Автоматизация разметки данных с помощью ML-моделей

Куцев Роман
Co-founder & CTO
TrainingData.Solutions

Несколько фактов о себе

- Окончил ВМК МГУ
- Занимаюсь сбором и разметкой данных больше 6 лет
- Являюсь сертифицированным партнером Толоки
- Реализовал более 300 проектов по разметке
- Преподаю краудсорсинг в ШАДе и ВШЭ

Зачем мы собираем и размечаем данные?

А сколько данных нужно собрать?

А сколько данных нужно собрать?

Как выбрать метрику? И как определить какой порог метрики нам нужен?

А сколько данных нужно собрать?

Как выбрать метрику? И как определить какой порог метрики нам нужен?

Какая самая главная метрика?

А сколько данных нужно собрать?

Как выбрать метрику? И как определить какой порог метрики нам нужен?

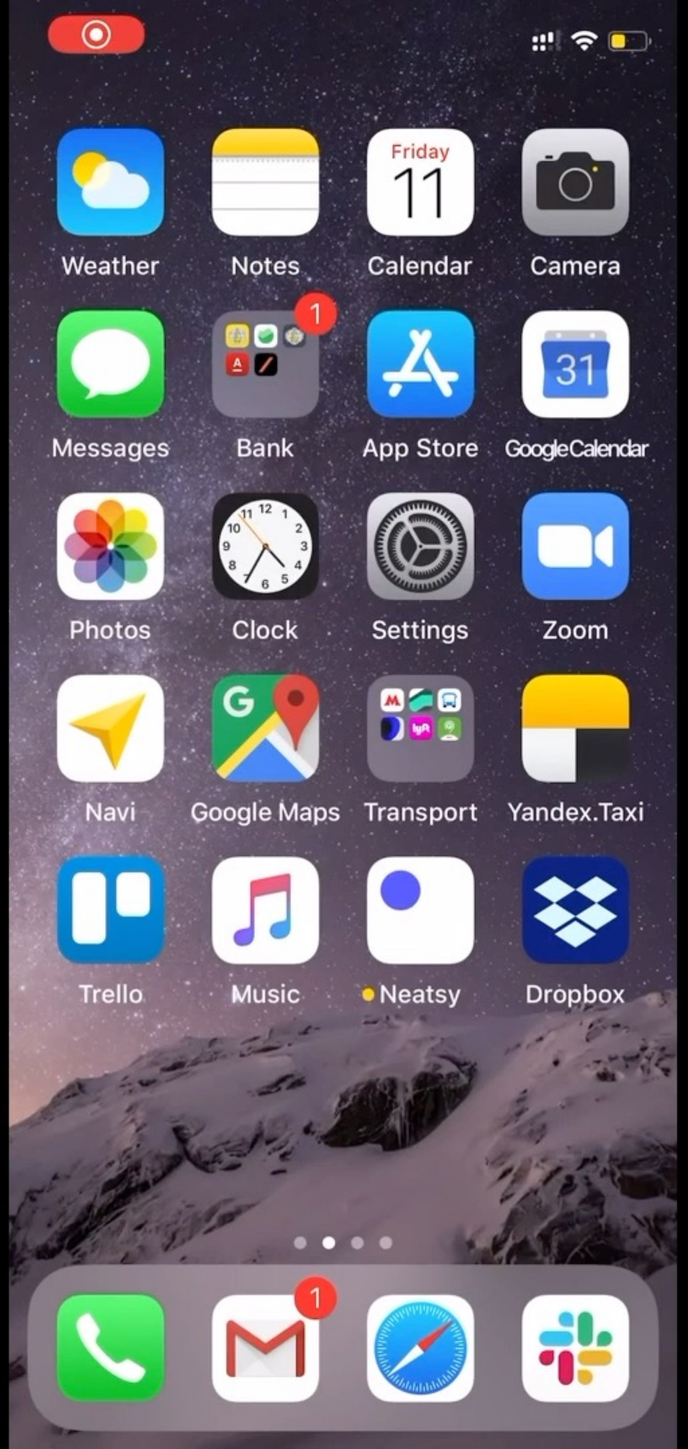
Какая самая главная метрика? **Деньги**

А сколько данных нужно собрать?

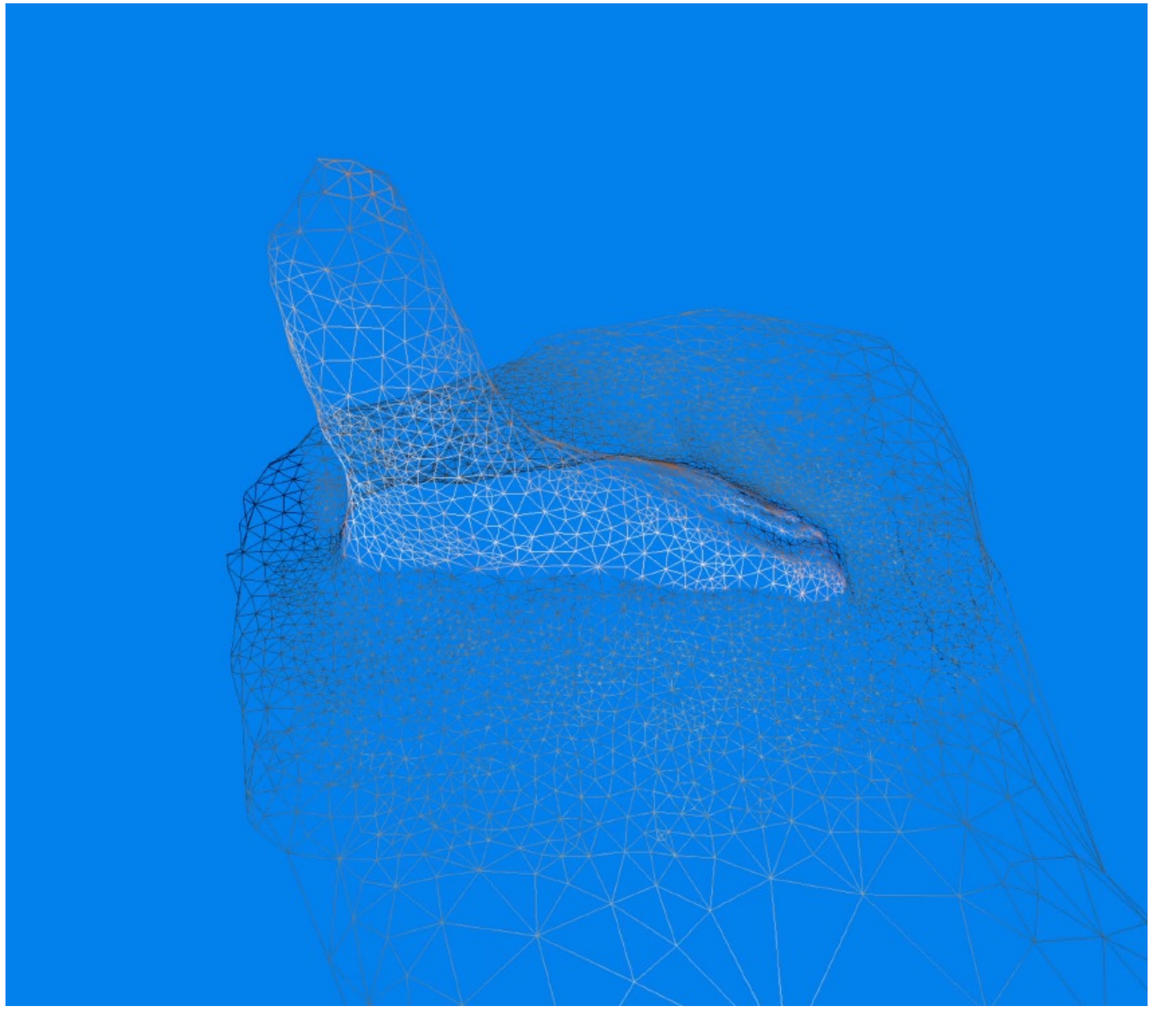
Как выбрать метрику? И как определить какой порог метрики нам нужен?

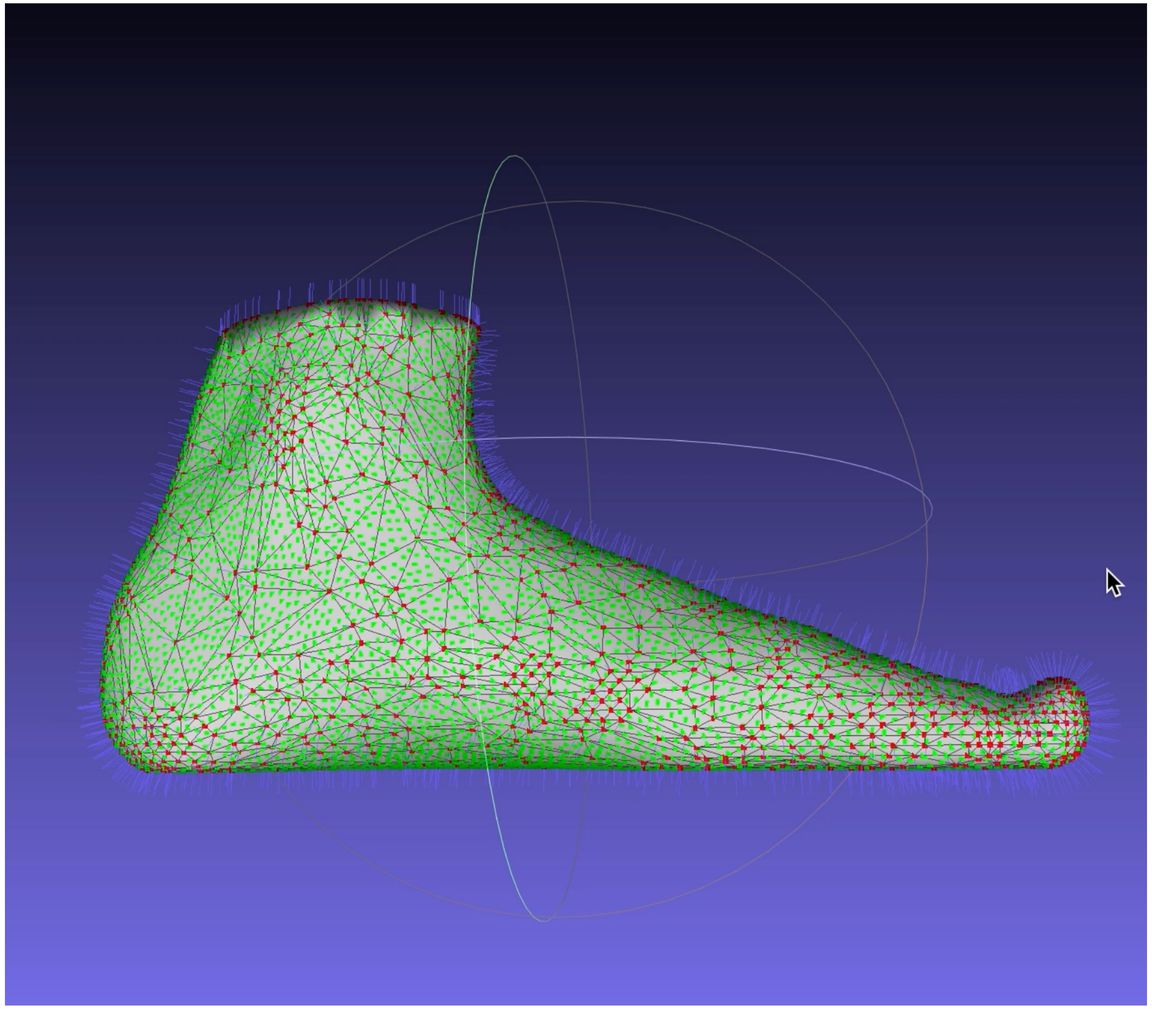
Какая самая главная метрика? **Деньги**

Как мы можем аппроксимировать метрику «Деньги» другими, понятными для нас метриками?


Weather
Notes
Friday
11
Calendar
Camera
Messages
1
Bank
App Store
31
GoogleCalendar
Photos
Clock
Settings
Zoom
Navi
Google Maps
Transport
Yandex.Taxi
Trello
Music
Neatsy
Dropbox
1

Neatsy





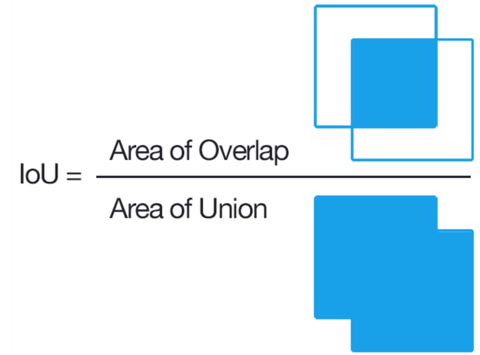
Neatsy

Какую выбрать метрику?

Какой порог метрики нам нужен?

Цель

Получить Intersection over Union (IoU) > 0.95

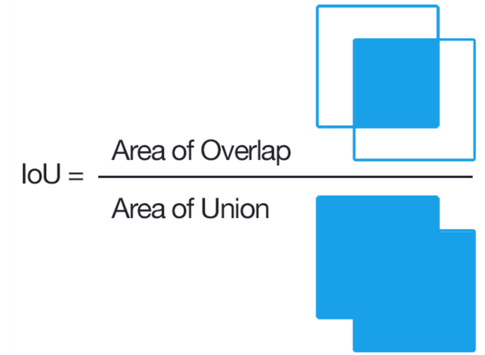


Вопросы

1. Как составить отражающий реальность валидационный датасет?
2. Какую архитектуру выбрать?
3. Сколько данных нужно?

Цель

Получить Intersection over Union (IoU) > 0.95



Вопросы

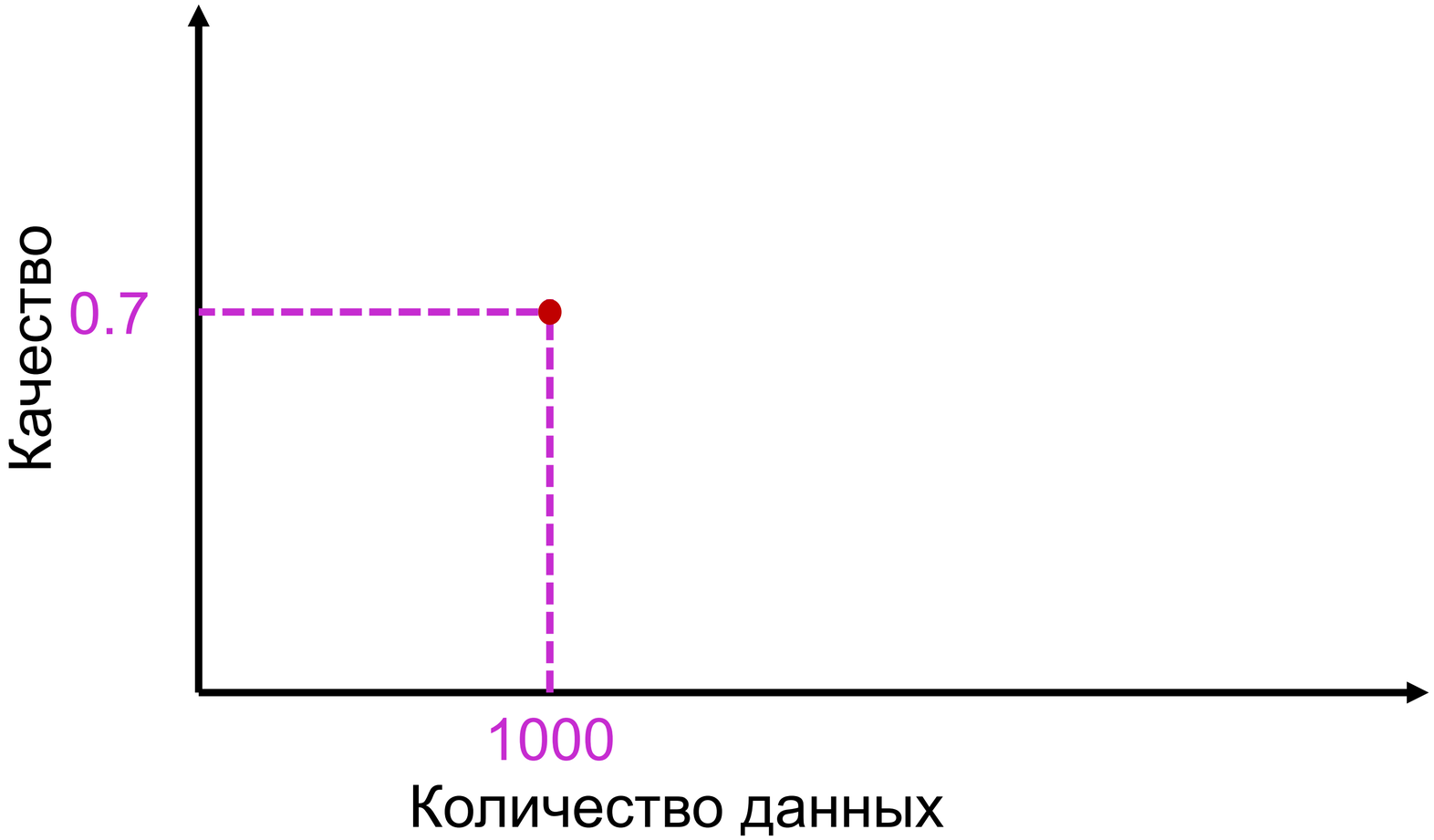
1. Как составить отражающий реальность валидационный датасет?
2. Какую архитектуру выбрать?
3. Сколько данных нужно?

Шаг 1. Разметим 1000 изображений

- Почему 1000? Эмпирически, по опыту в других задачах.

