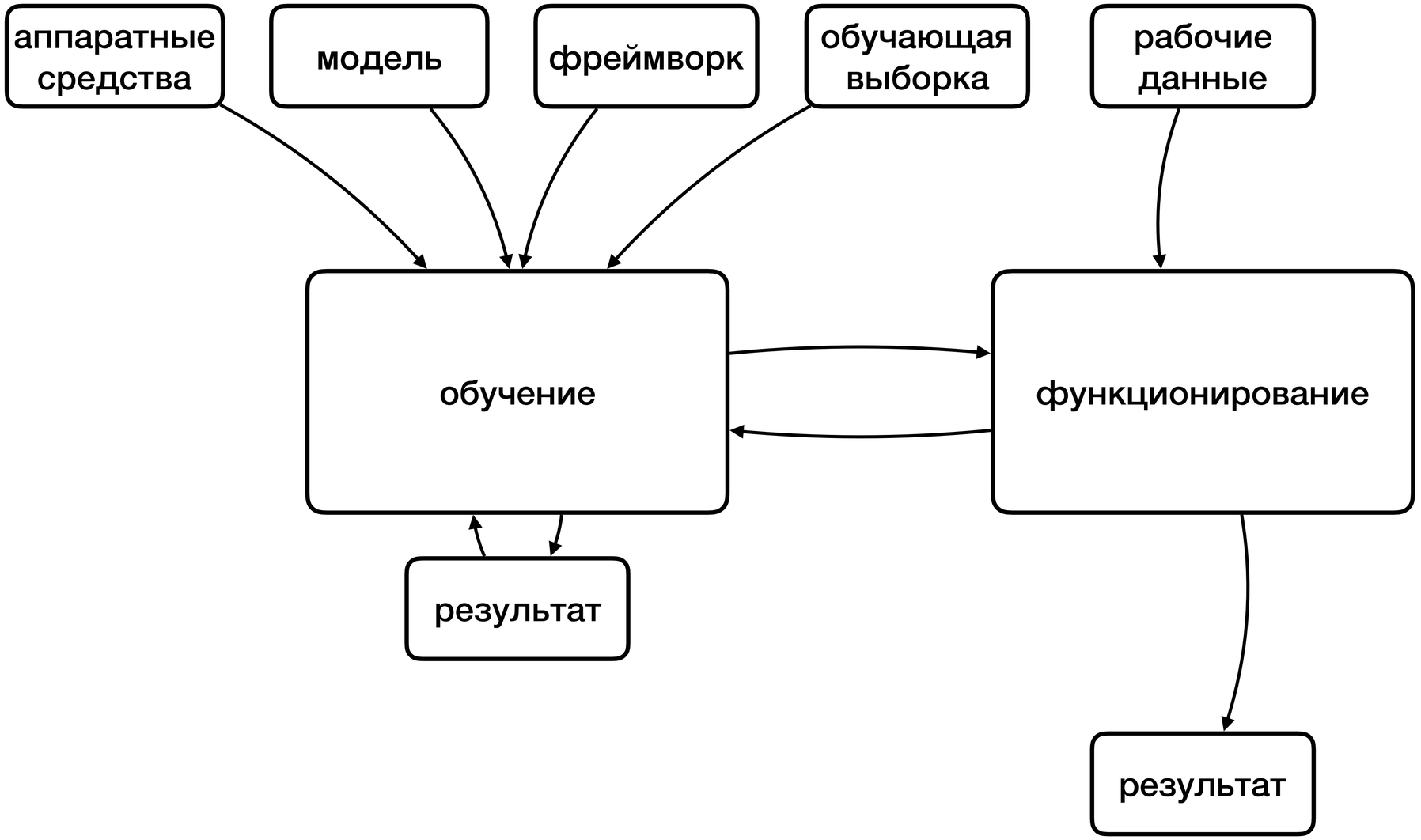
К вопросу построения требований к системам доверенного искусственного интеллекта

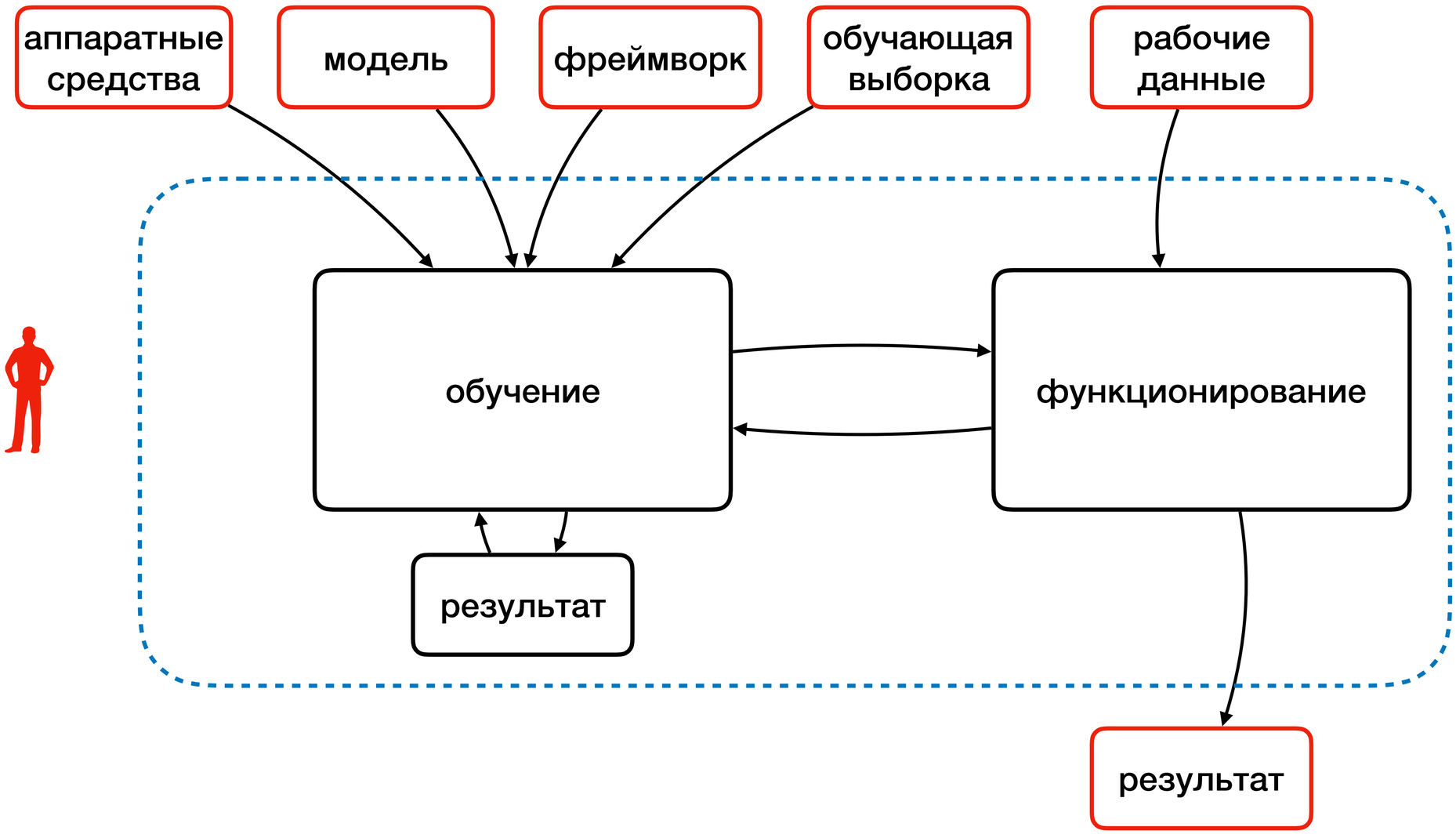
Система ИИ


аппаратные средства
модель
фреймворк
обучающая выборка
рабочие данные
обучение
функционирование
результат
результат