Шаг 1. Разметим 1000 изображений

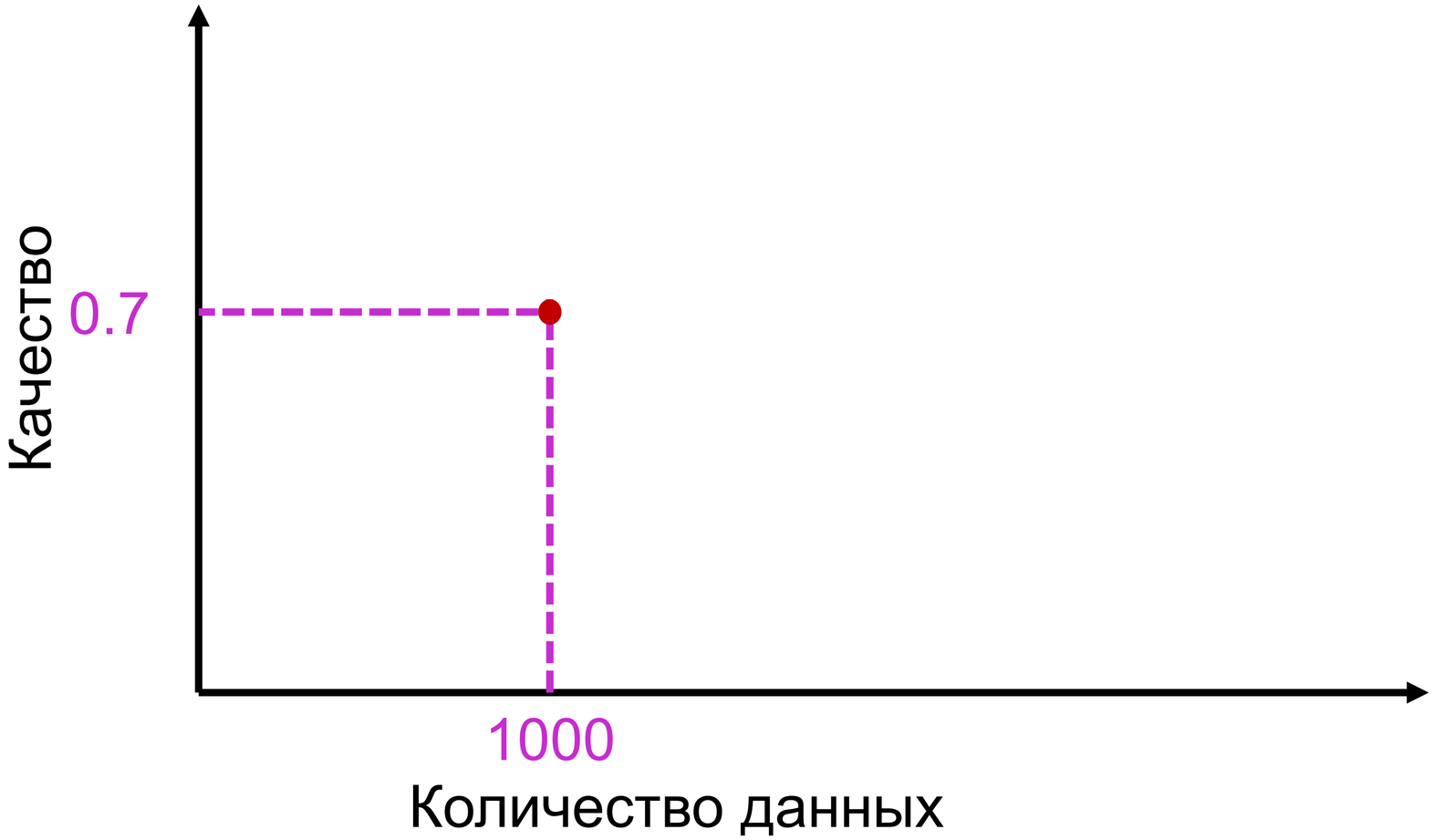
- Почему 1000? Эмпирически, по опыту в других задачах.



Шаг 1. Разметим 1000 изображений

- Почему 1000? Эмпирически, по опыту в других задачах.

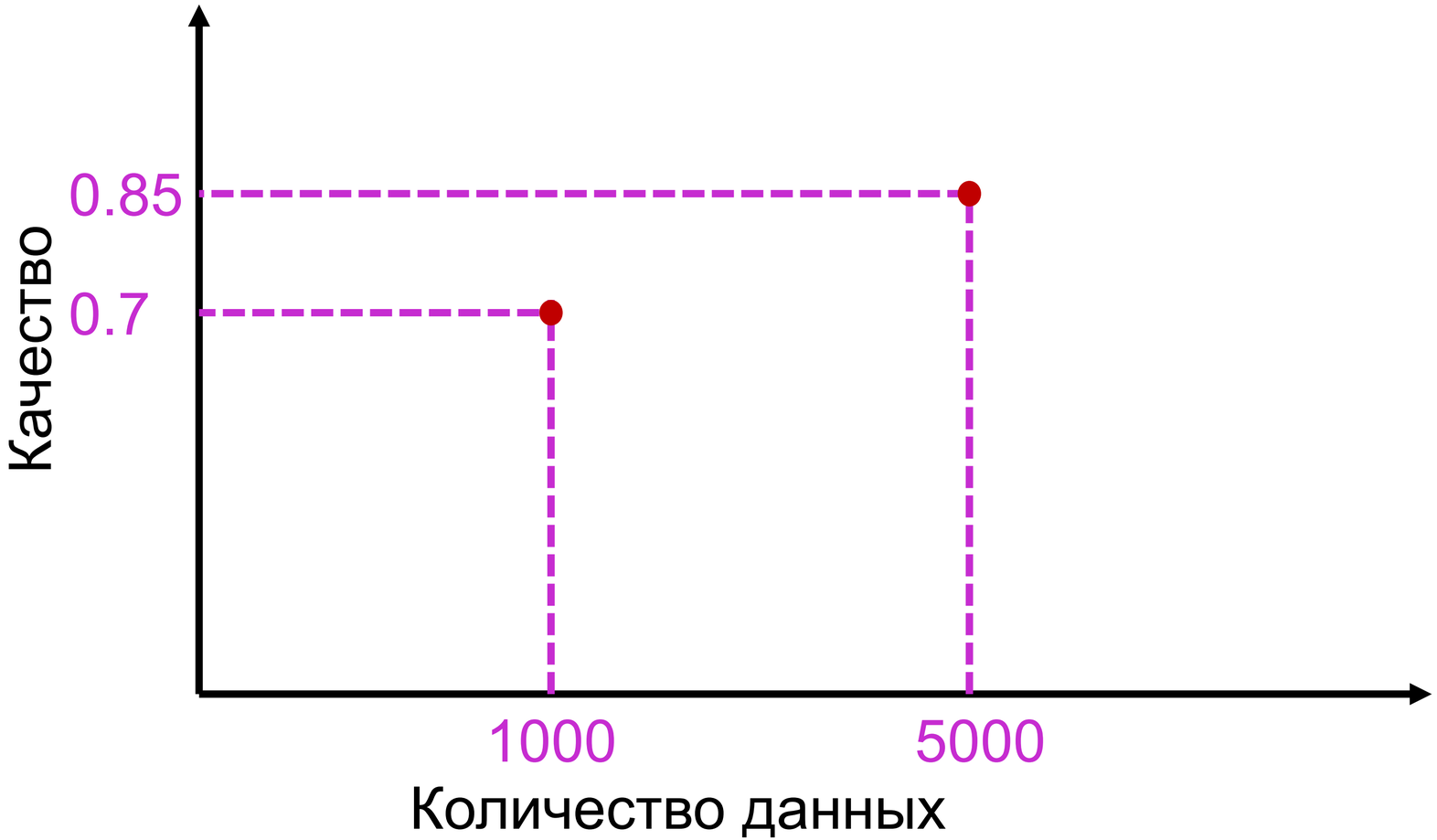
Шаг 2. Разметим 5000 изображений



Шаг 1. Разметим 1000 изображений

- Почему 1000? Эмпирически, по опыту в других задачах.

Шаг 2. Разметим 5000 изображений

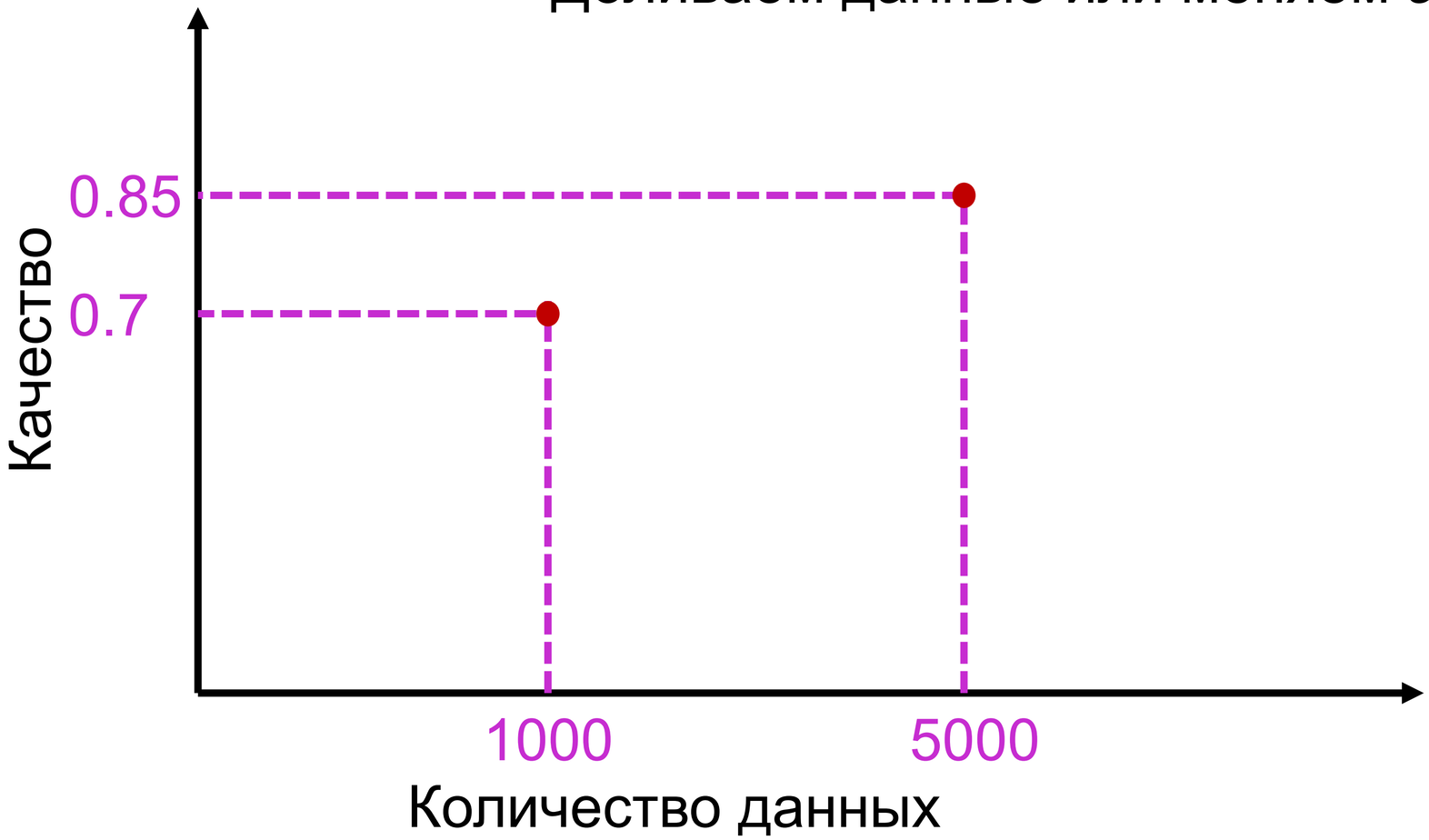


Шаг 1. Разметим 1000 изображений

- Почему 1000? Эмпирически, по опыту в других задачах.

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

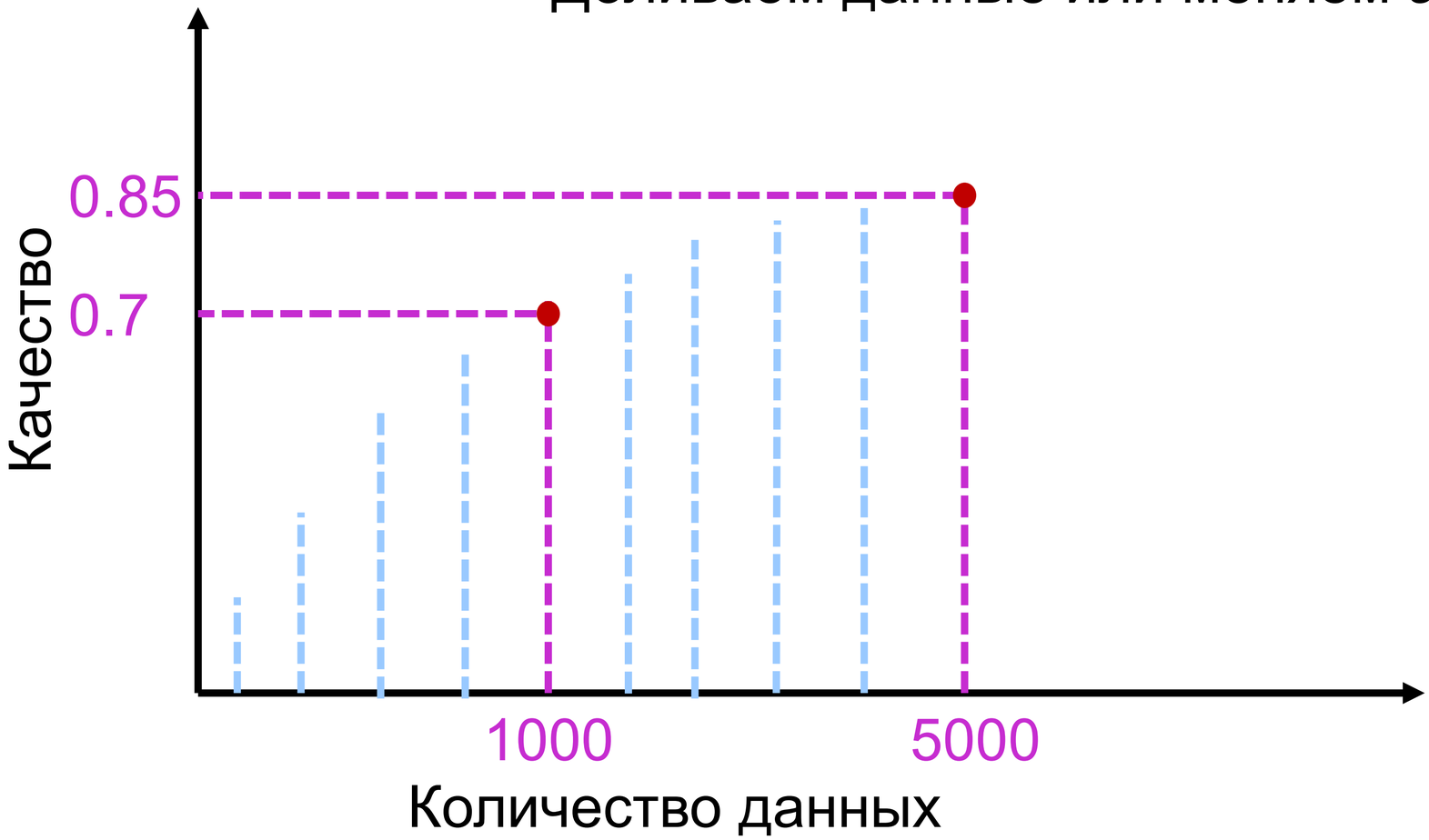


Шаг 1. Разметим 1000 изображений

- Почему 1000? Эмпирически, по опыту в других задачах.

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

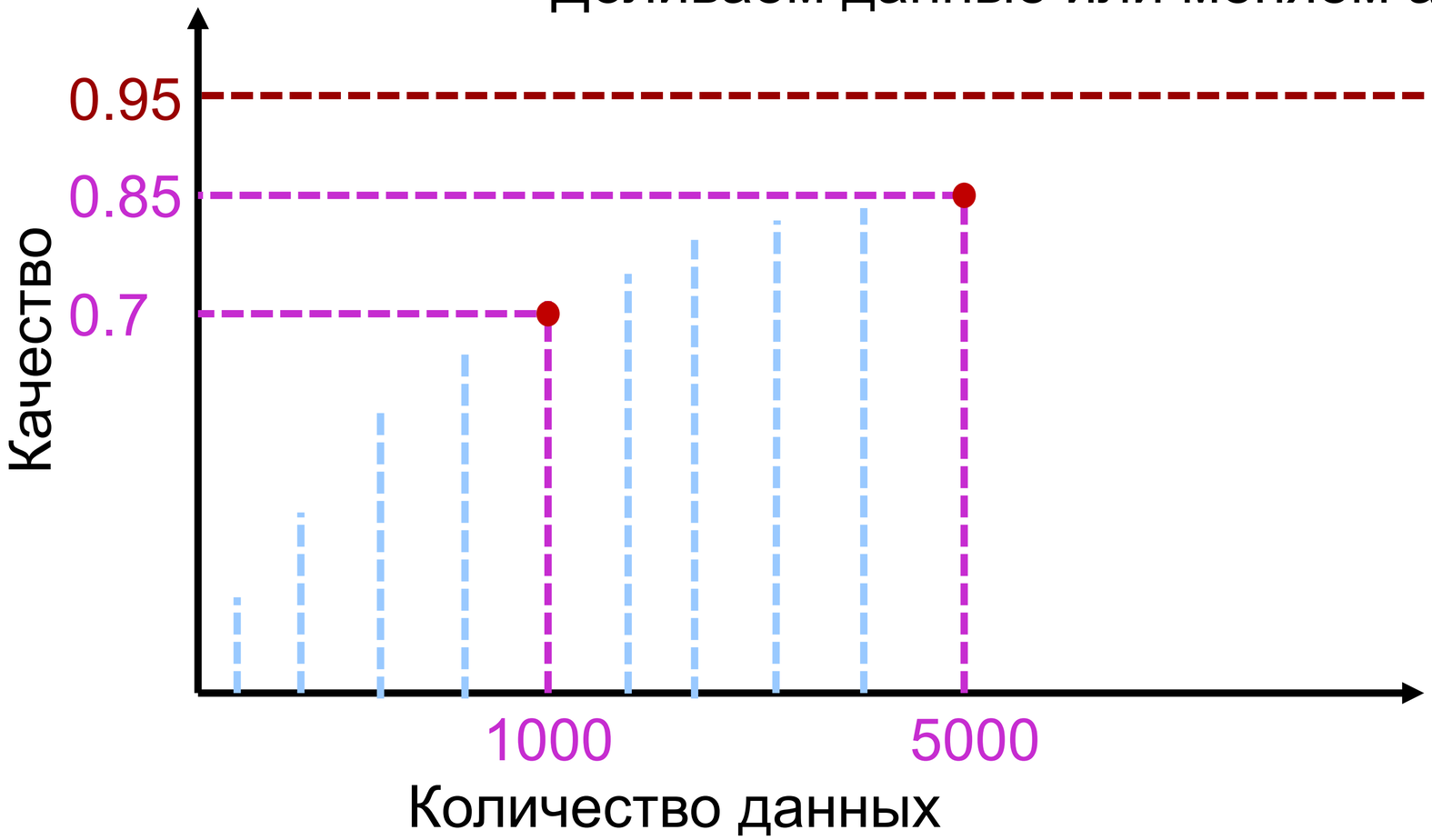


Шаг 1. Разметим 1000 изображений

- Почему 1000? Эмпирически, по опыту в других задачах.

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

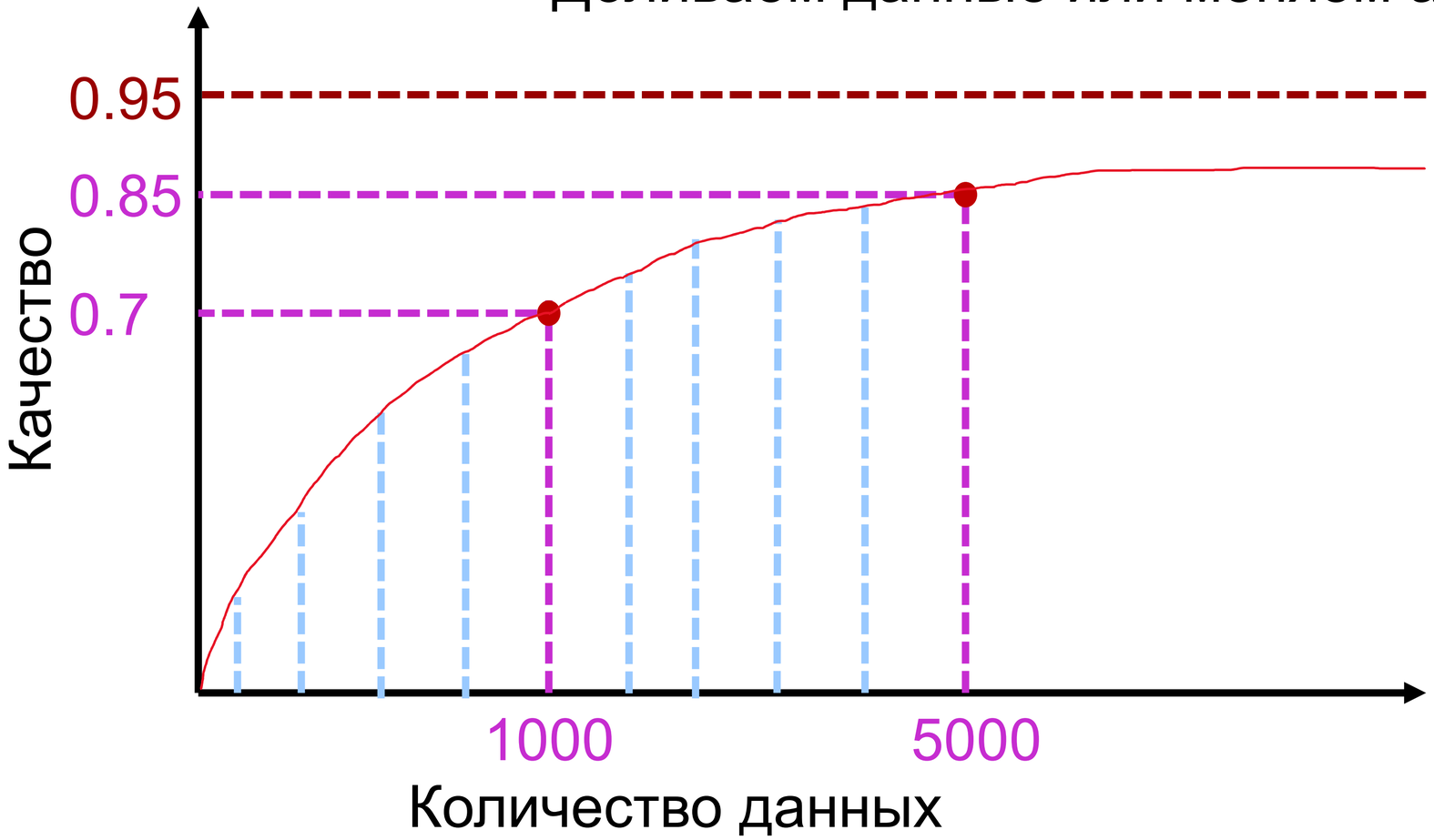


Шаг 1. Разметим 1000 изображений

- Почему 1000? Эмпирически, по опыту в других задачах.

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?



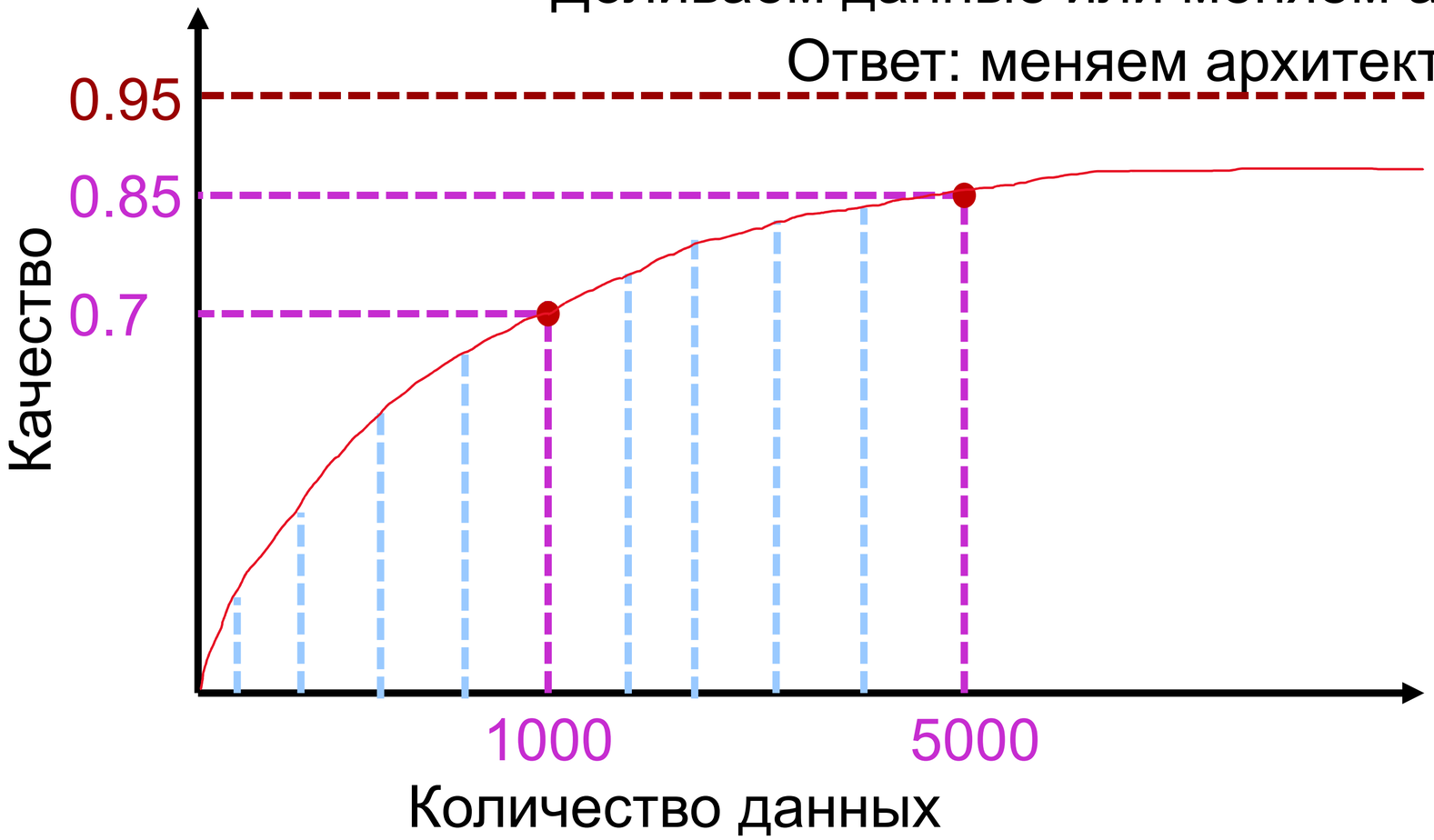
Шаг 1. Разметим 1000 изображений

- Почему 1000? Эмпирически, по опыту в других задачах.

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

Ответ: меняем архитектуру



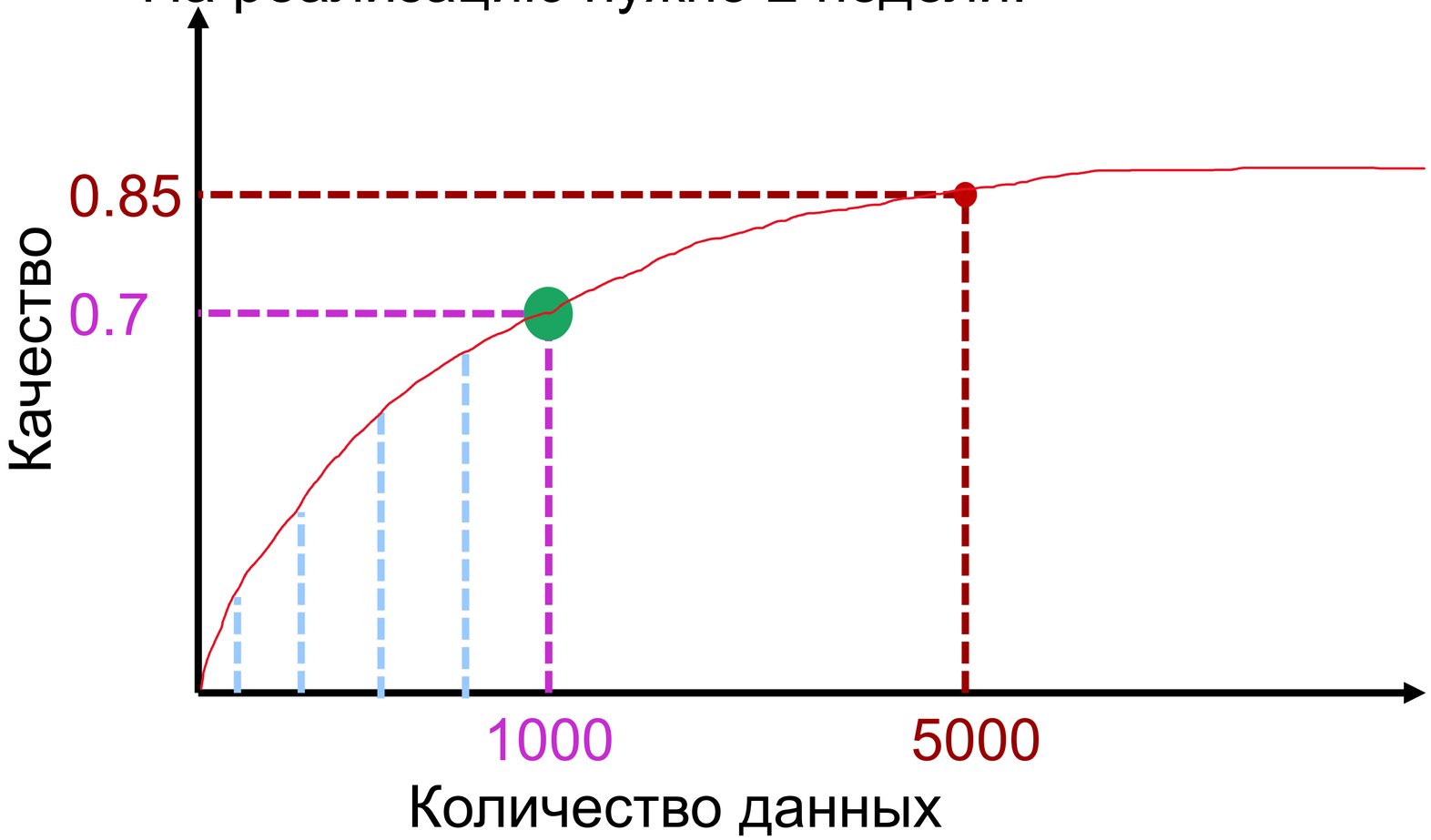
Кейс 2:

Сейчас имеем 0.7, нужно 0.85.

Вася прочитал статью о новом алгоритме.

Уверен, что получится достичь 0.85.

На реализацию нужно 2 недели.



Кейс 2:

Сейчас имеем 0.7, нужно 0.85.

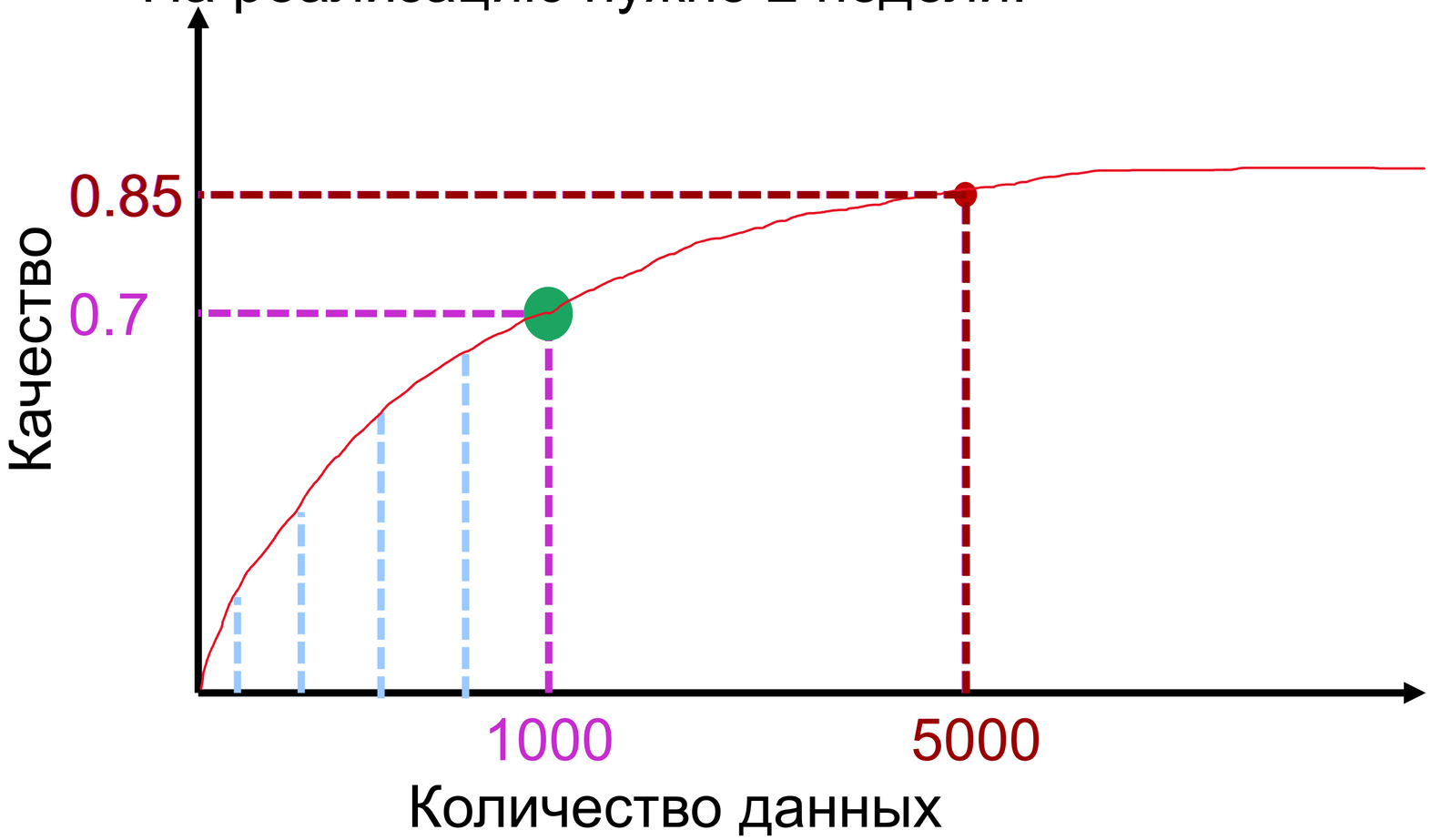
Вася прочитал статью о новом алгоритме.

Уверен, что получится достичь 0.85.

На реализацию нужно 2 недели.

Варианты:

1) Вася реализует новый алгоритм
2) Размечаем больше данных



Кейс 2:

Сейчас имеем 0.7, нужно 0.85.