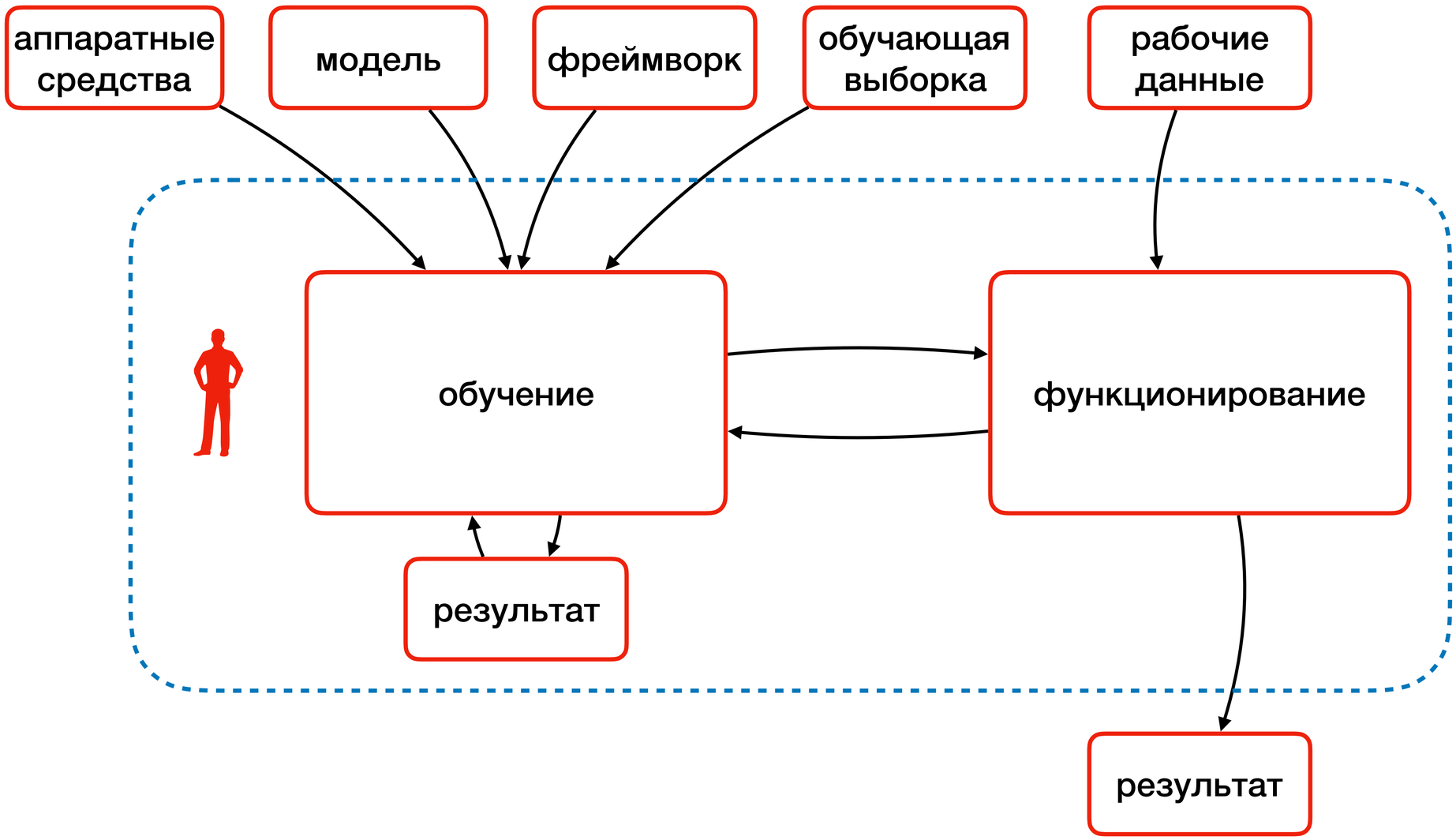
Проблема

- Доверие — степень уверенности пользователя или другой заинтересованной стороны в том, что продукт или система ведет себя так, как задумано (ISO/IEC 25010:2011)
- Классические информационные системы — безопасная разработка, верификация, аудит сертификация и тп.
- Системы с ИИ — алгоритм решения не является фиксированным и разрабатывается вместе с поиском решения. Высокая зависимость от данных и опыта специалиста по данным/инженера

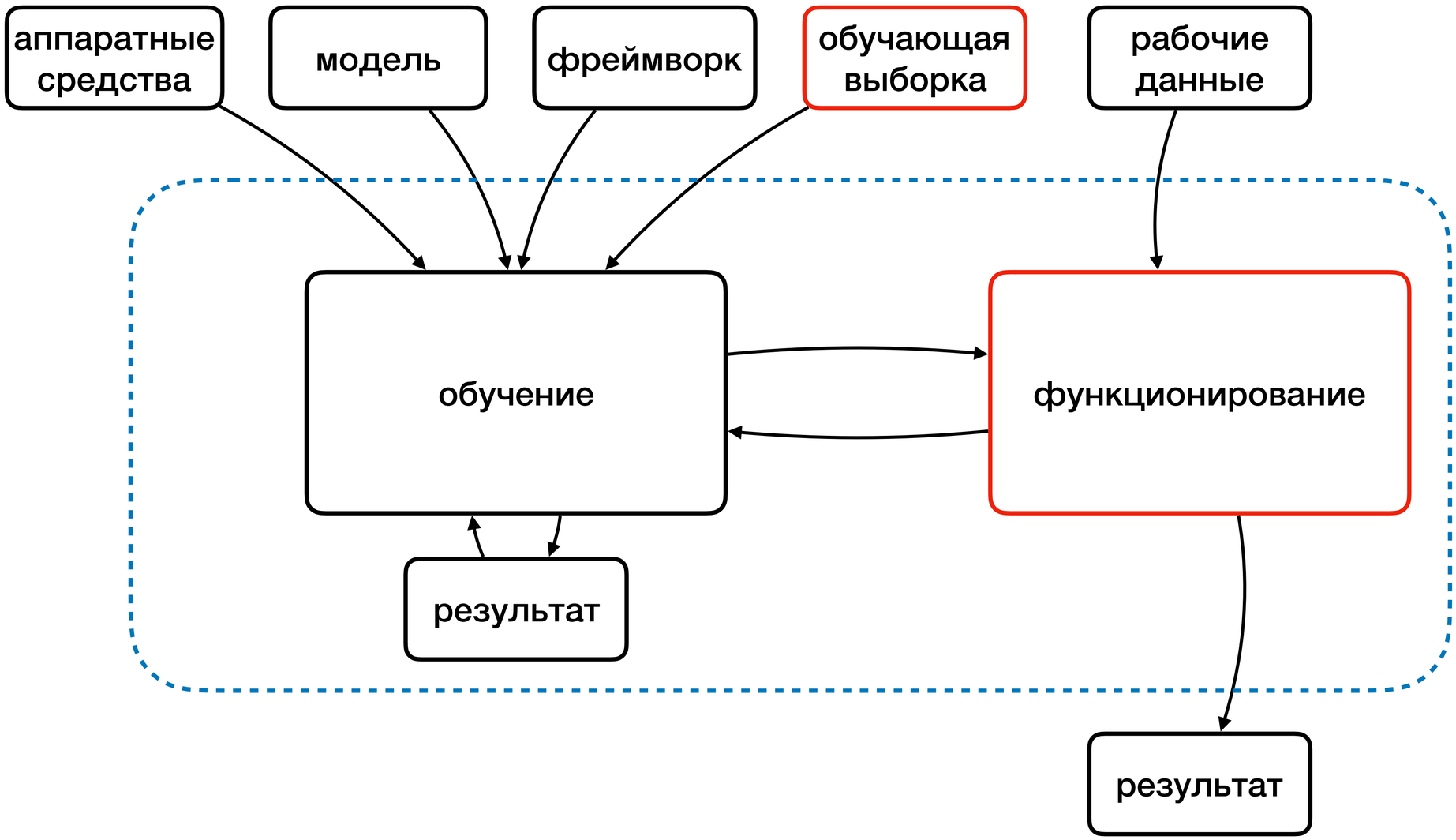
Возможности нарушителя



Возможности нарушителя



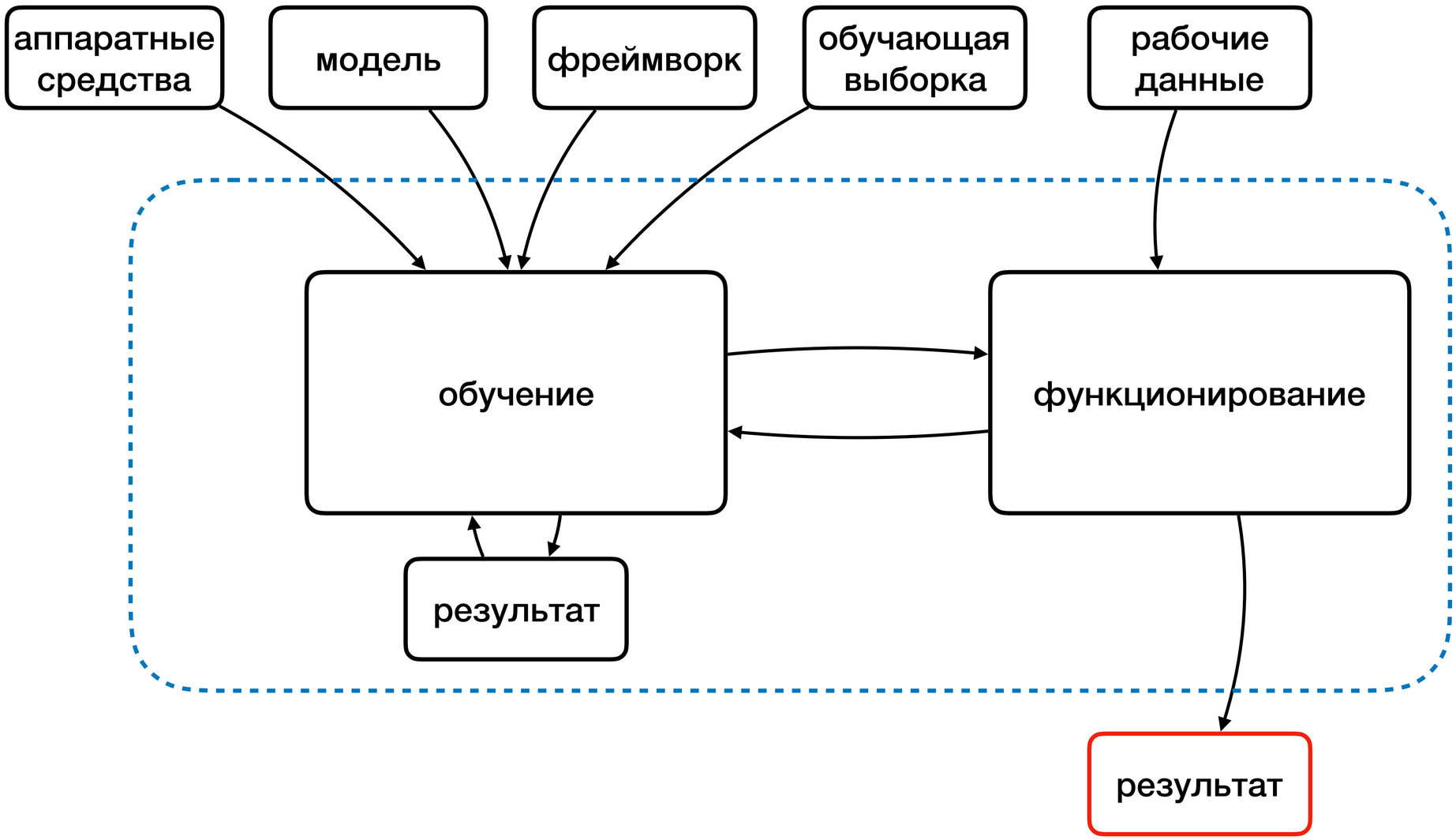
Цели нарушителя



Нарушение свойства конфиденциальности данных:

- о параметрах обученной модели
- об обучающей выборке

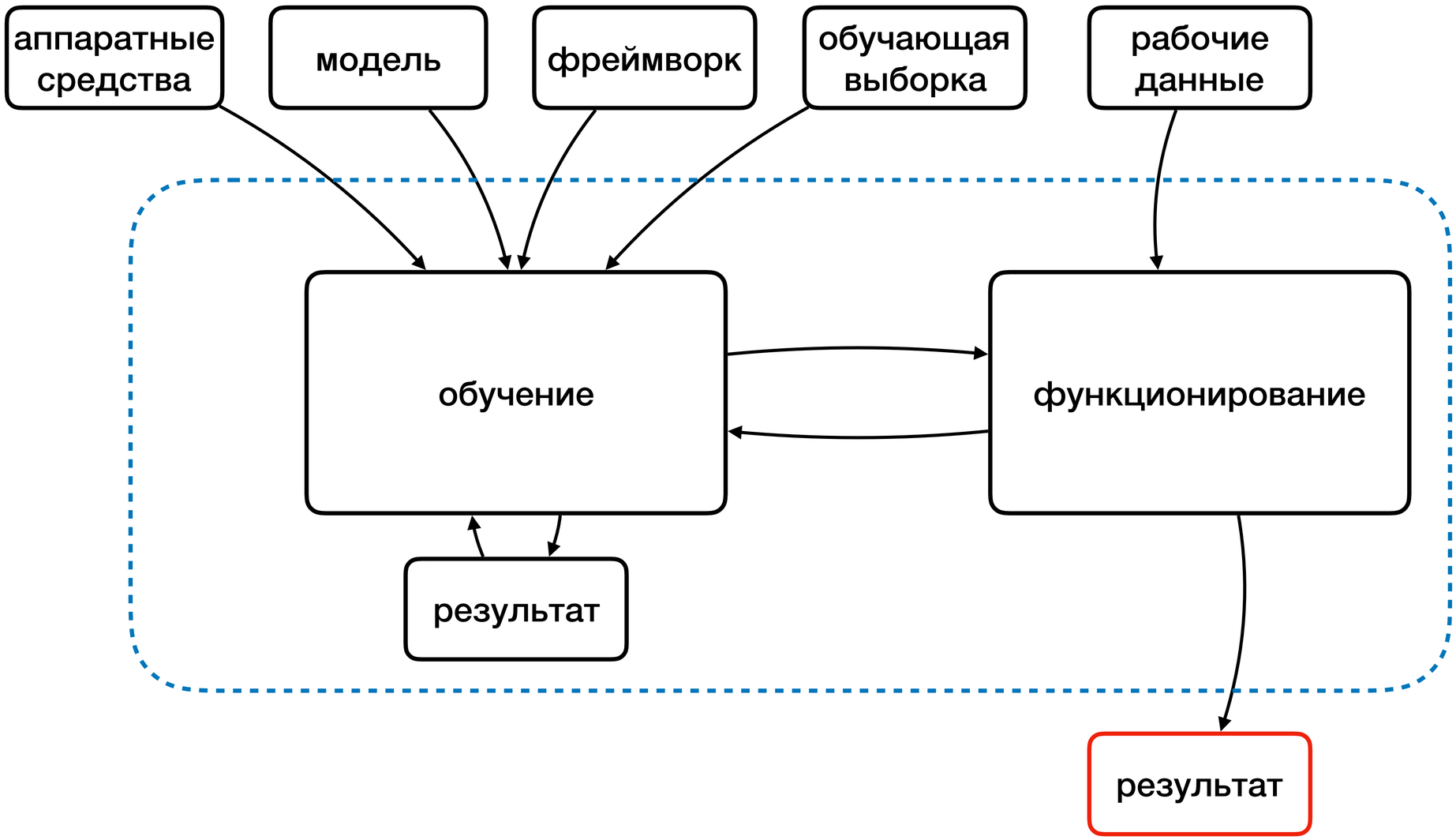
Цели нарушителя



Нарушение свойства доступности модели:

- ухудшение качества модели

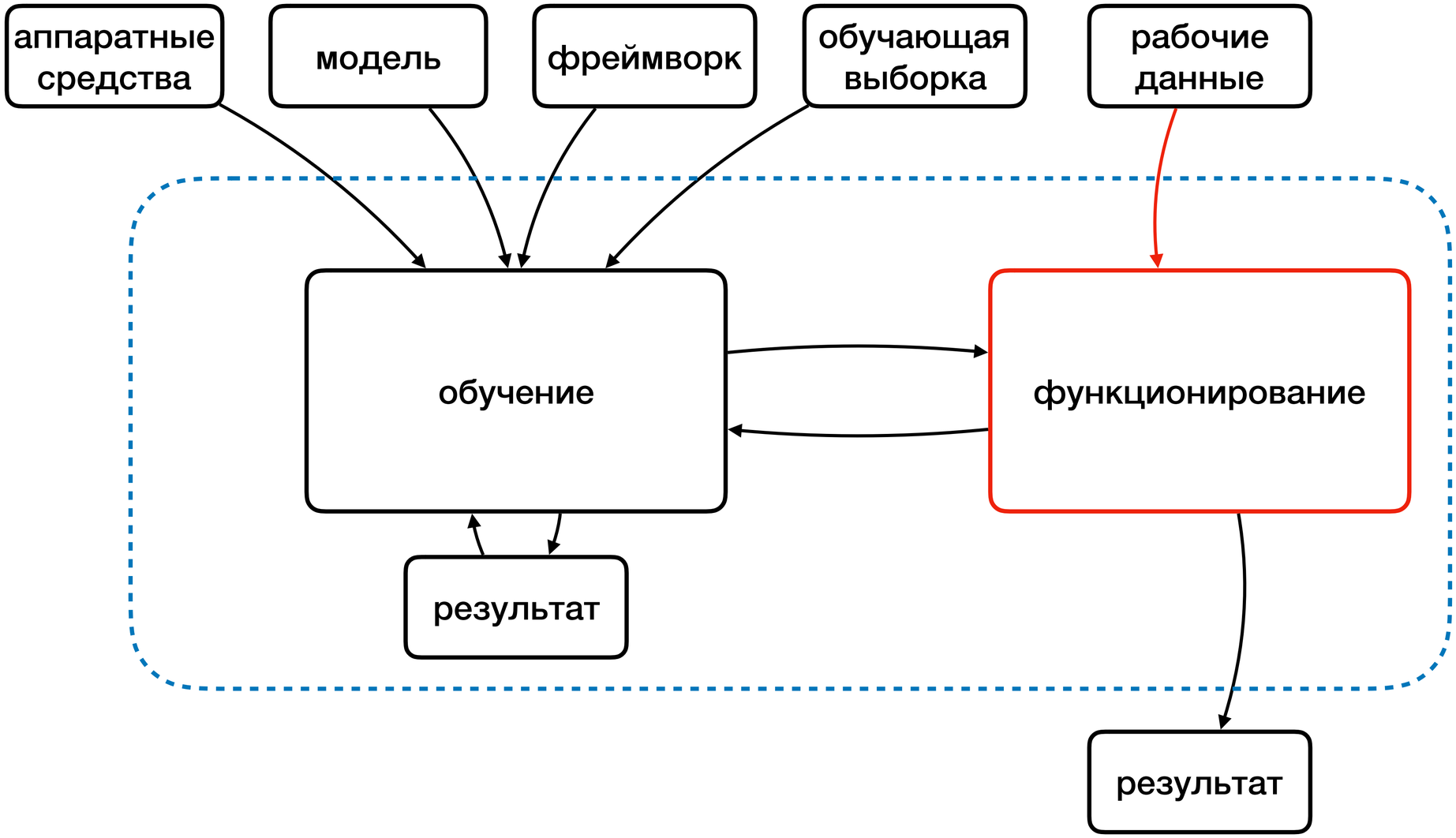
Цели нарушителя



Нарушение свойства целостности модели:

- целенаправленное изменение результатов работы

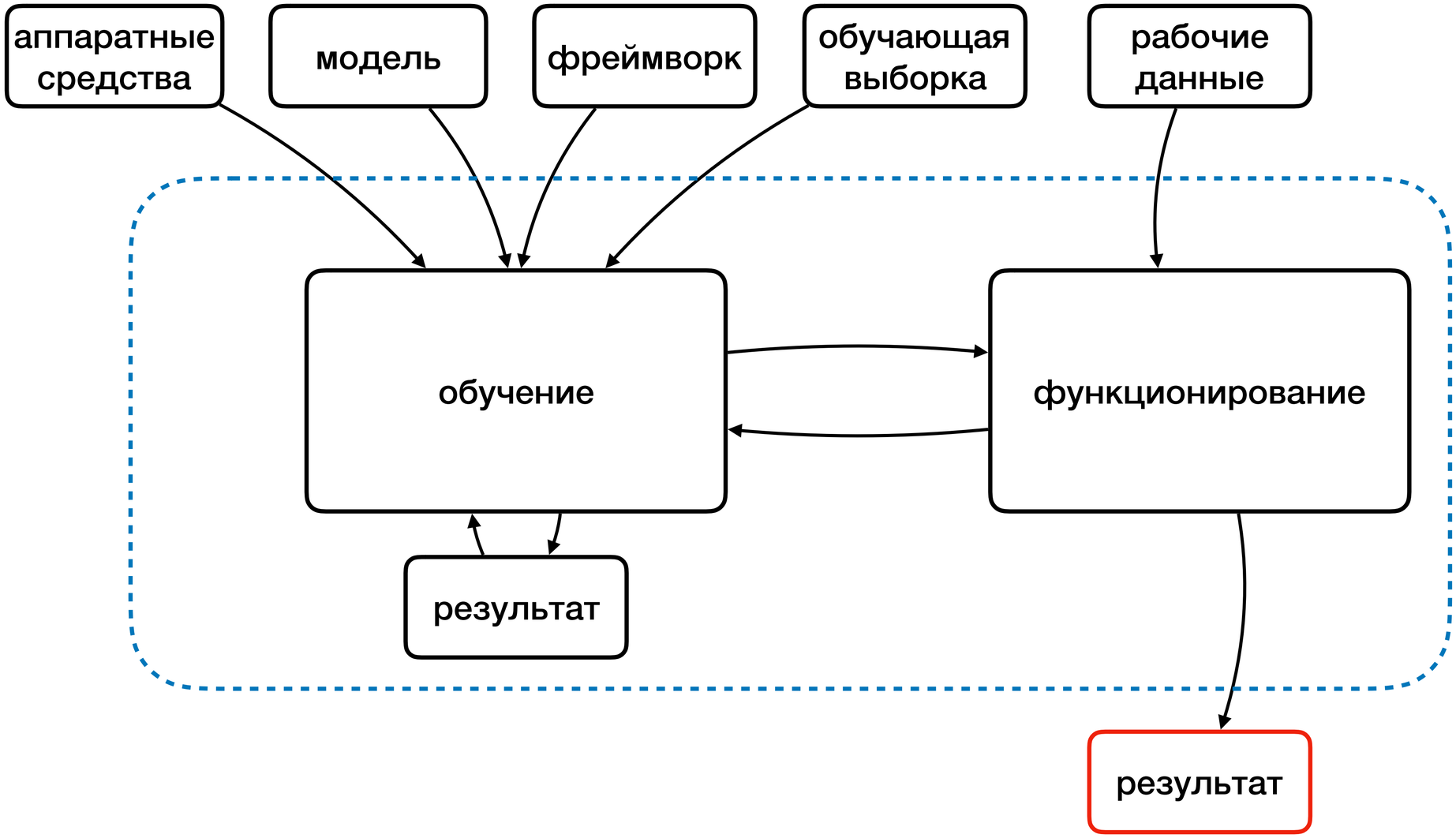
Подходы к защите



Организационно-технические (контроль доступа)

- к модели
- к результату работы
- к попыткам ввода данных
- контроль происхождения и качества данных

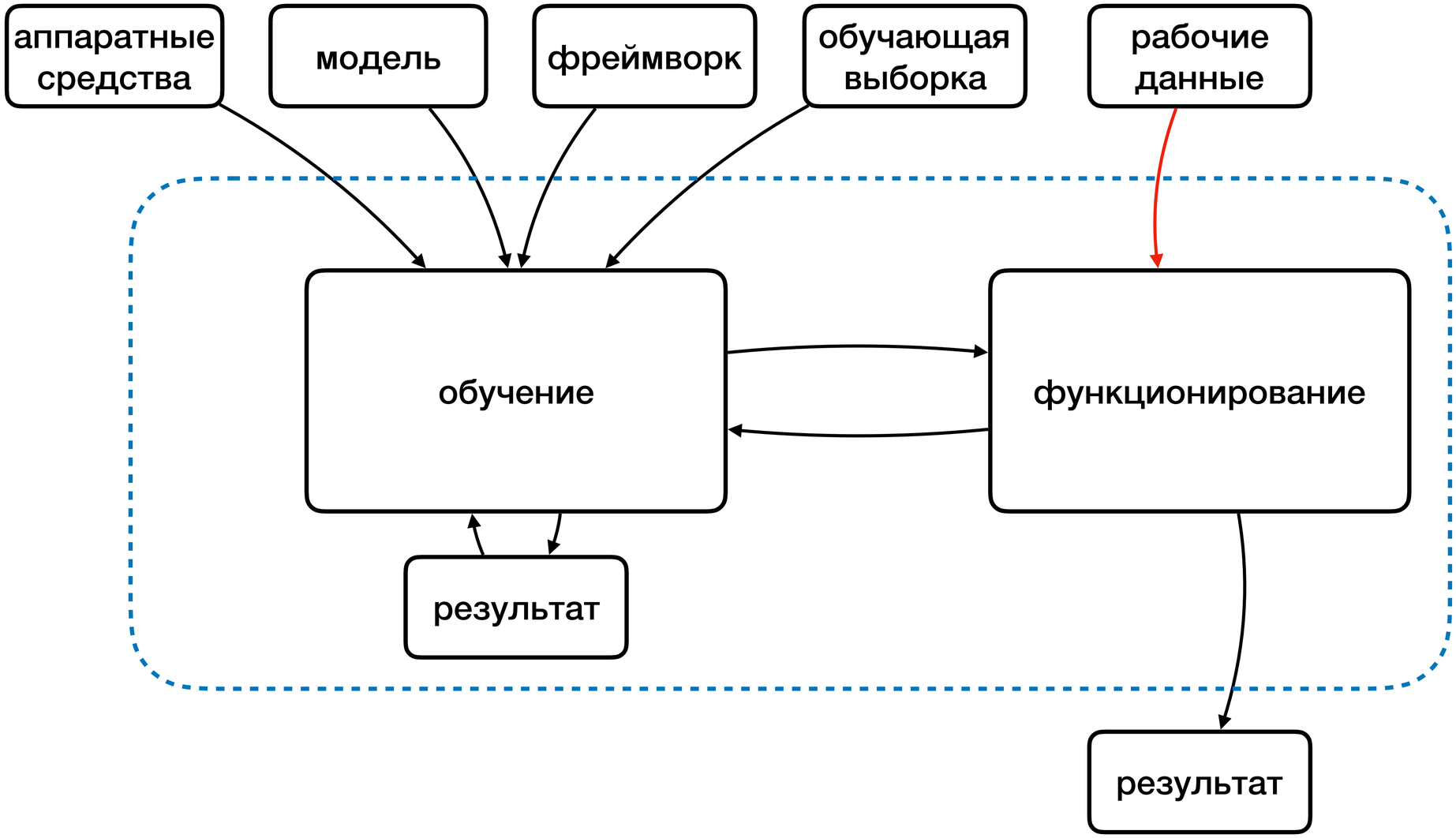
Подходы к защите



Организационно-технические (контроль доступа)

- к модели
- к результату работы
- к попыткам ввода данных
- контроль происхождения и качества данных

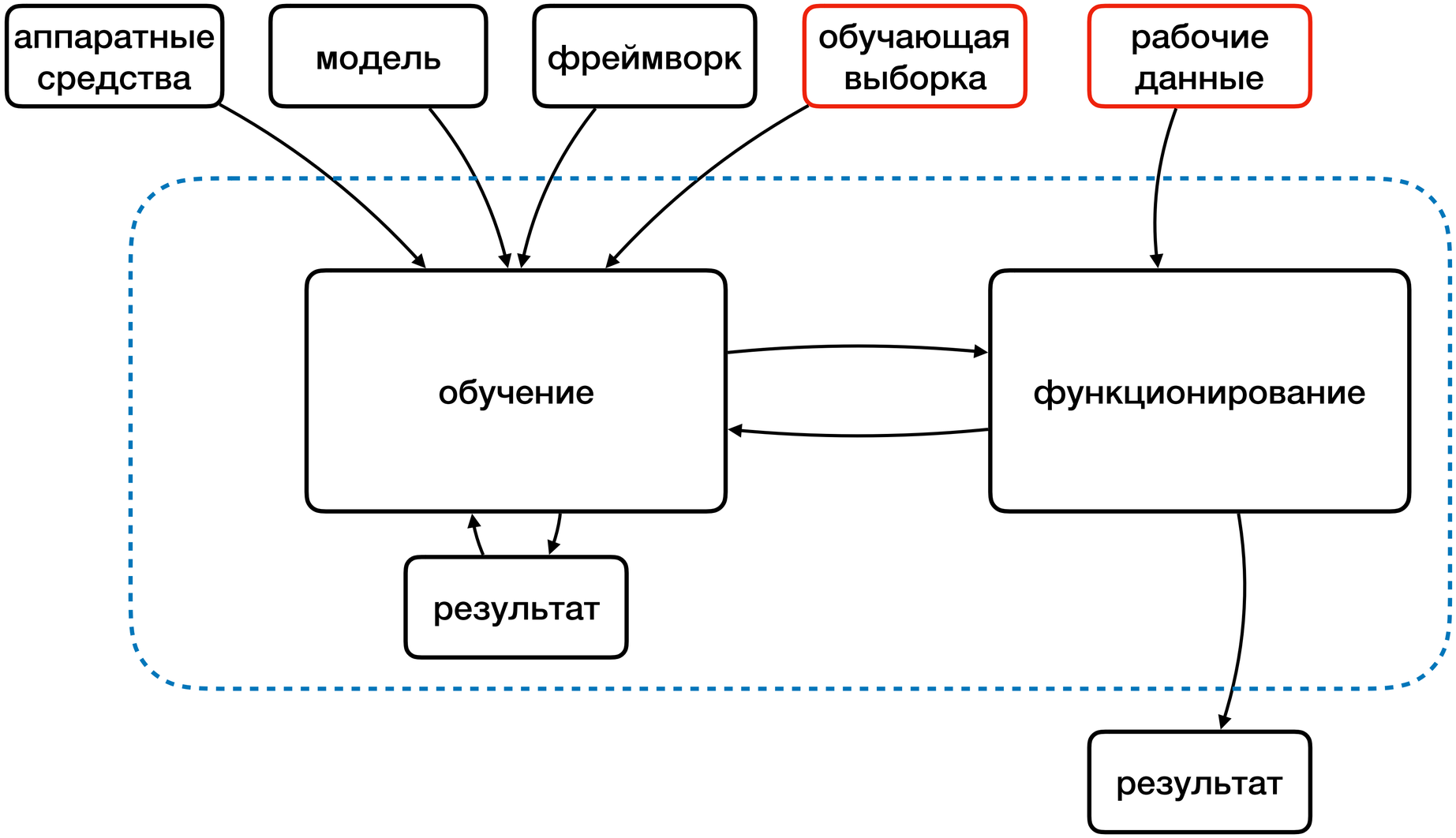
Подходы к защите



Организационно-технические (контроль доступа)

- к модели
- к результату работы
- к попыткам ввода данных
- контроль происхождения и качества данных

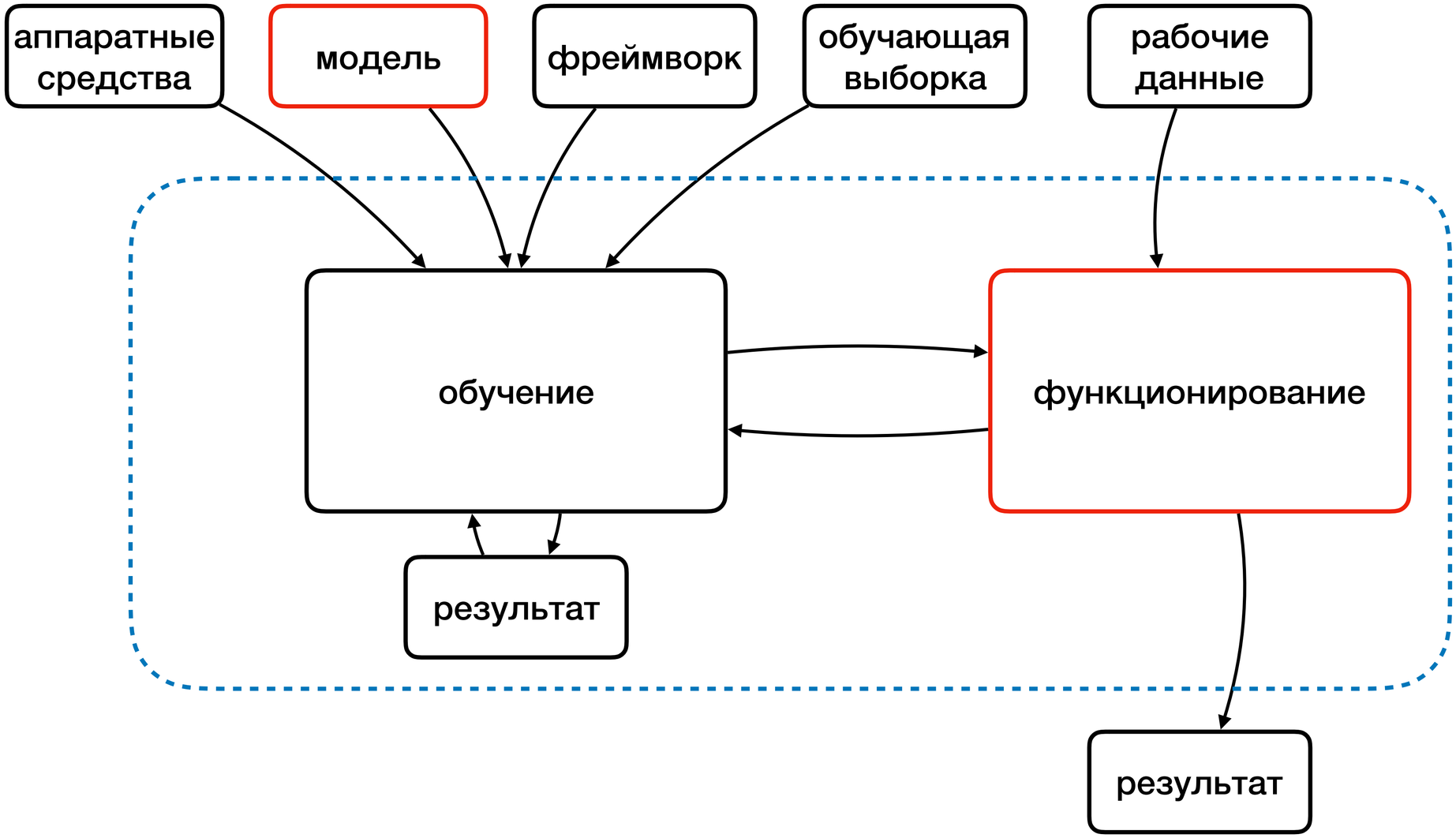
Подходы к защите



Организационно-технические (контроль доступа)

- к модели
- к результату работы
- к попыткам ввода данных
- контроль происхождения и качества данных