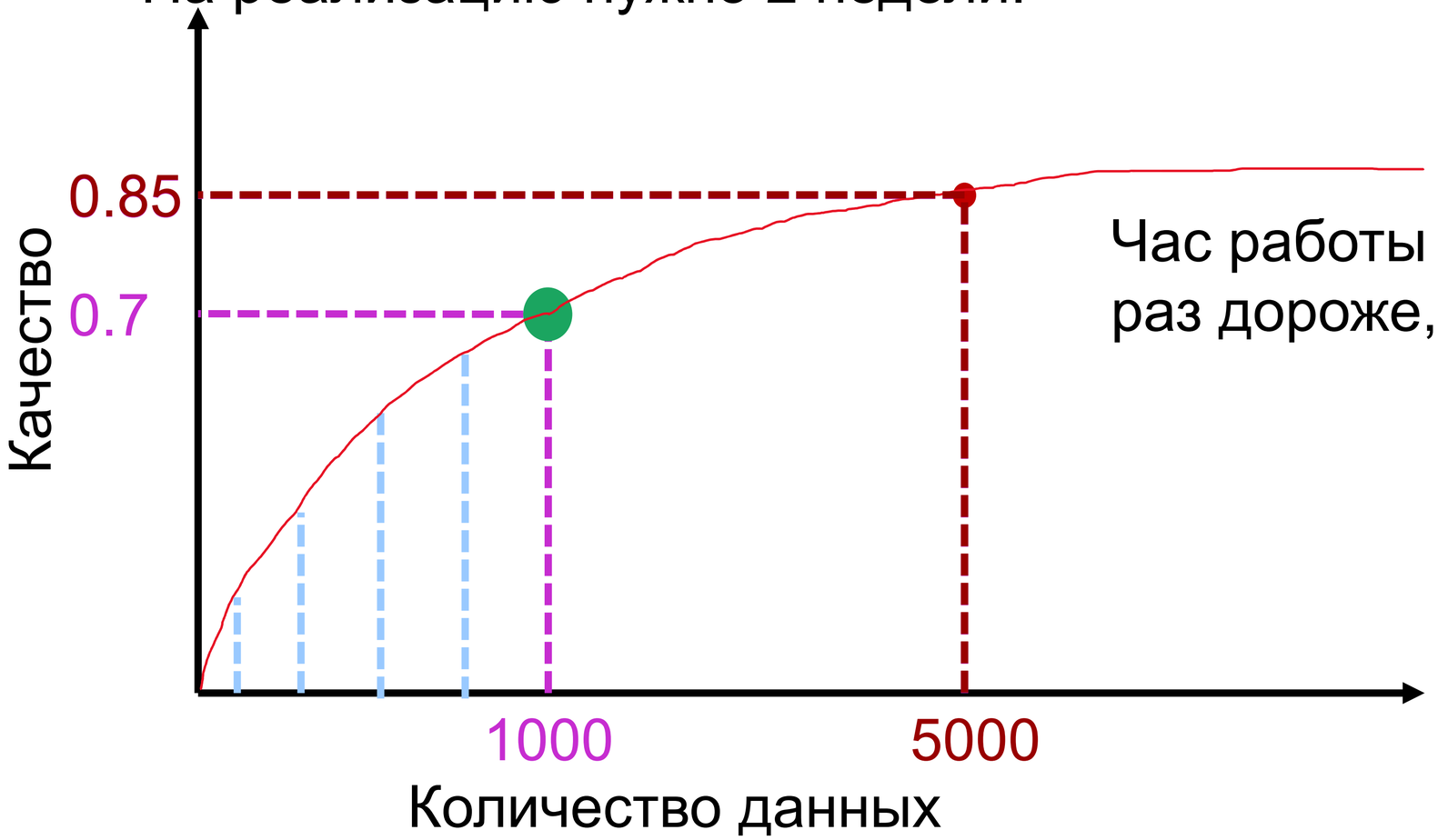
Вася прочитал статью о новом алгоритме.

Уверен, что получится достичь 0.85.

На реализацию нужно 2 недели.

Варианты:

1) Вася реализует новый алгоритм
2) Размечаем больше данных



Час работы ML специалиста стоит в **10-15** раз дороже, чем час работы асессора

Кейс 2:

Сейчас имеем 0.7, нужно 0.85.

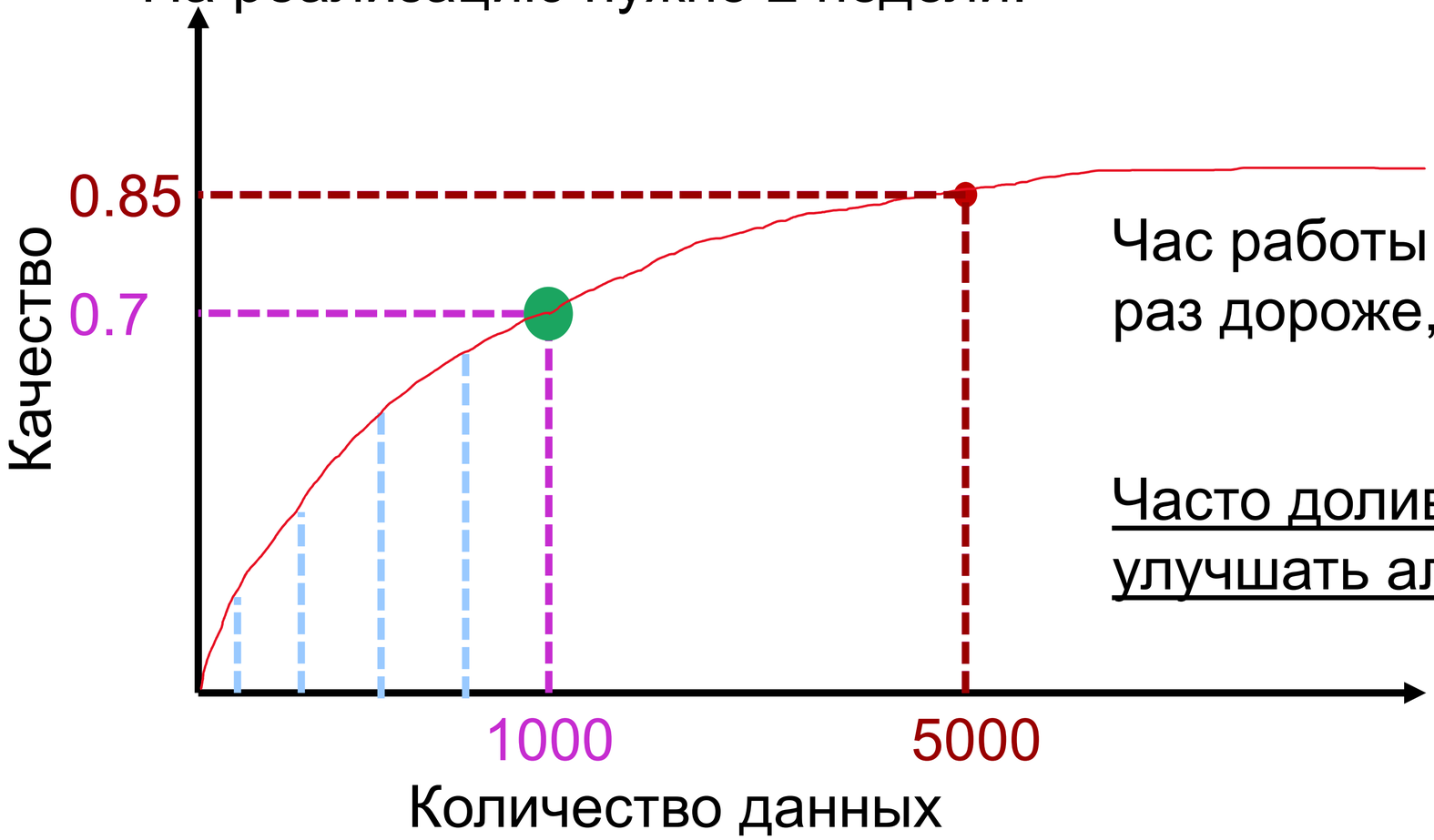
Вася прочитал статью о новом алгоритме.

Уверен, что получится достичь 0.85.

На реализацию нужно 2 недели.

Варианты:

1) Вася реализует новый алгоритм
2) Размечаем больше данных



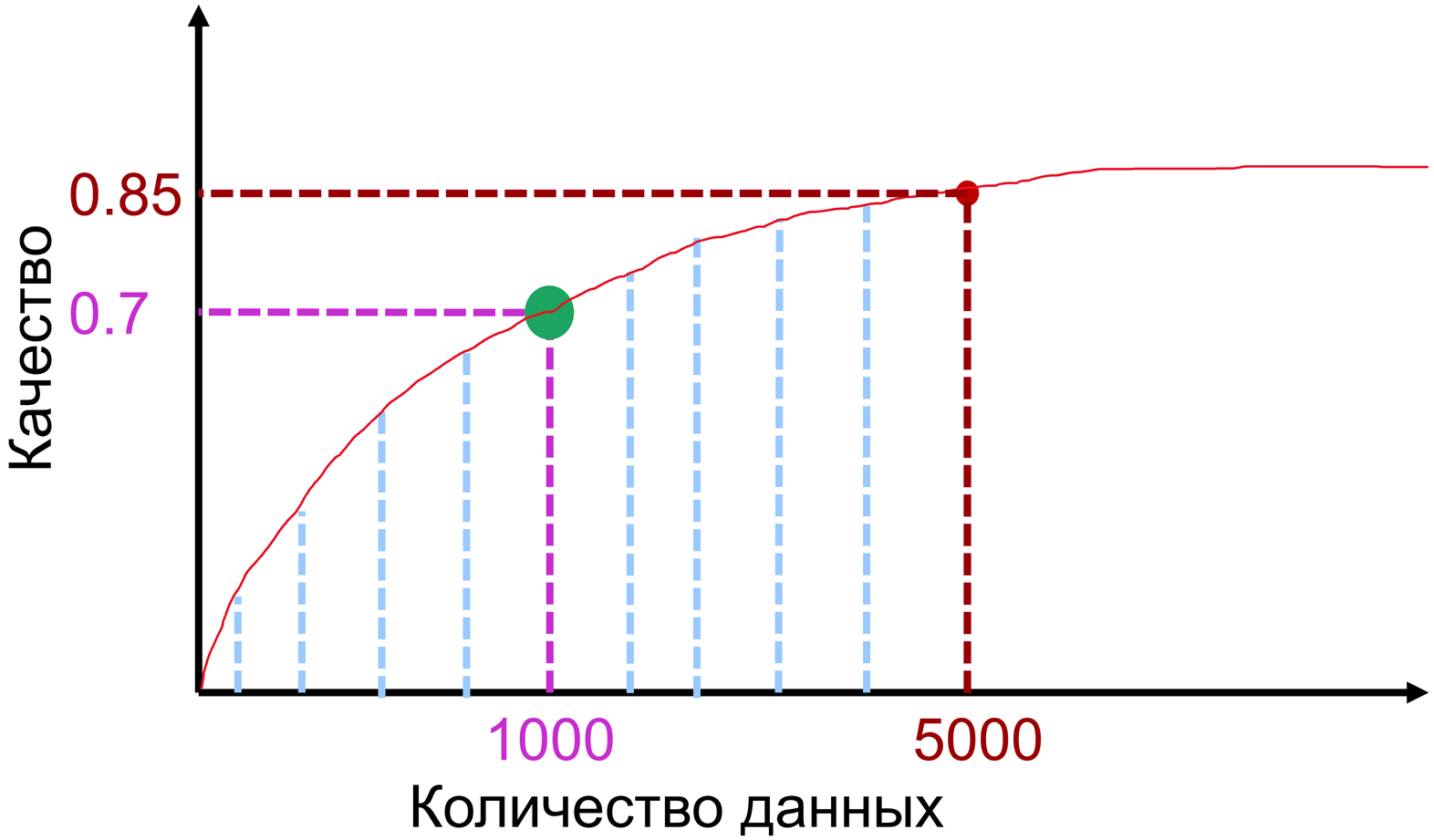
Час работы ML специалиста стоит в **10-15** раз дороже, чем час работы асессора

<u>Часто доливать данные дешевле, чем улучшать алгоритмы</u>

Кейс 3:

Сейчас имеем 0.7, нужно 0.85.

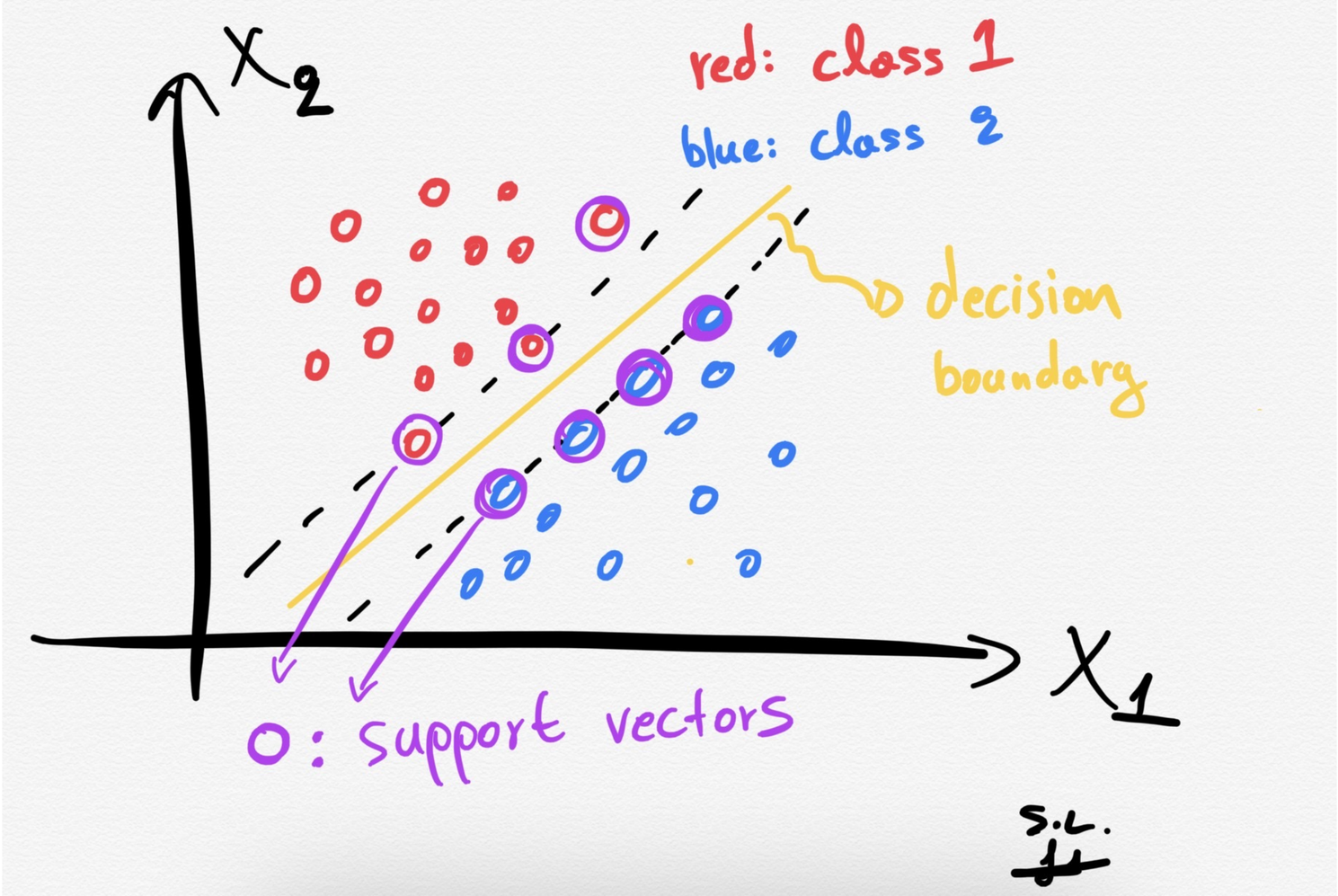
Можем ли мы меньше, чем за 5000 картинок достичь качества 0.85?



Кейс 3:

Сейчас имеем 0.7, нужно 0.85.

Можем ли мы меньше, чем за 5000 картинок достичь качества 0.85?



Active Learning

Цель: достичь как можно лучшего качества модели, использовав как можно меньше обучающих примеров.