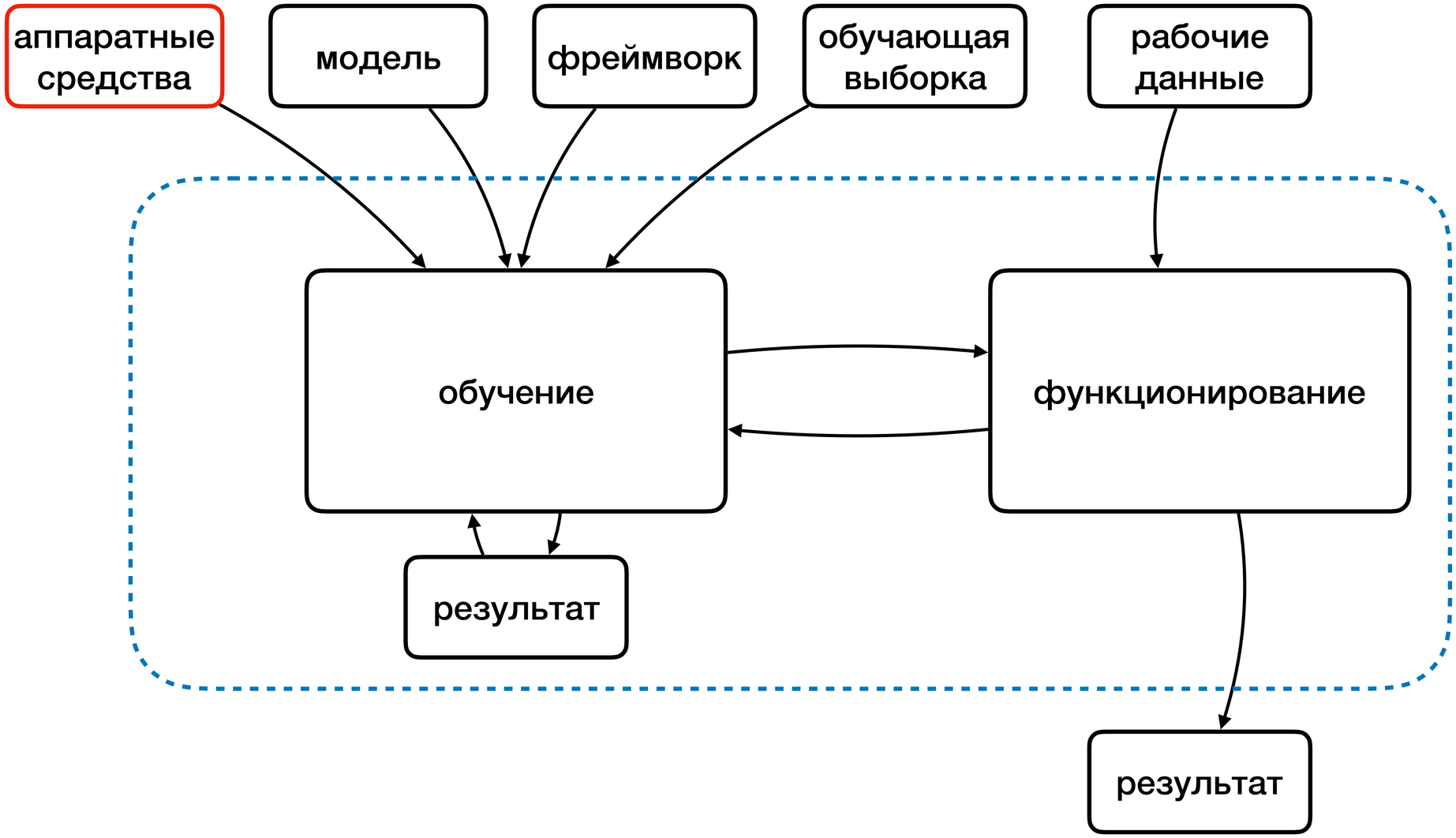
Подходы к защите



Аудит

- модели
- аппаратных средств
- фреймворка
- интерпретация моделей

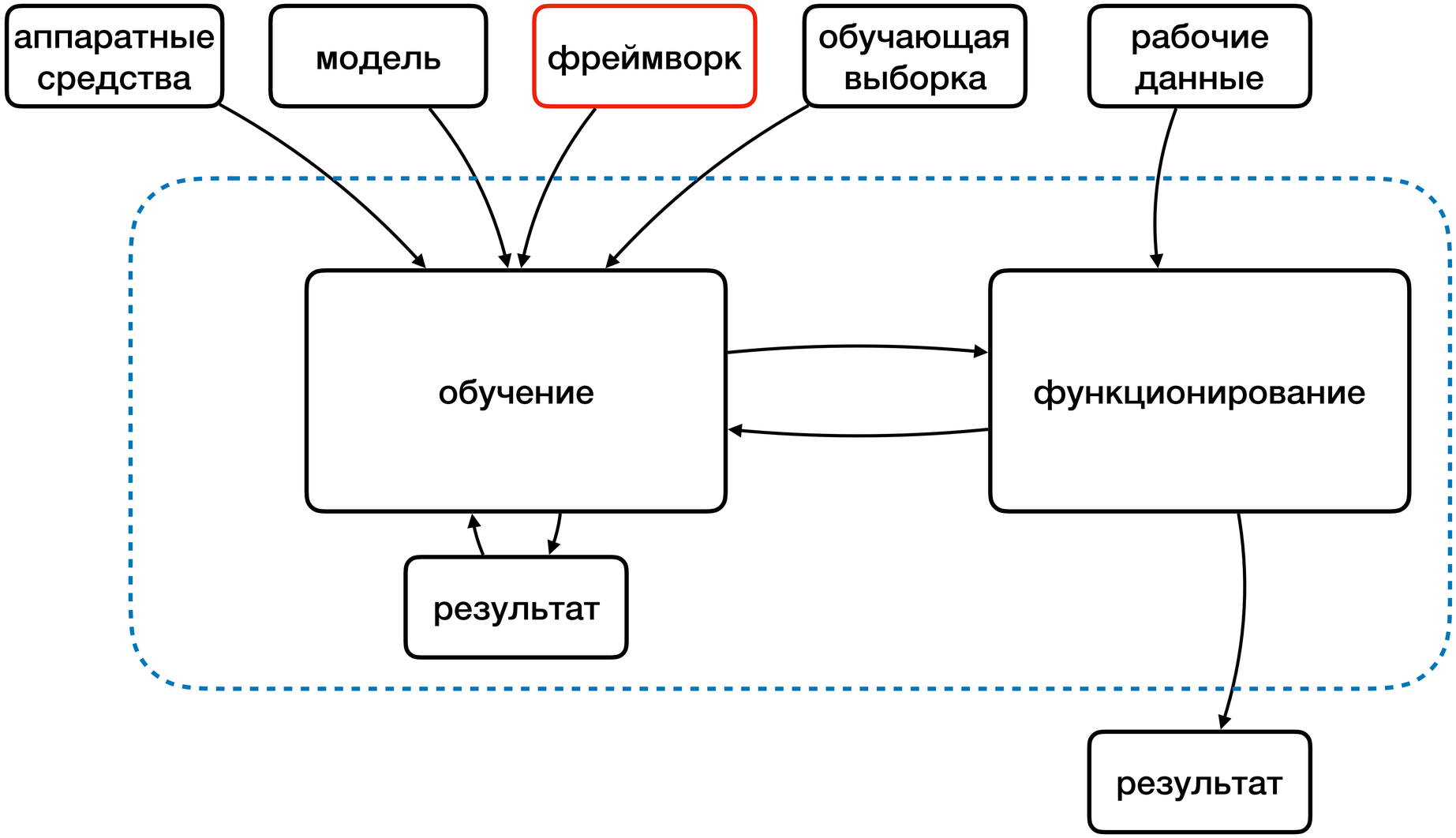
Подходы к защите



Аудит

- модели
- аппаратных средств
- фреймворка
- интерпретация моделей

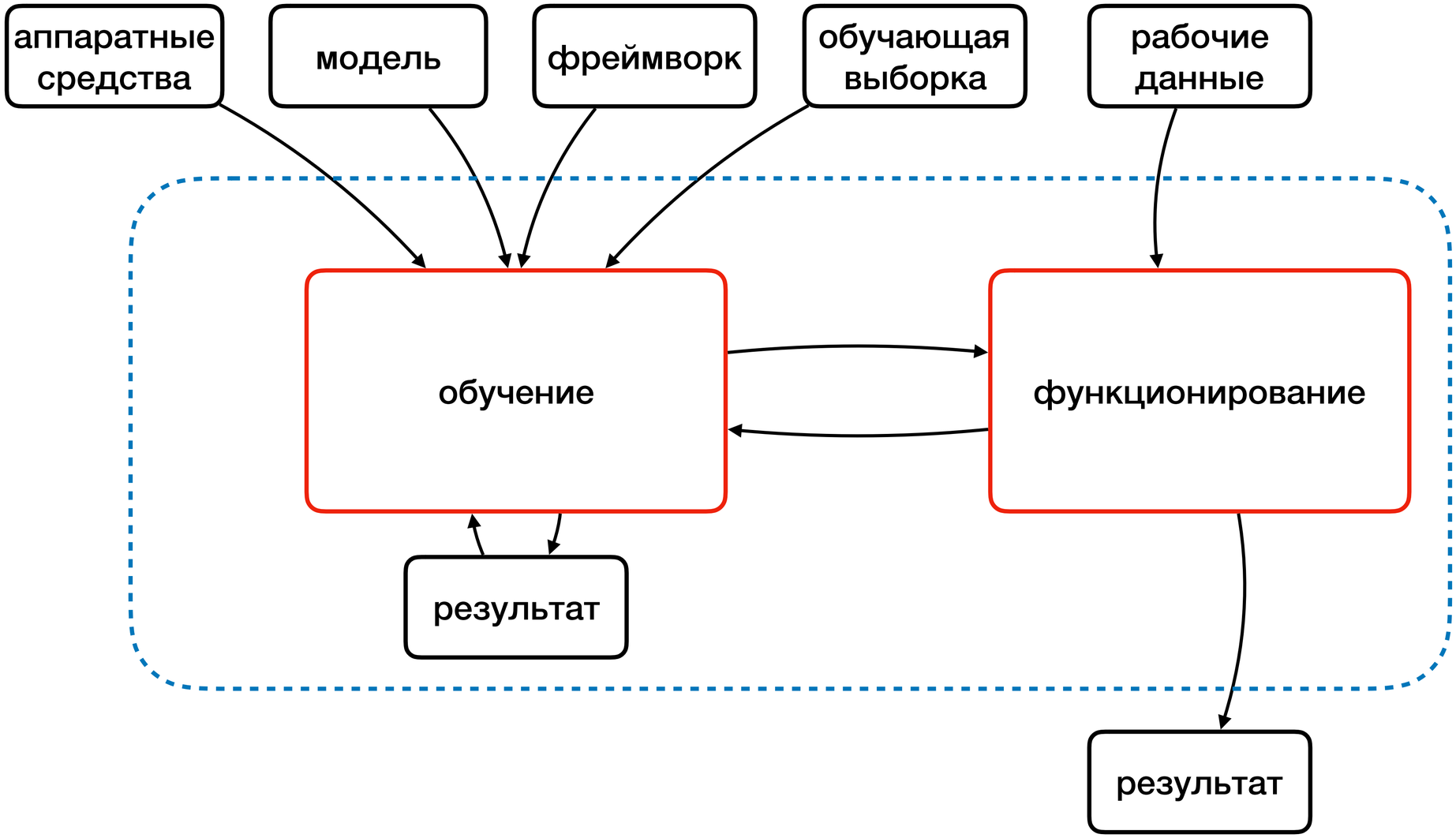
Подходы к защите



Аудит

- модели
- аппаратных средств
- фреймворка
- интерпретация моделей

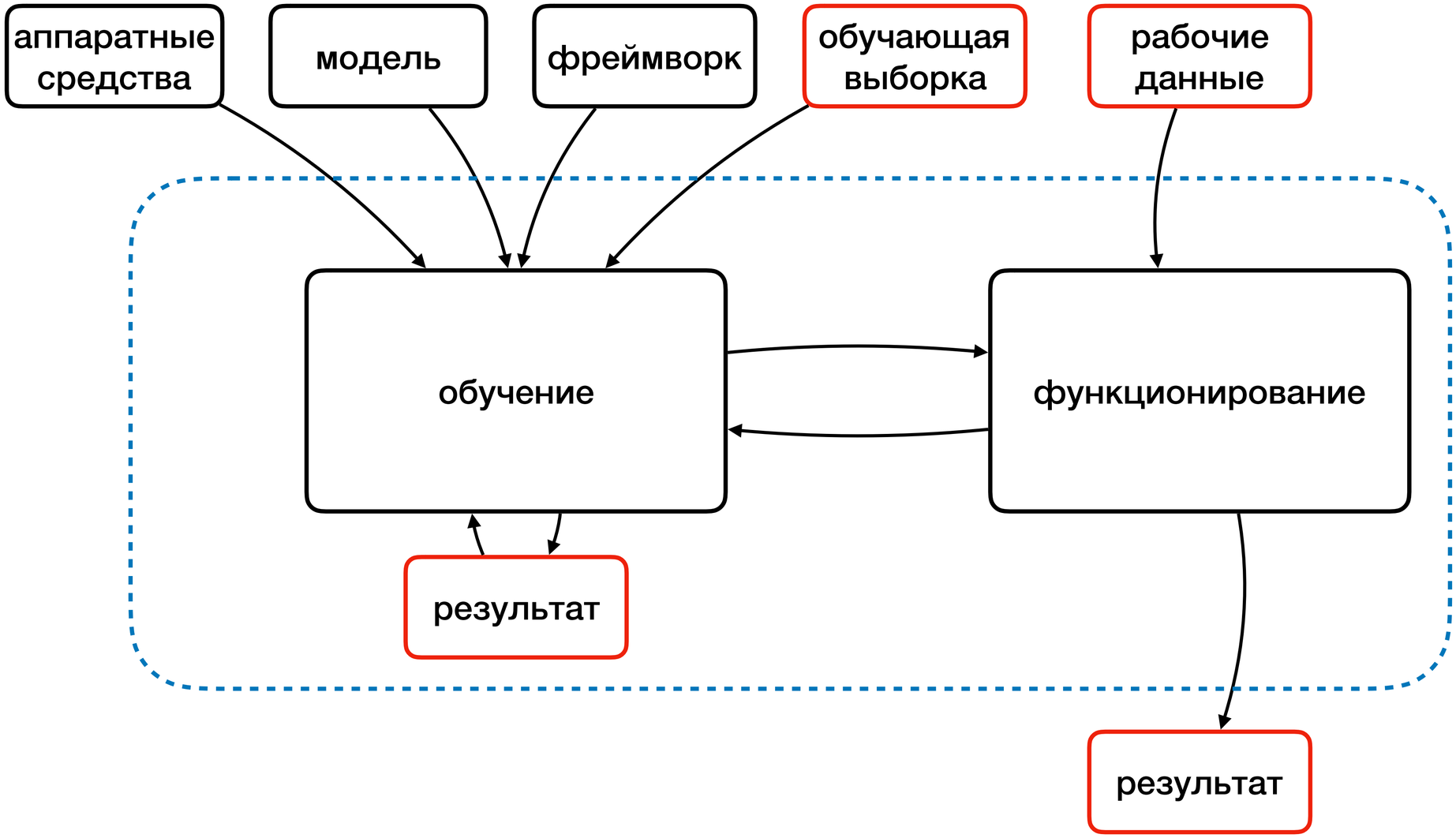
Подходы к защите



Аудит

- модели
- аппаратных средств
- фреймворка
- интерпретация моделей

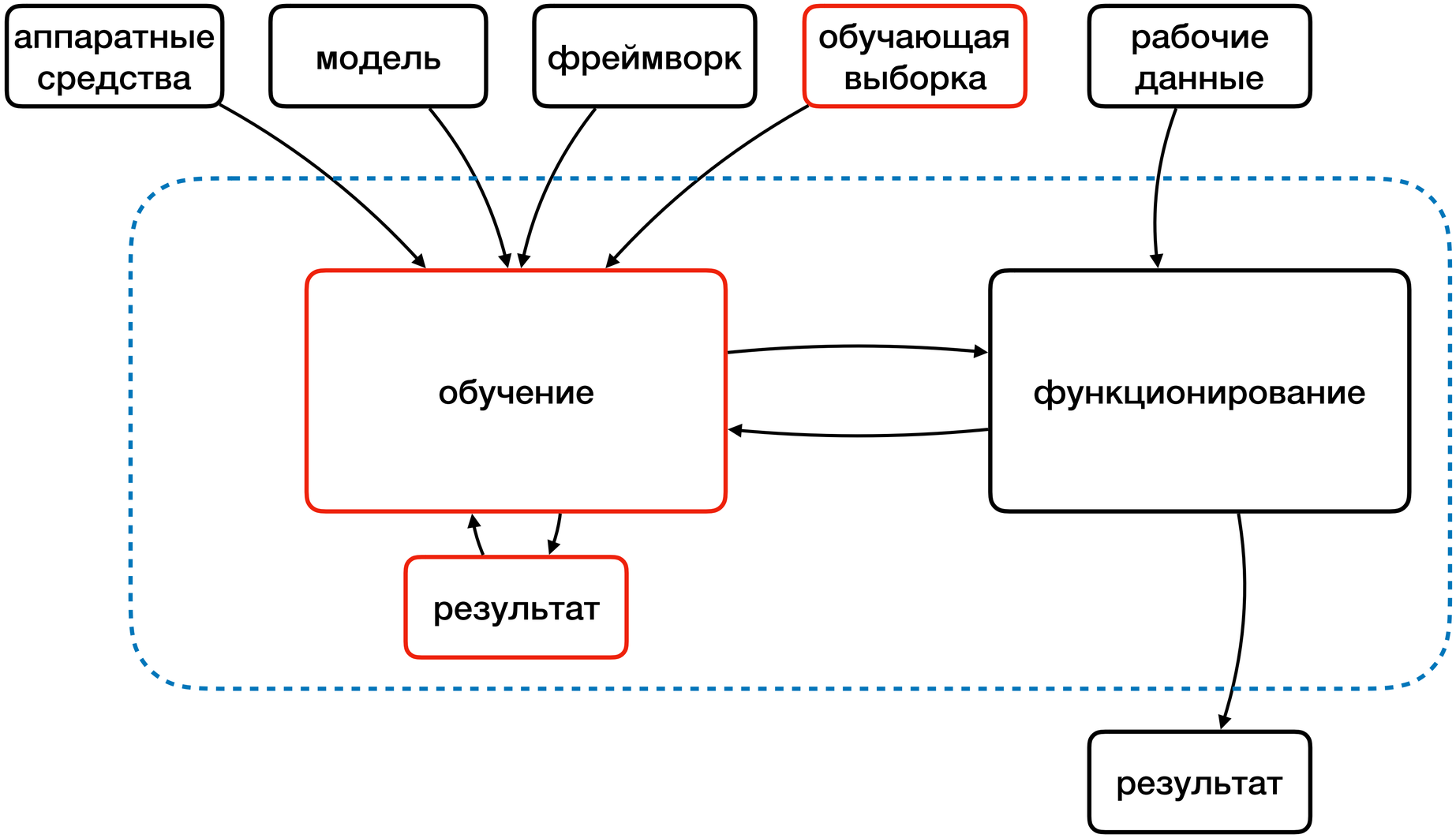