Подробнее тут: https://youtu.be/M-geBlVIUXY

Есть различные методики:

- uncertainty sampling
- query by committee
- version space reduction
- variance reduction

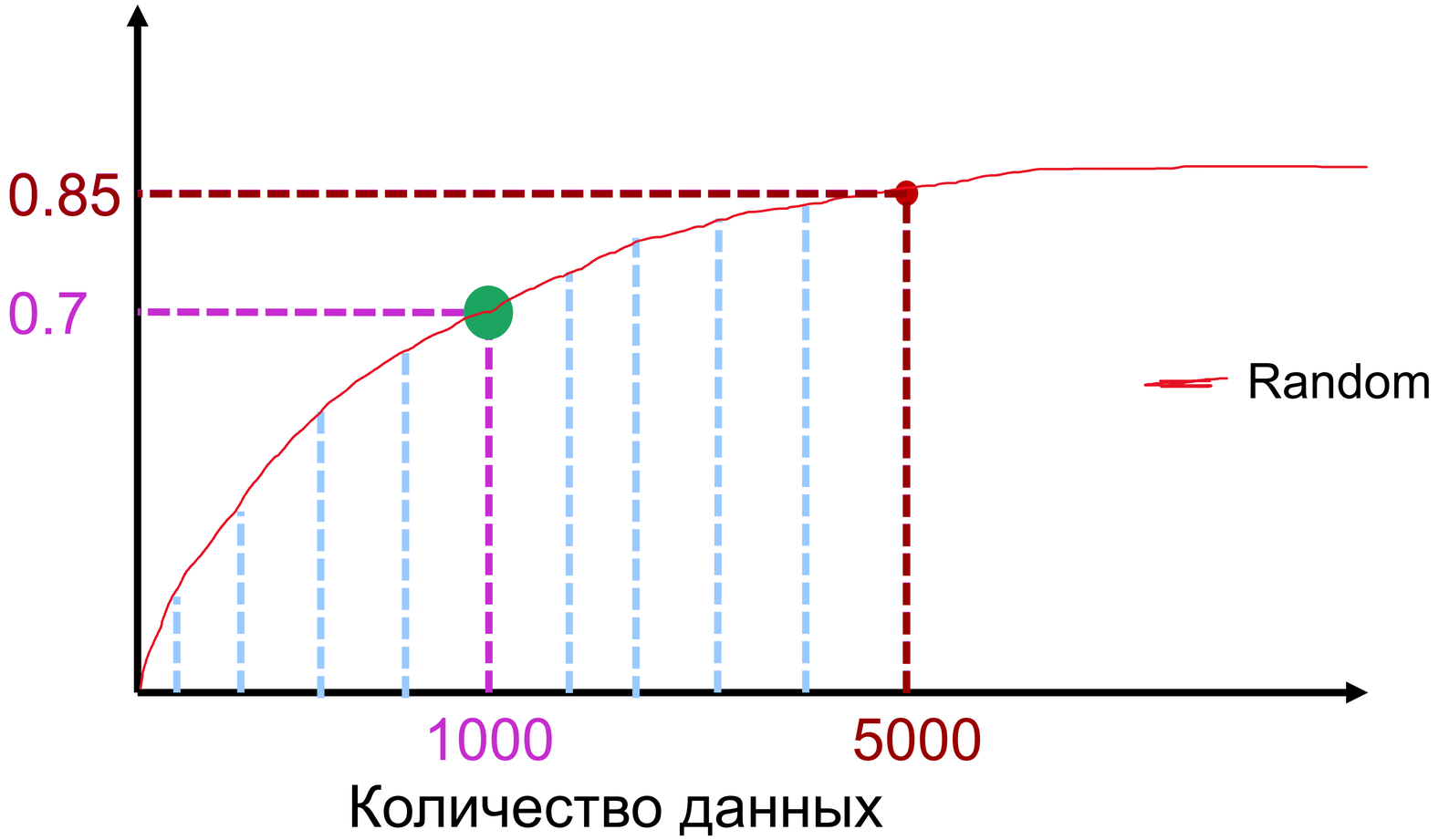
Active Learning

- Неразмечанные данные прогоняются через алгоритм машинного обучения.

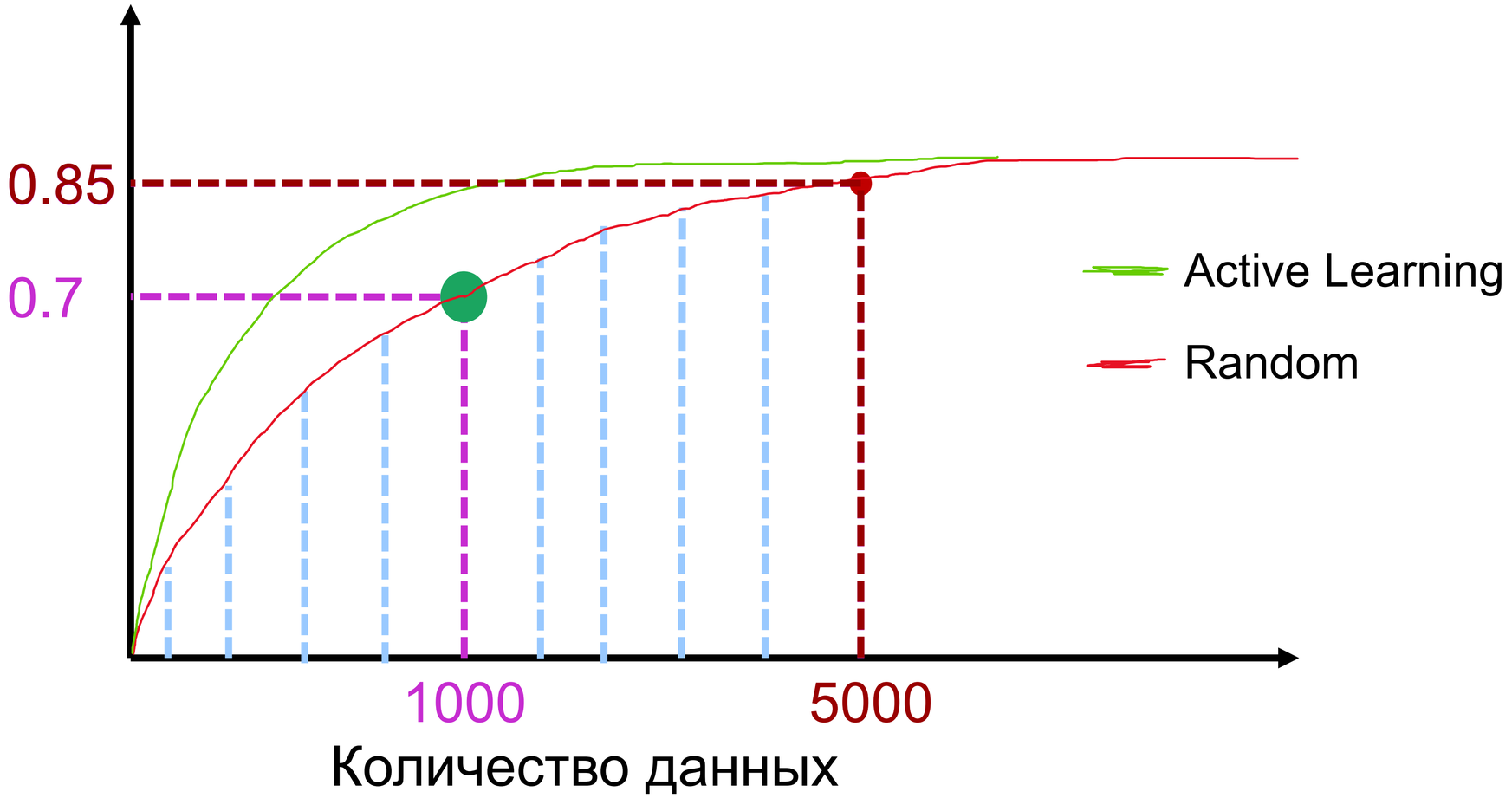
- Сначала размечаются данные с низкой уверенностью в ответе. Как правило они оказываются сложными и алгоритм с ними плохо справляется.

- При добавлении именно этих данных мы вносим максимальную информацию в наш алгоритм

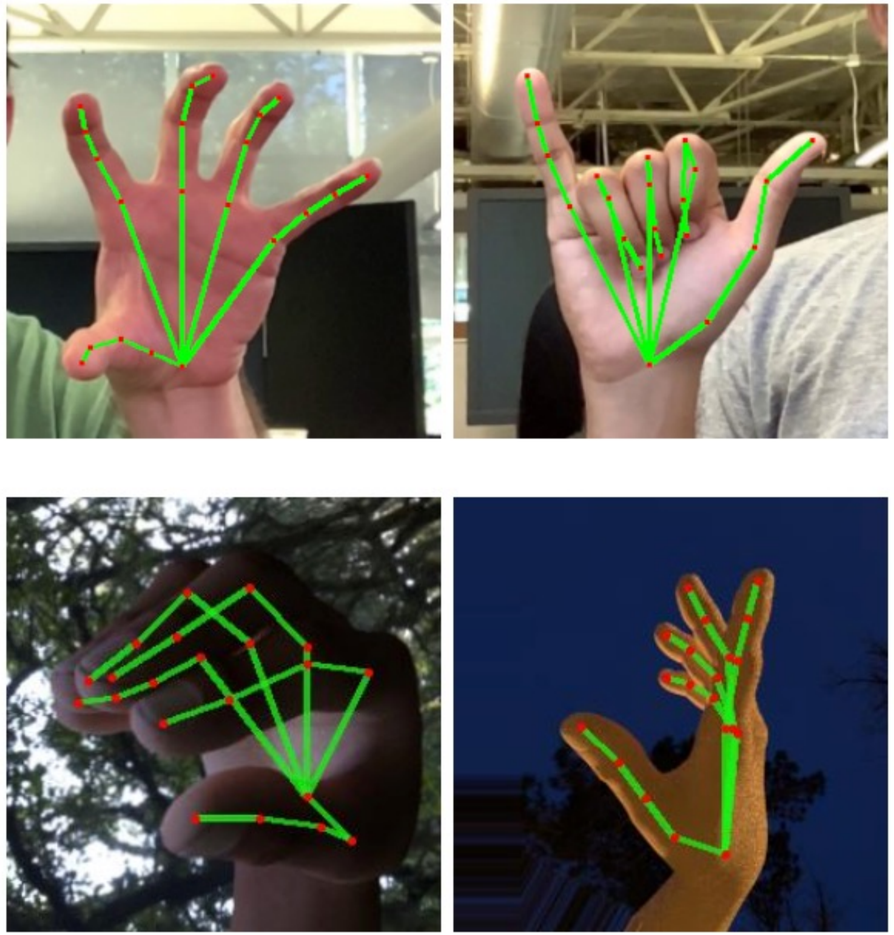
Active Learning



Active Learning



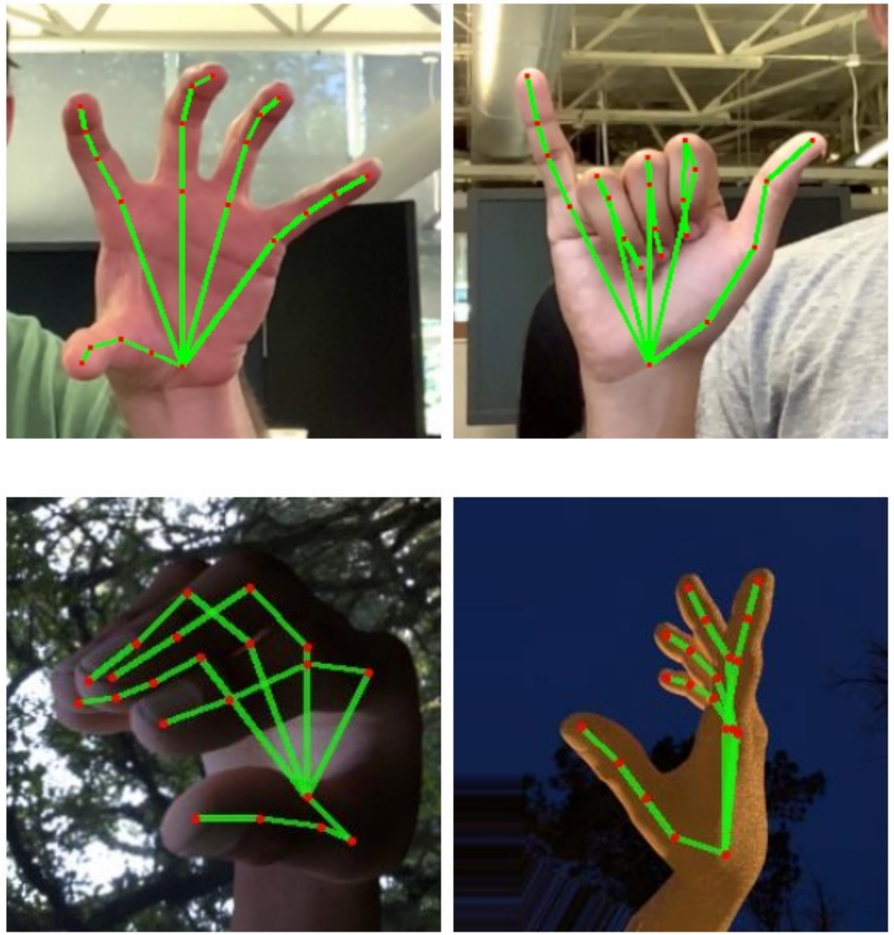
Synthetic datasets



Synthetic datasets

Используется, когда:

- Реальные данные собрать невозможно
- Сбор реальных данных очень дорогой



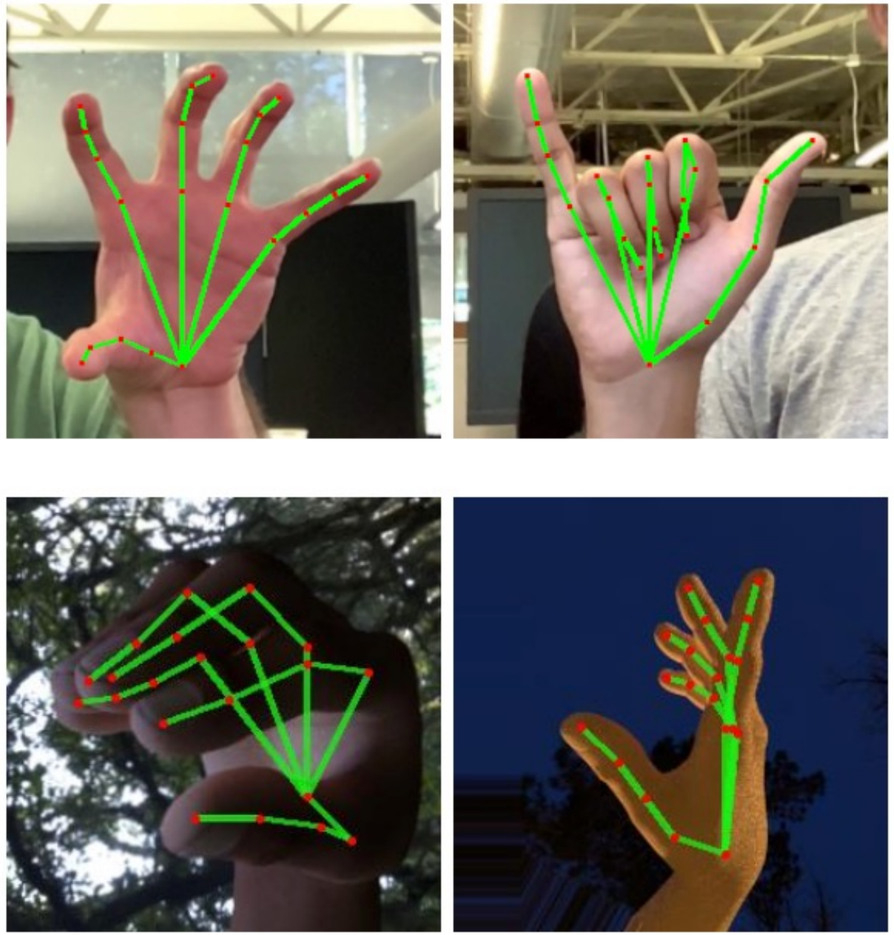
Synthetic datasets

Используется, когда:

- Реальные данные собрать невозможно
- Сбор реальных данных очень дорогой

Особенности:

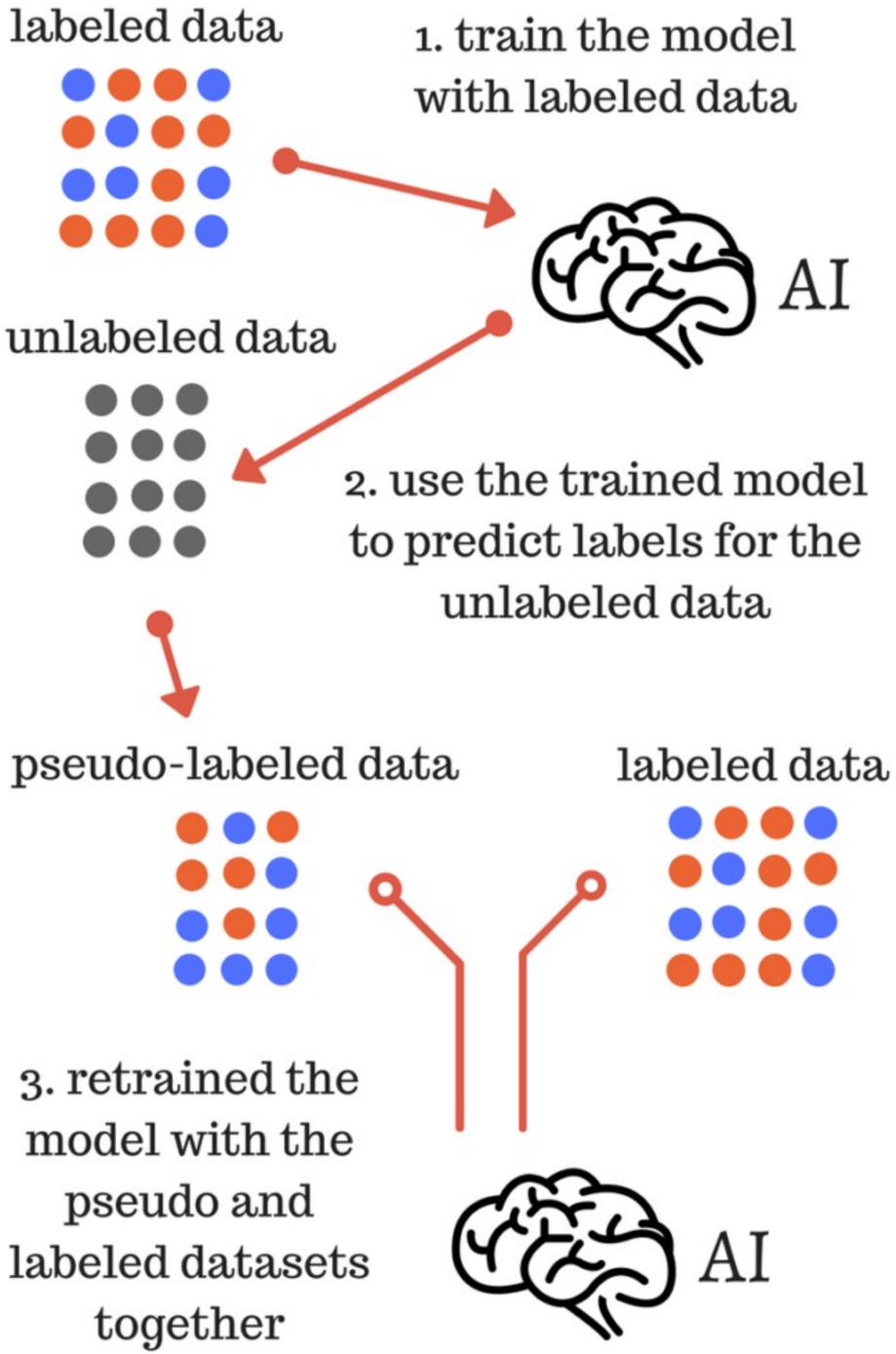
- Трудно сделать синтетику похожей на реальные данные
- Сильное переобучение на синтетику
- Часто нестабильный результат
- Синтетику можно за очень дешево очень много нарендерить



Synthetic datasets

Pseudo Labeling

Pseudo Labeling

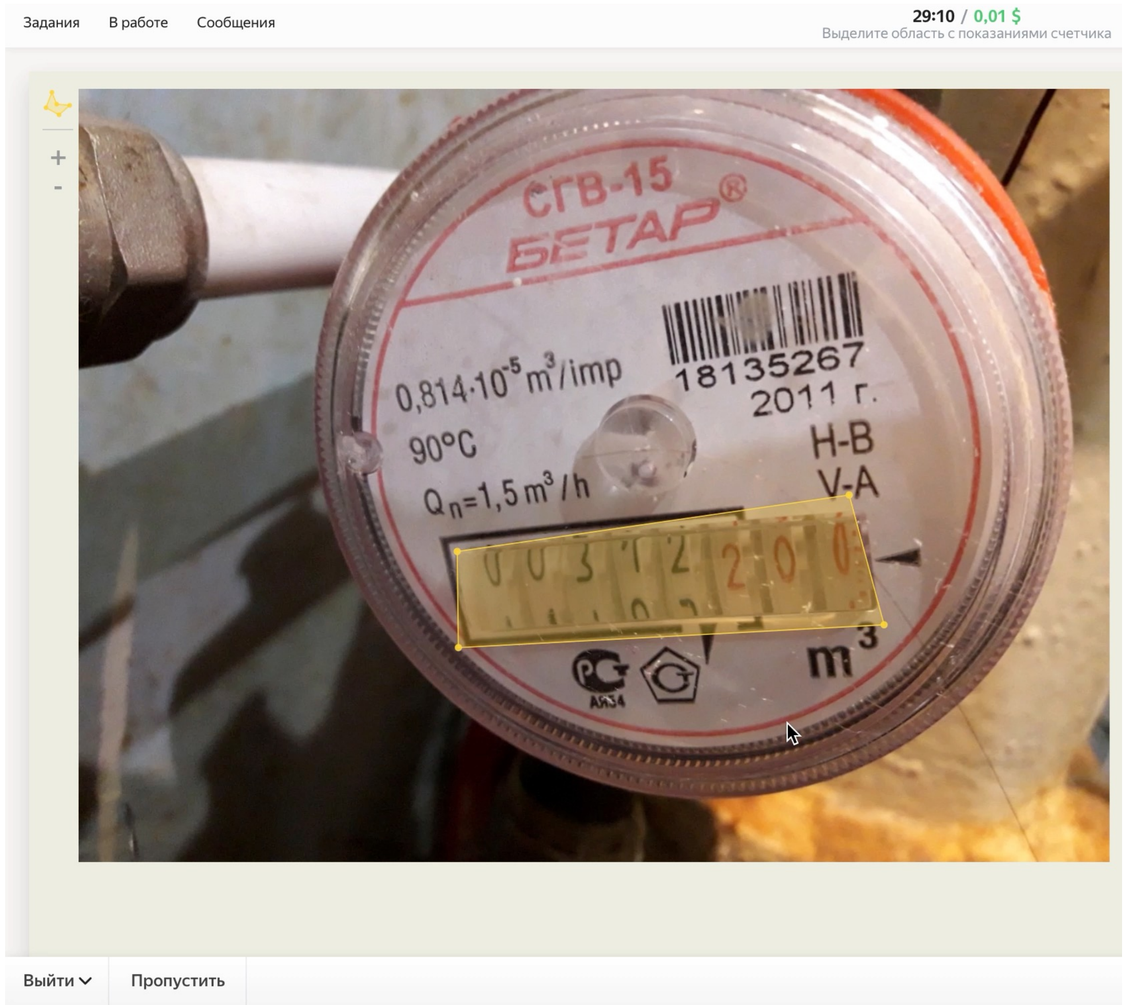


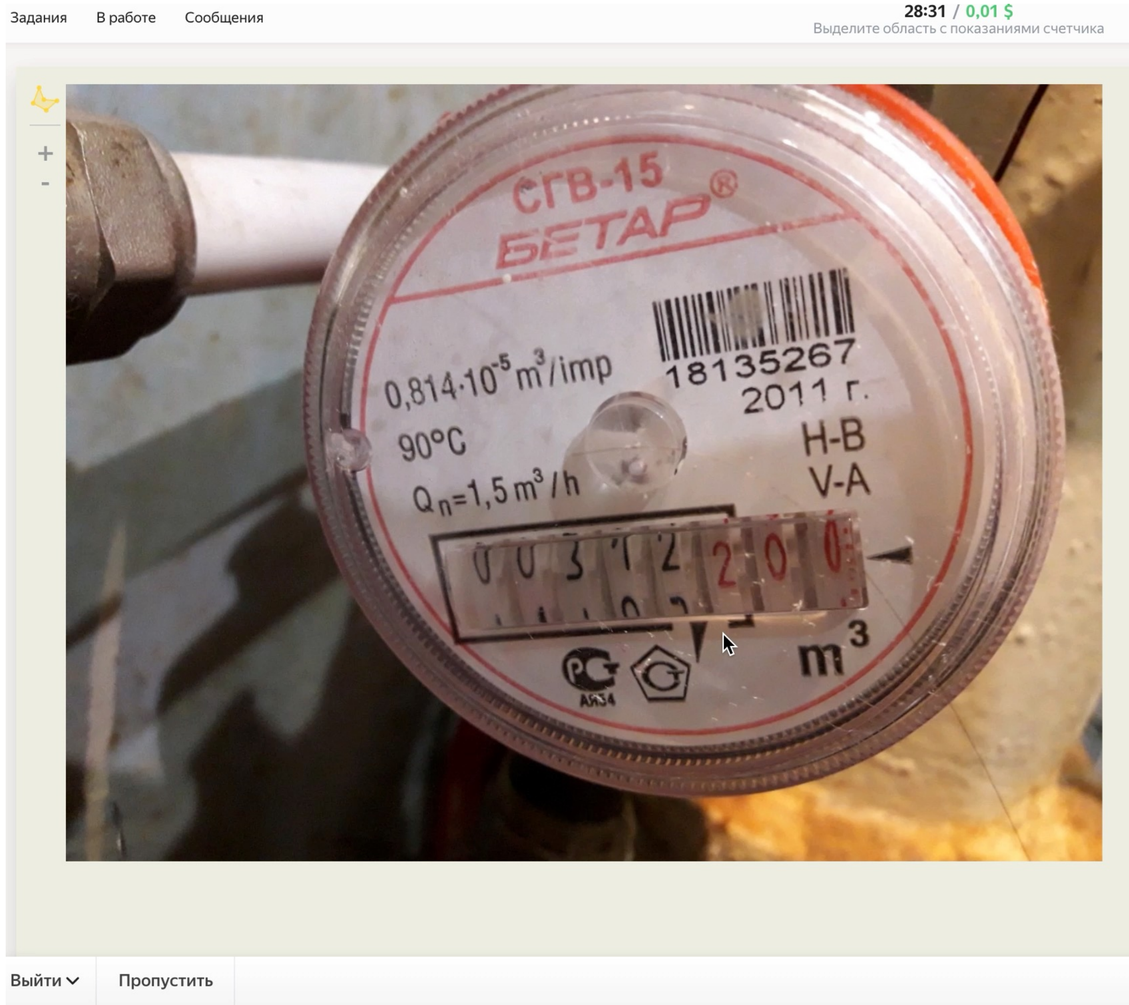
towardsdatascience.com/pseudo-labeling-to-deal-with-small-datasets-what-why-how-fd6f903213af

Pre-labelling

Pre-labelling

Данные перед разметкой прогоняются через алгоритм машинного обучения.
Разметчики поправляют ошибки алгоритма.





Pre-labelling

Данные перед разметкой прогоняются через алгоритм машинного обучения. Разметчики поправляют ошибки алгоритма.

- Ускорение разметки до x10 раз
- Необходимо иметь хоть как-то работающую модель

Pre-labelling

Данные перед разметкой прогоняются через алгоритм машинного обучения. Разметчики поправляют ошибки алгоритма.

- Ускорение разметки до x10 раз
- Необходимо иметь хоть как-то работающую модель
- Немного смещаются ответы в разметке

Pre-labelling

Какой номер у автомобиля?

Pre-labelling

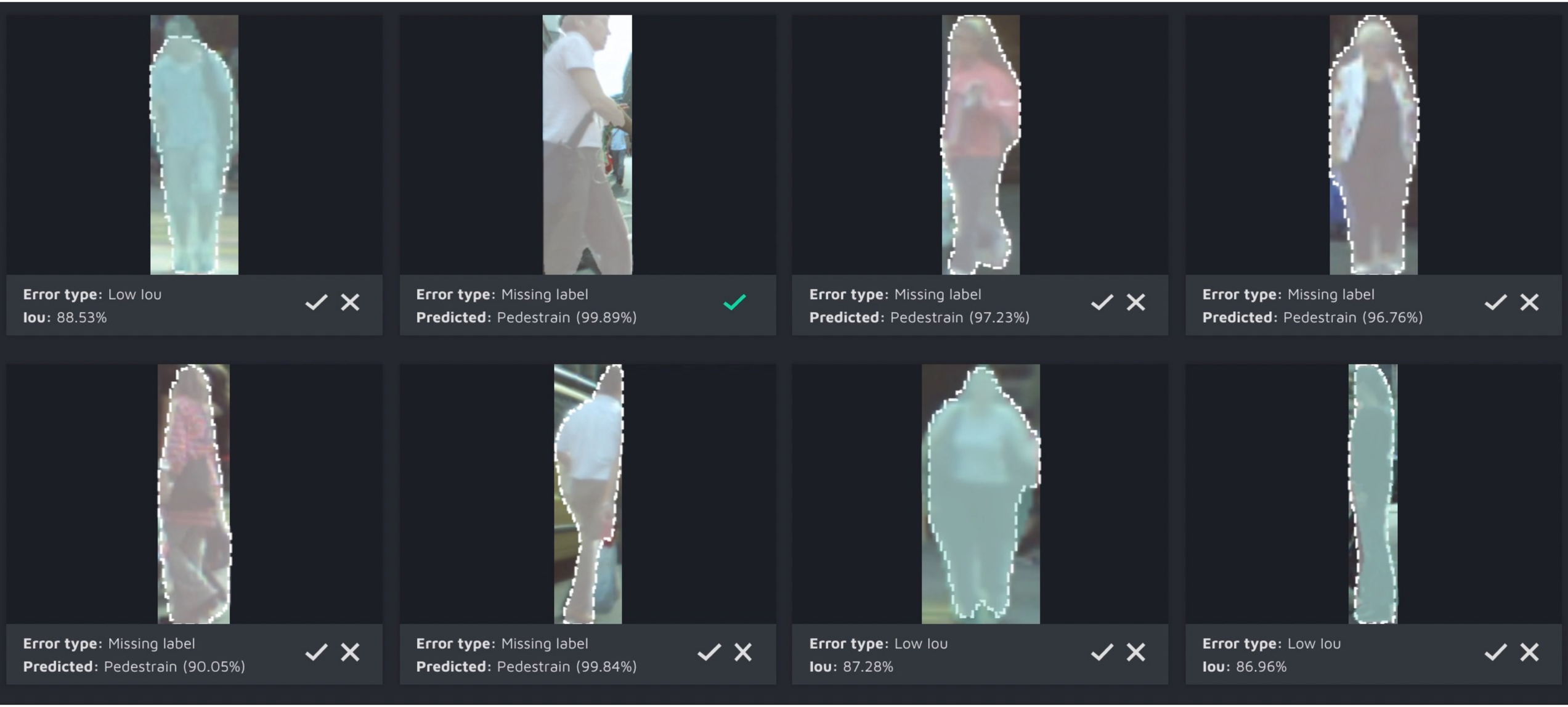
Какой номер у автомобиля?

Если бы нейронная сеть предзаполнила «B999BX74», вы бы стали исправлять?

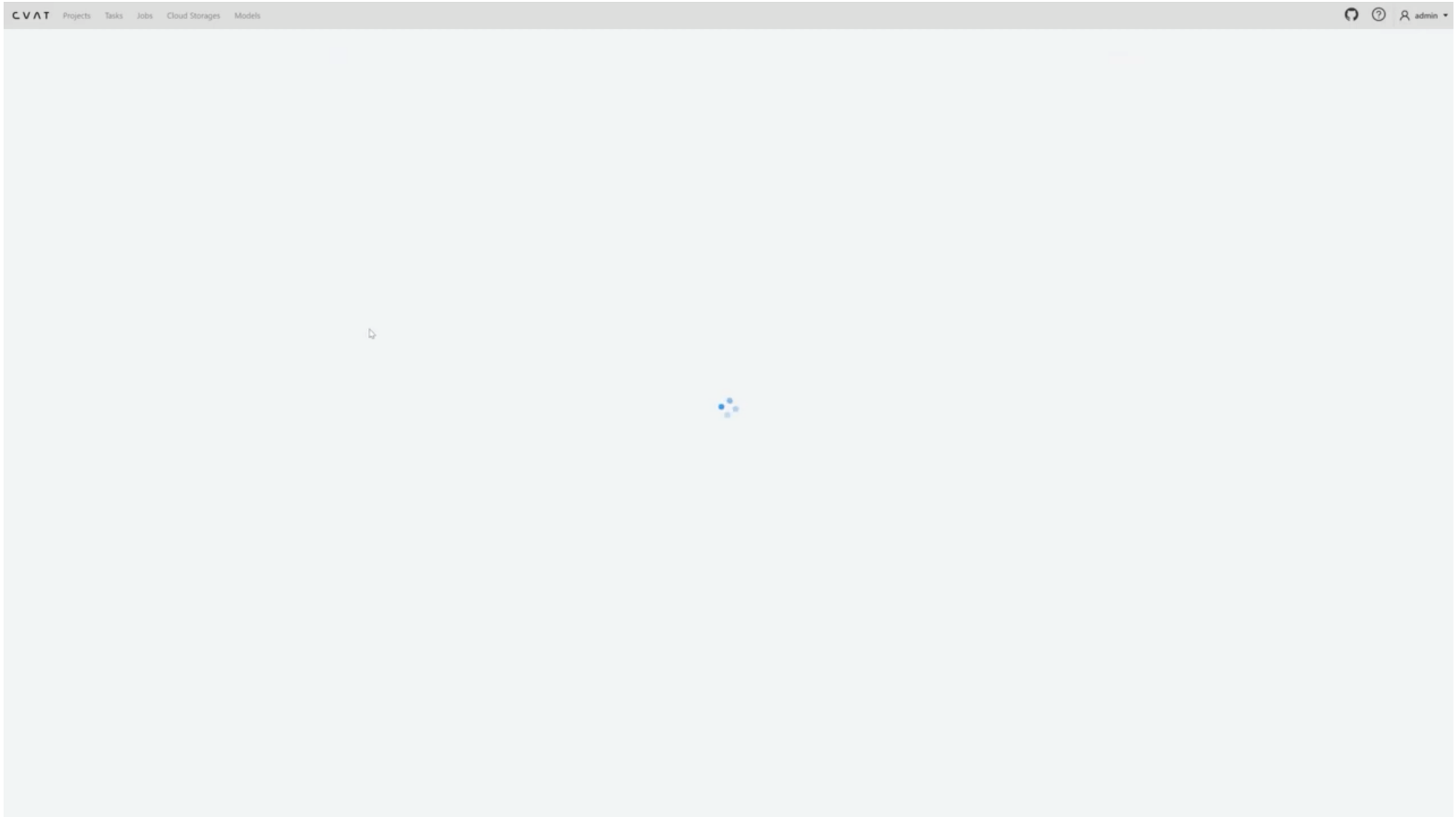
Добавление ответов ML модели в консенсус

Добавление ответов ML модели в консенсус

Поиск ошибок в разметке с помощью ML модели



Интерактивная сегментация SAM и HRnet



Semi-automatic annotation: Point2Label

Semi-automatic annotation: Bbox2Label

Human in the Loop

- Концепция, которая объединяет искусственный и человеческий интеллект
- Основную работу выполняет ML модель, люди размечают только самые сложные случаи
- Такой подход позволяет достигнуть точности 99.9%
- Применяется в задачах, где стоимость ошибки очень высока