Подходы к защите



Алгоритмические

- обезличивание данных
- федеративное обучение
- устойчивые (робастные) модели

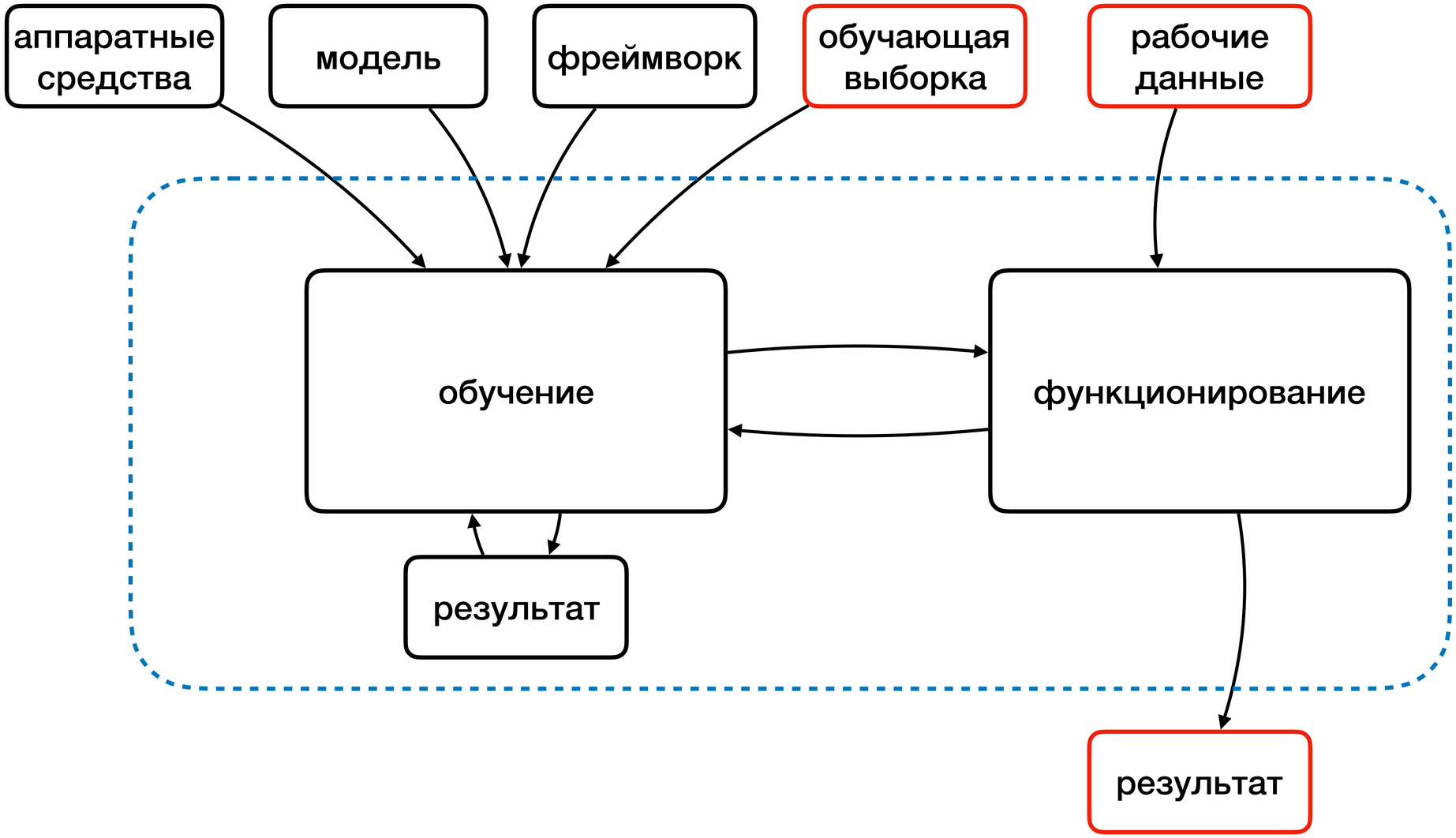
Подходы к защите



Алгоритмические

- обезличивание данных
- федеративное обучение
- устойчивые (робастные) модели

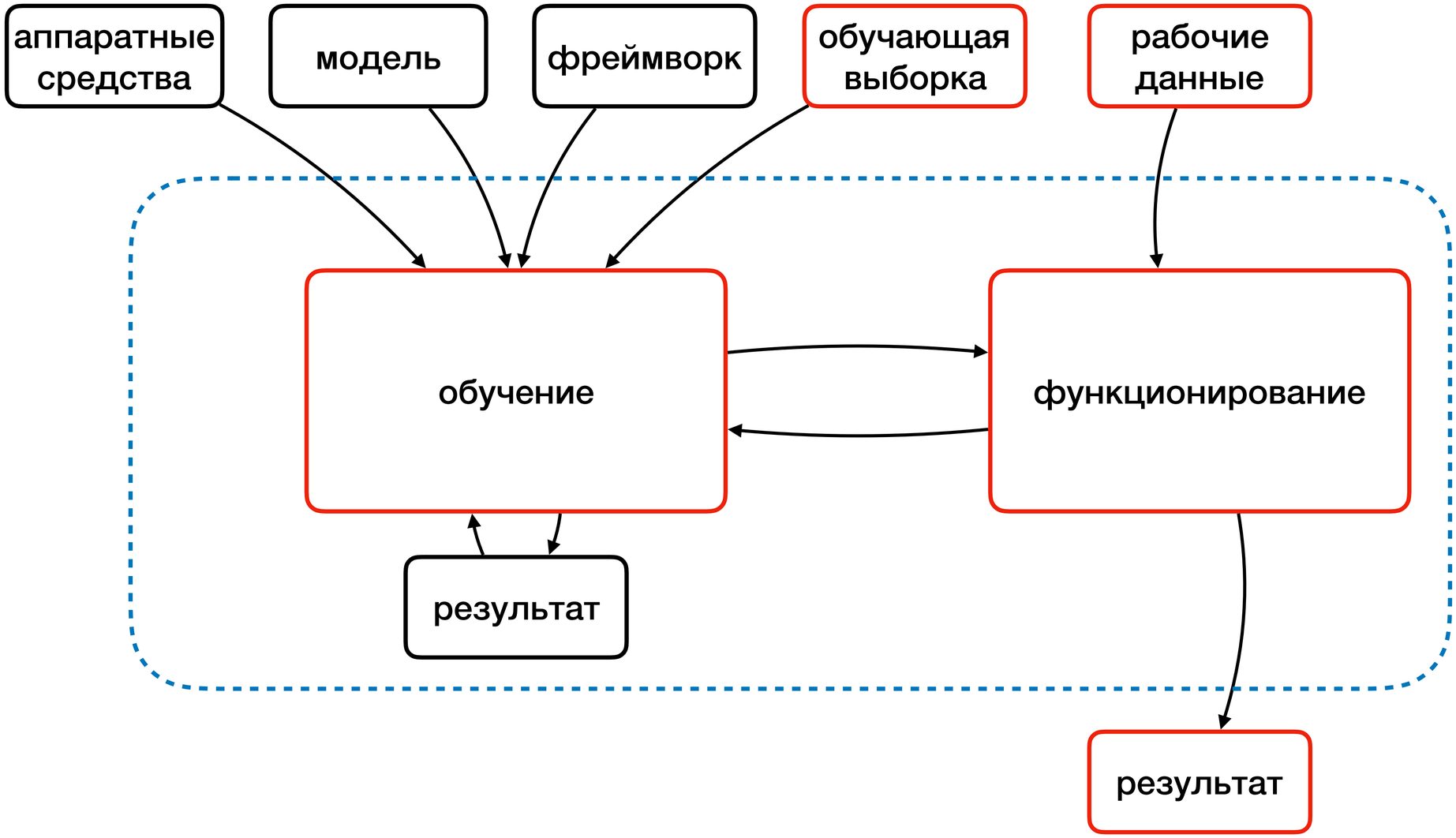
Подходы к защите



Алгоритмические

- обезличивание данных
- федеративное обучение
- устойчивые (робастные) модели

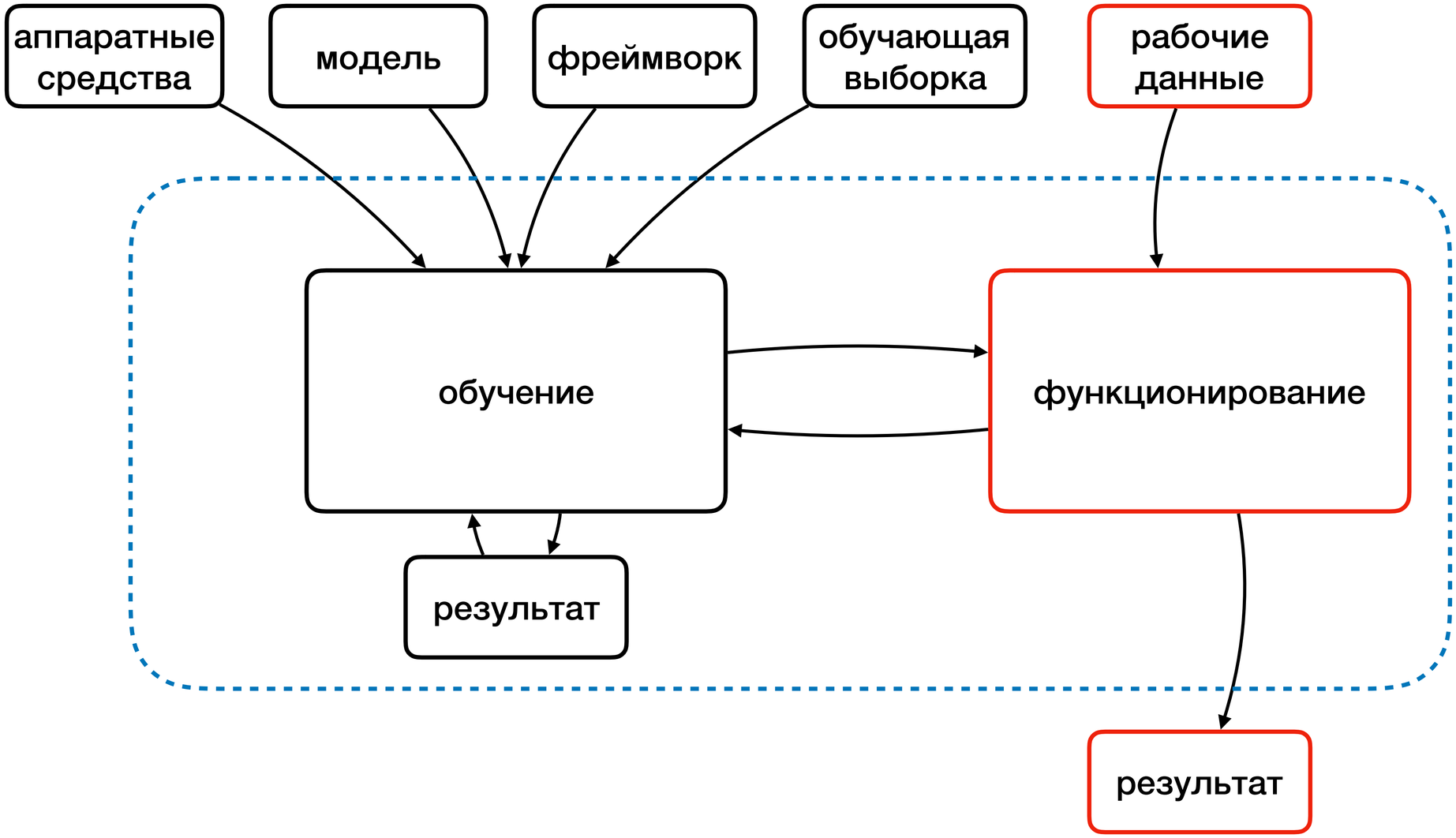
Подходы к защите



Криптографичес-кие (конфиденциаль-ное машинное обучение)

- протоколы распределен-ных вычисления
- гомоморфное шифрование

Подходы к защите



Криптографичес-кие (конфиденциаль-ное машинное обучение)

- протоколы распределен-ных вычисления
- гомоморфное шифрование

Разрабатываемая концепция требований

- Наследует подходы и общую структуру существующий требований к безопасности информации (см., например, Р 1323565.1.012.-2017): обеспечение безопасности всего цикла разработки и эксплуатации системы ИИ
- Предполагает использование различных механизмов обеспечения безопасности
- Формулирует частные требования к каждому механизму