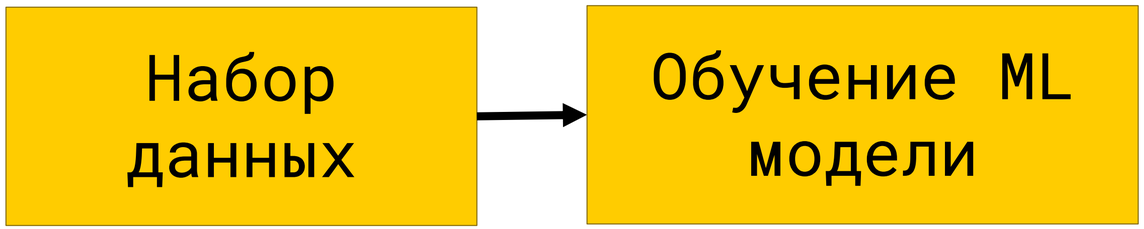
Human in the Loop

Система фиксации нарушений

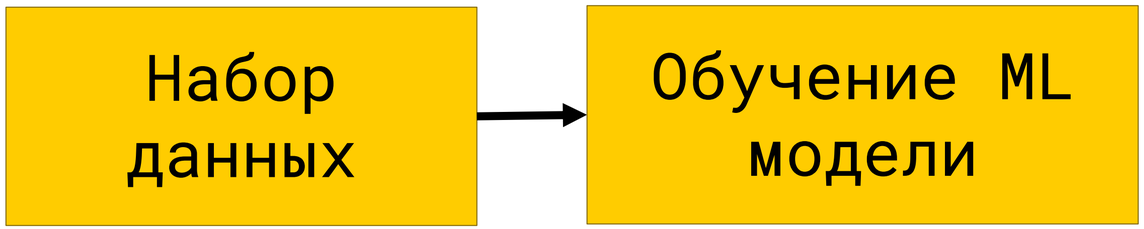
Human in the Loop

Набор данных

Human in the Loop

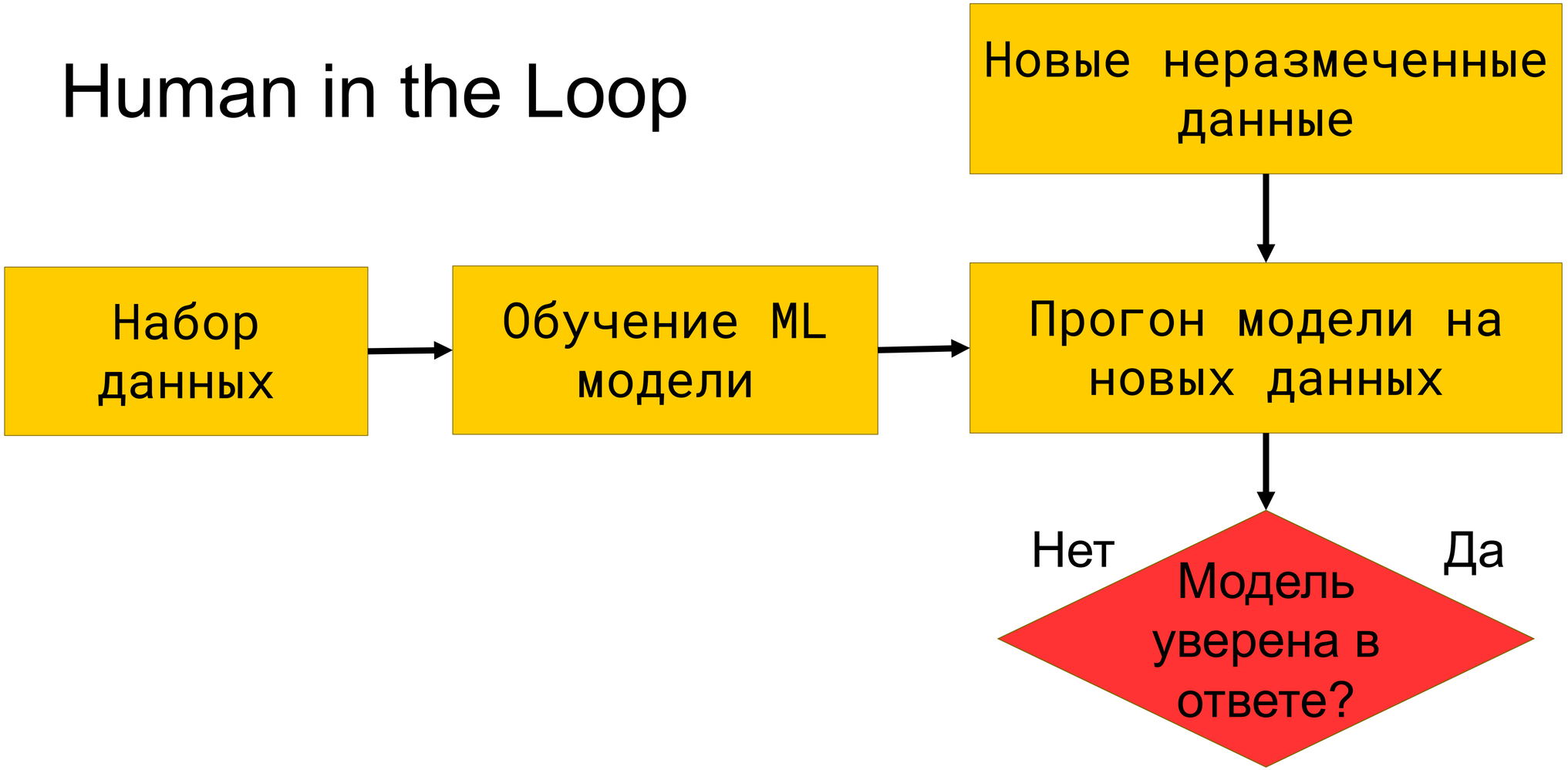


Human in the Loop

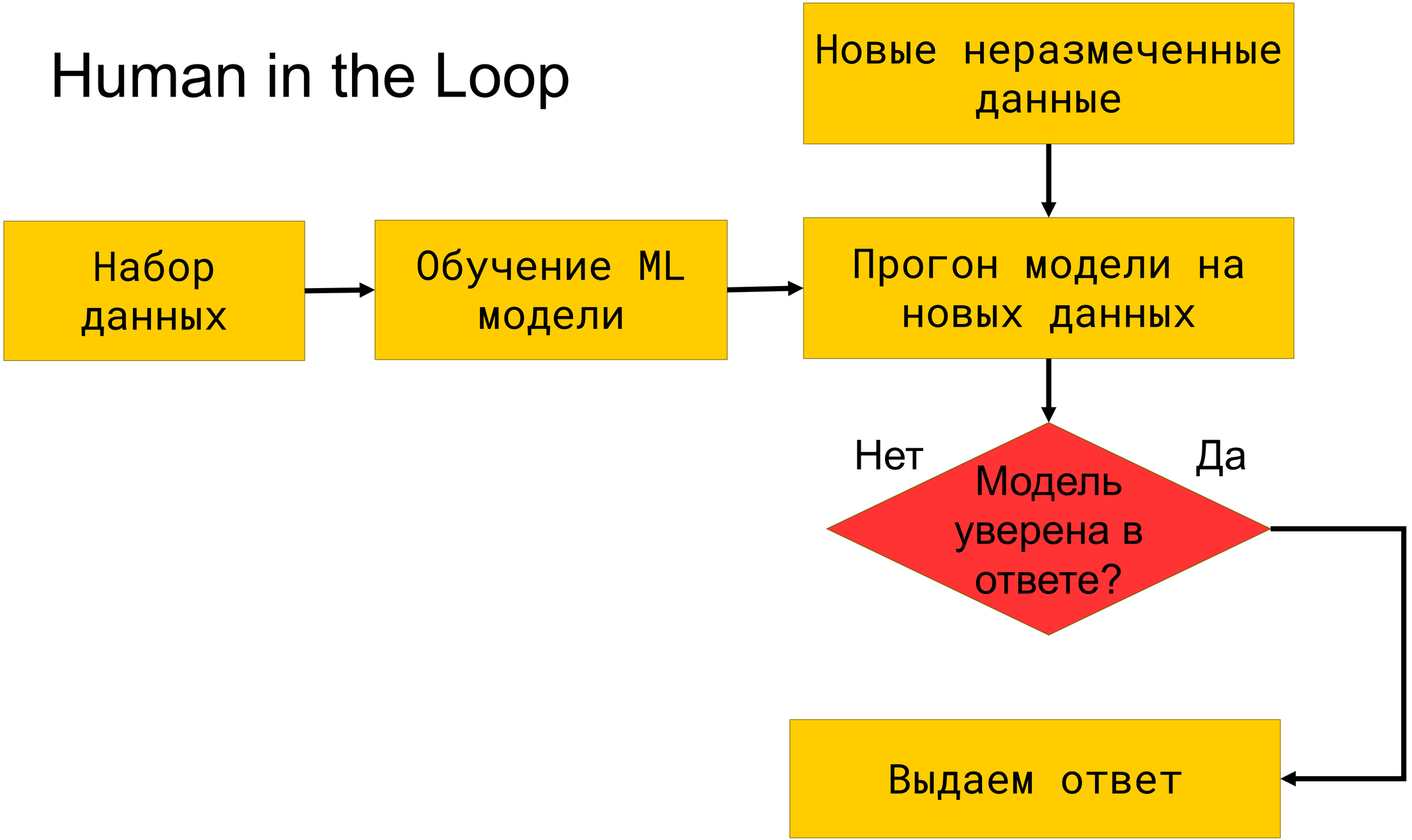


Новые неразмеченные данные

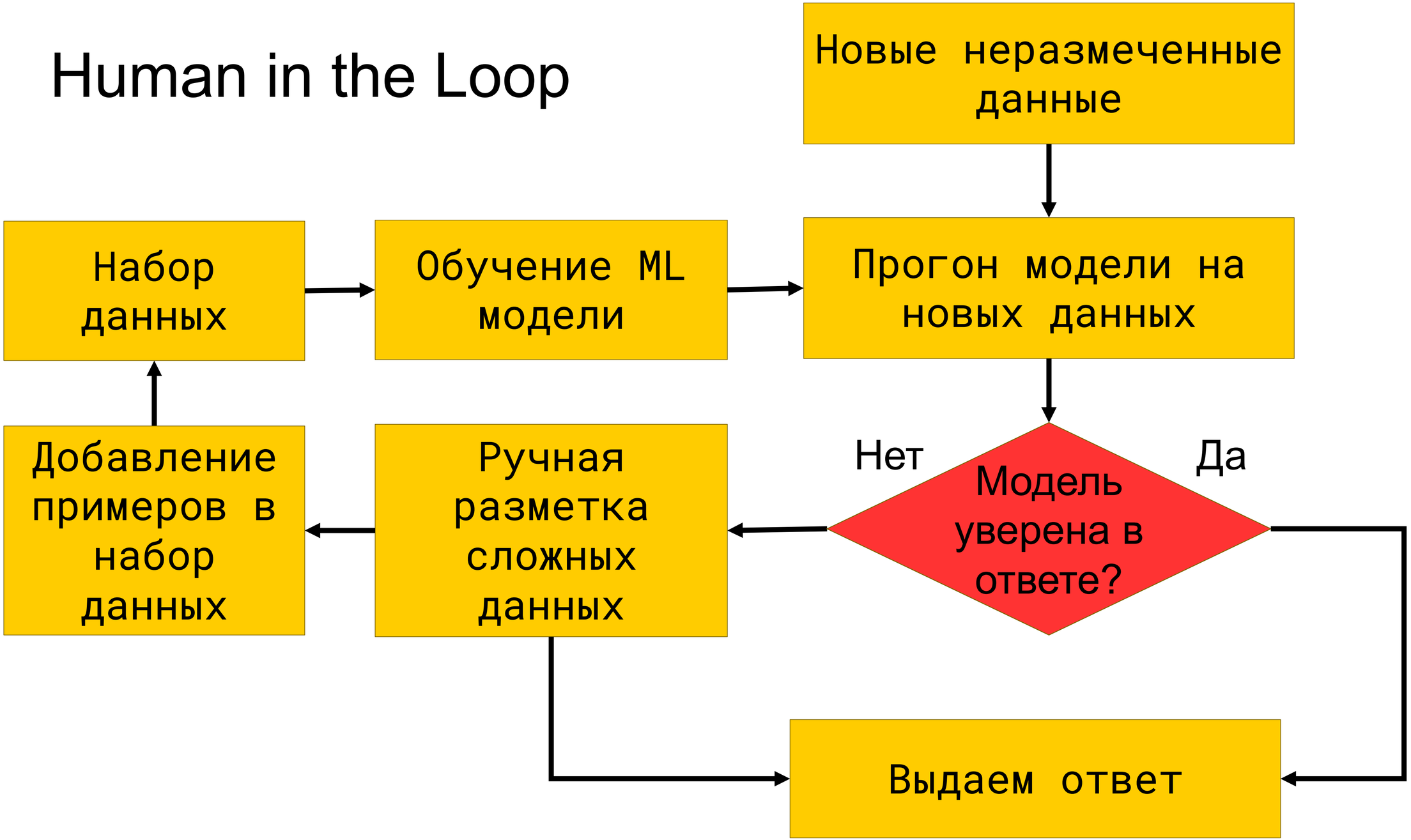
Human in the Loop



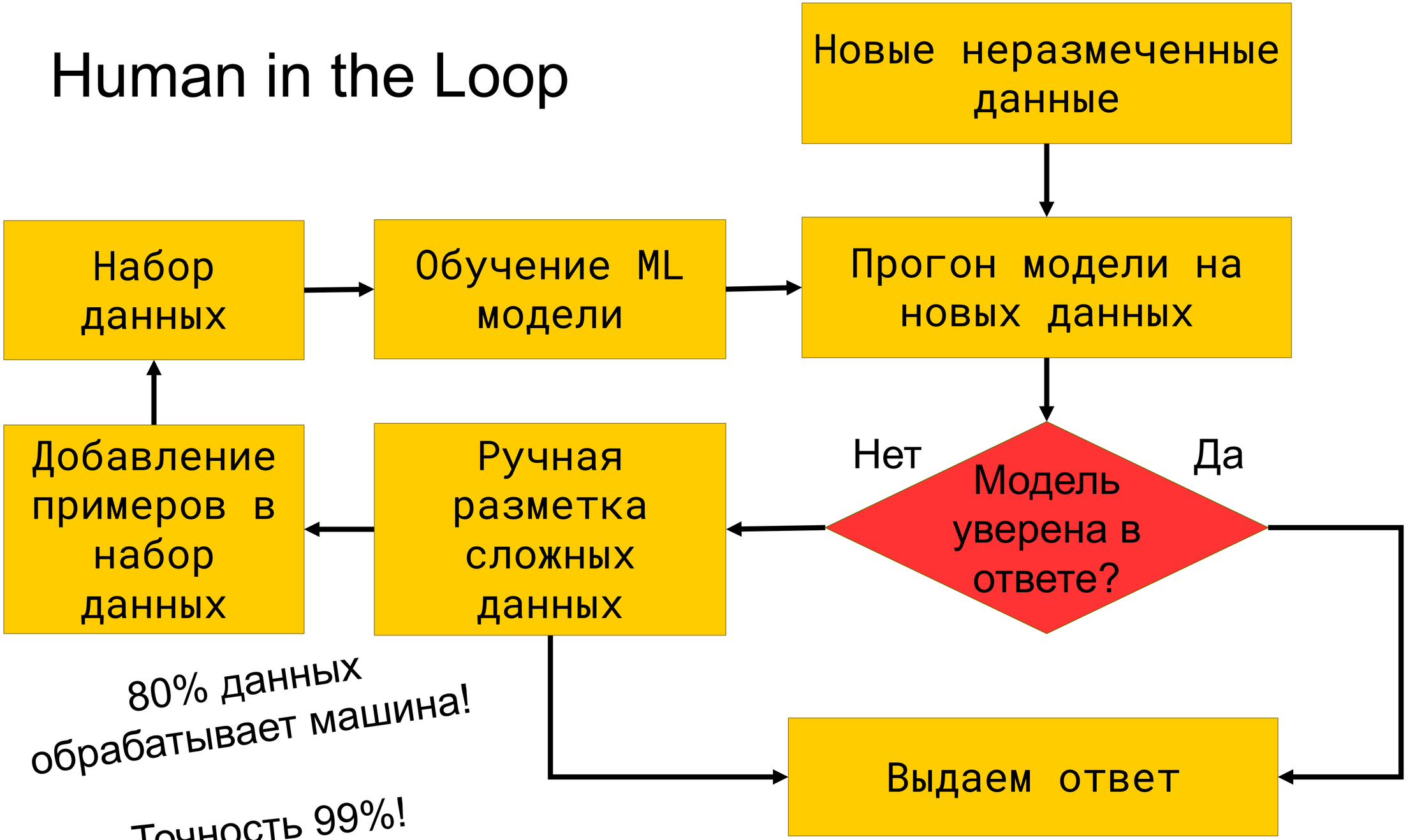
Human in the Loop



Human in the Loop



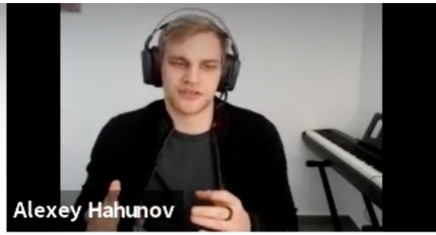
Human in the Loop



80% данных обрабатывает машина!

Точность 99%!

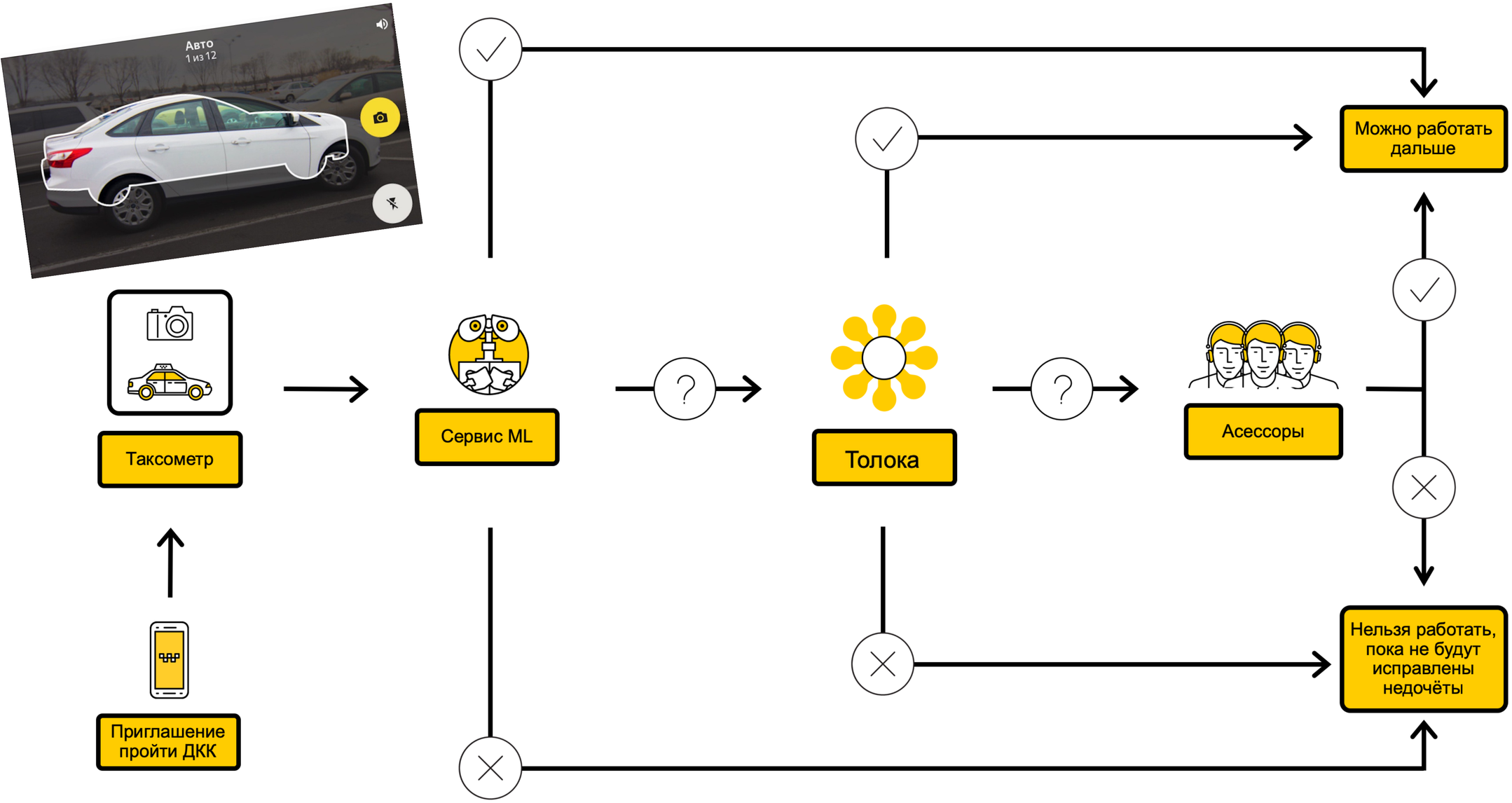
Human in the Loop в Dbrain



Отдельные сервисы для решения любой задачи

https://youtu.be/yyybisXHFc0

Human in the Loop в Яндекс.Такси



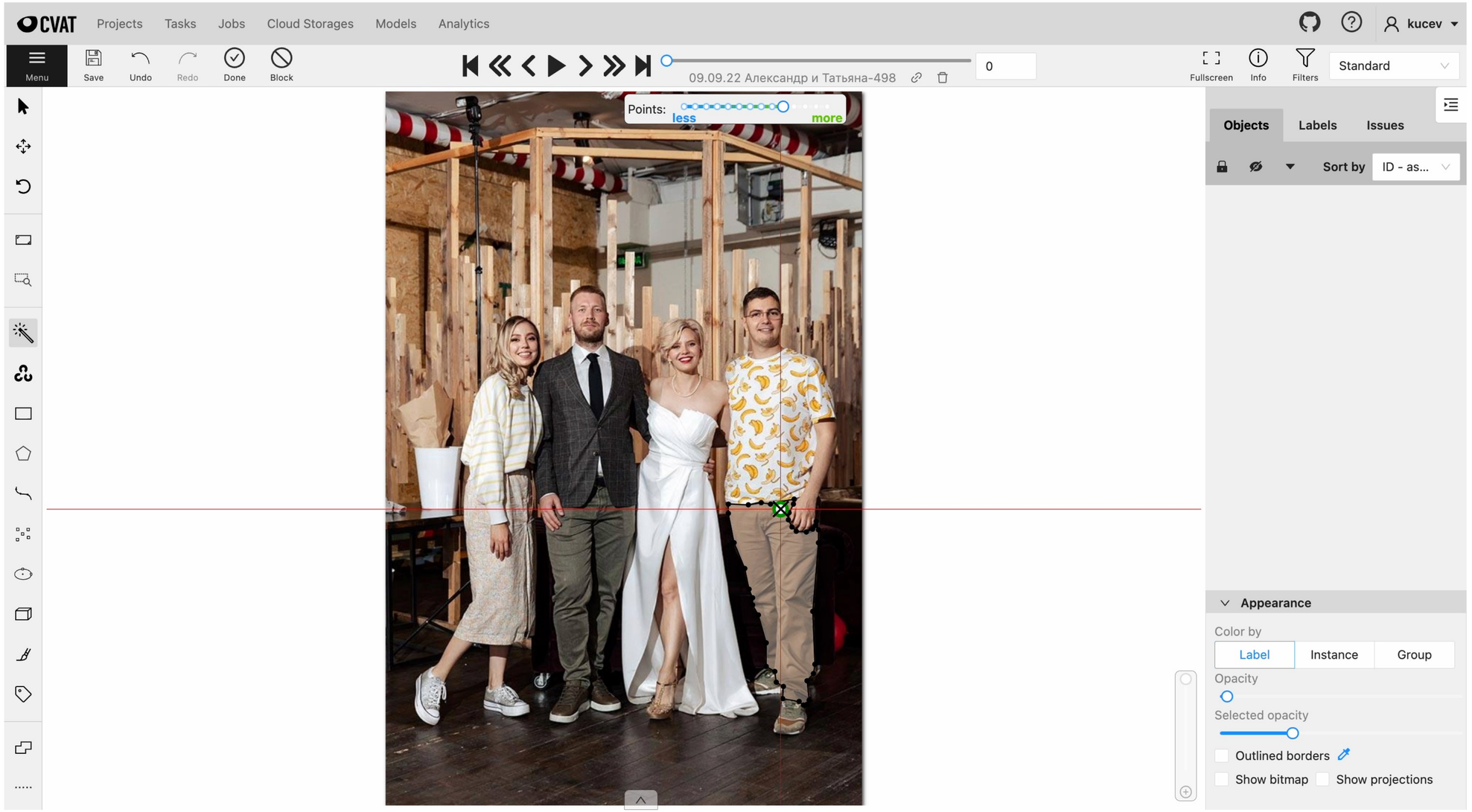
https://habr.com/ru/company/yandex/blog/433386/

Выводы

- Существует множество методов и техник, которые позволяют увеличить скорость и качество разметки
- Данные техники требуют времени на внедрение и настройку, но позволяют сильно снизить стоимость разметки

Тренды в разметке данных

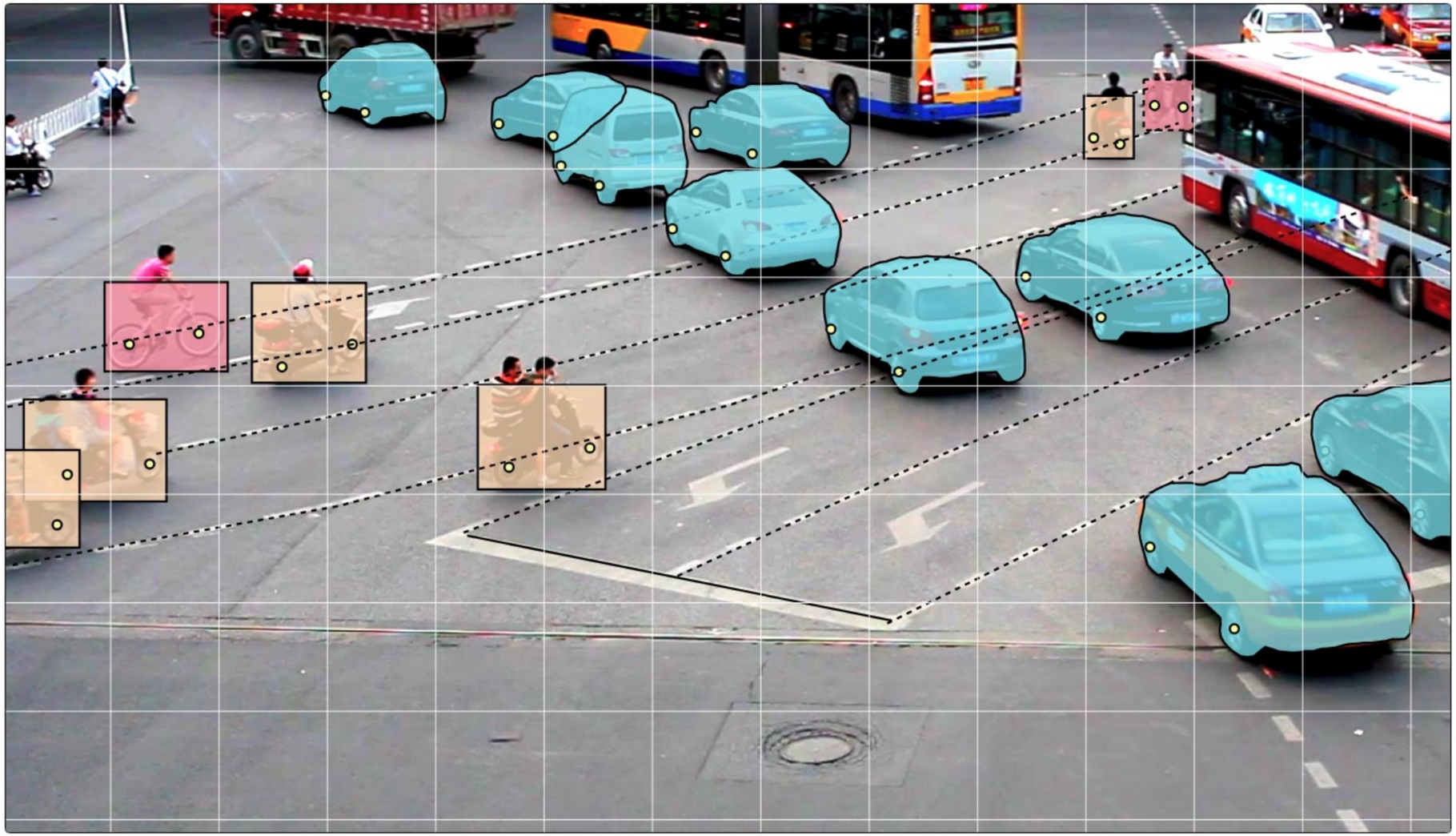
Автоматизация



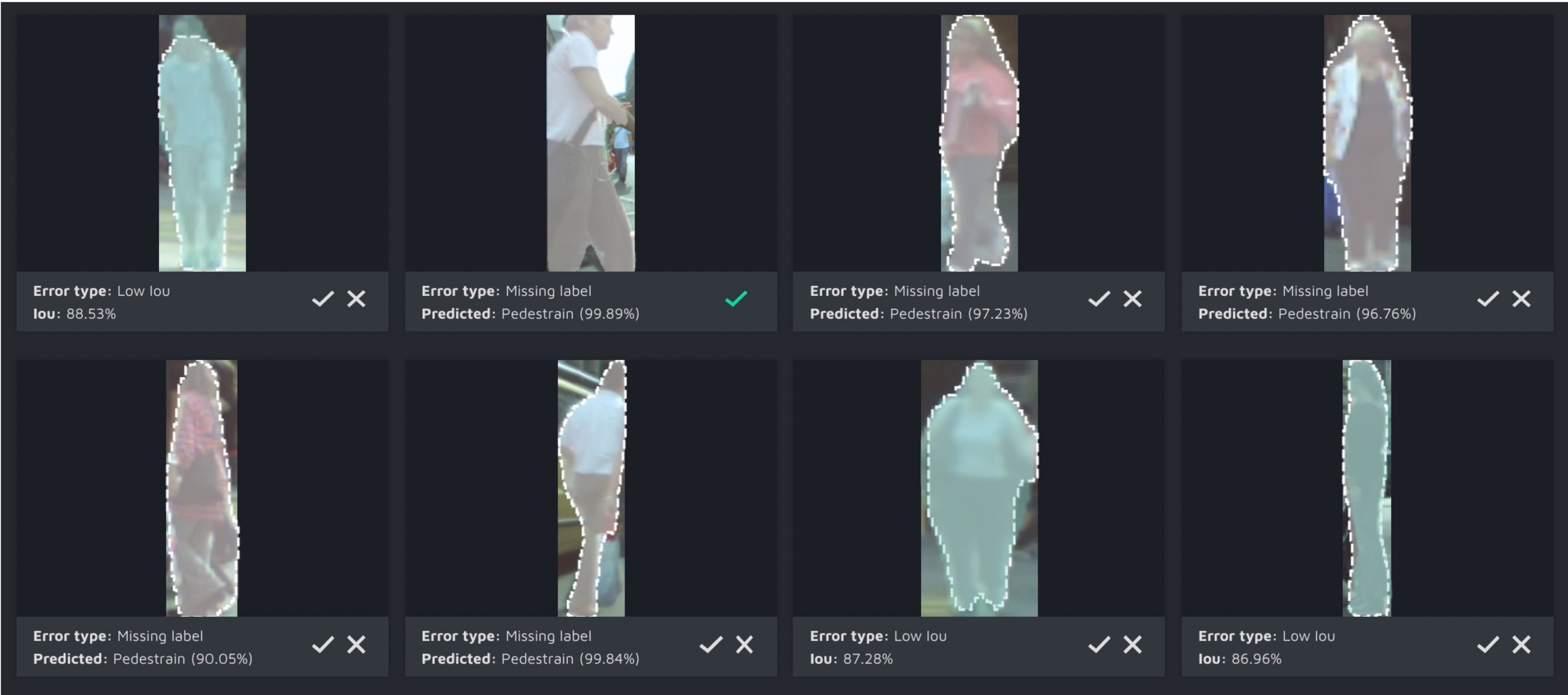
https://youtu.be/bBqersGW0ic

Повышение квалификации разметчика

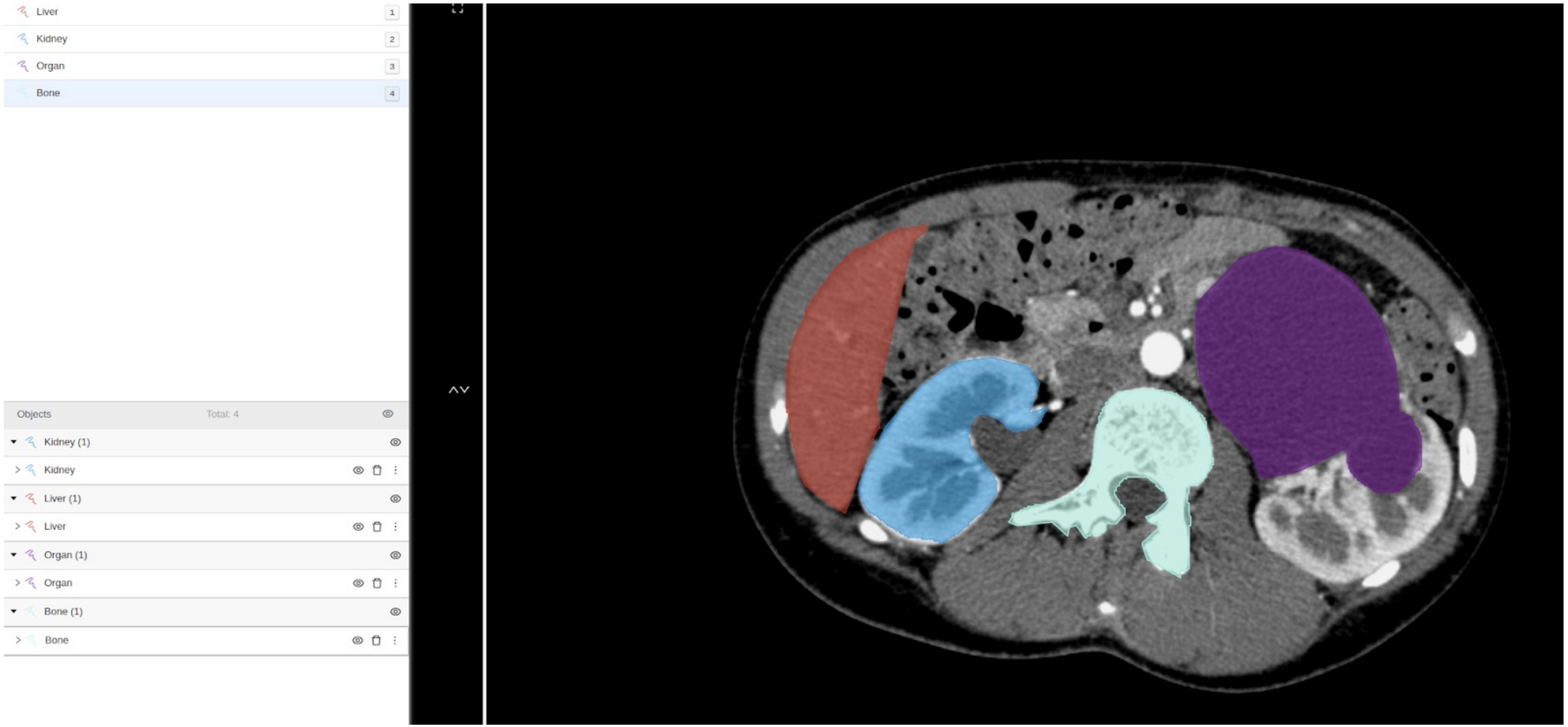
РАЗМЕТЧИК ДАННЫХ: КАК СТАТЬ СПЕЦИАЛИСТОМ В РАСТУЩЕЙ ОБЛАСТИ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ



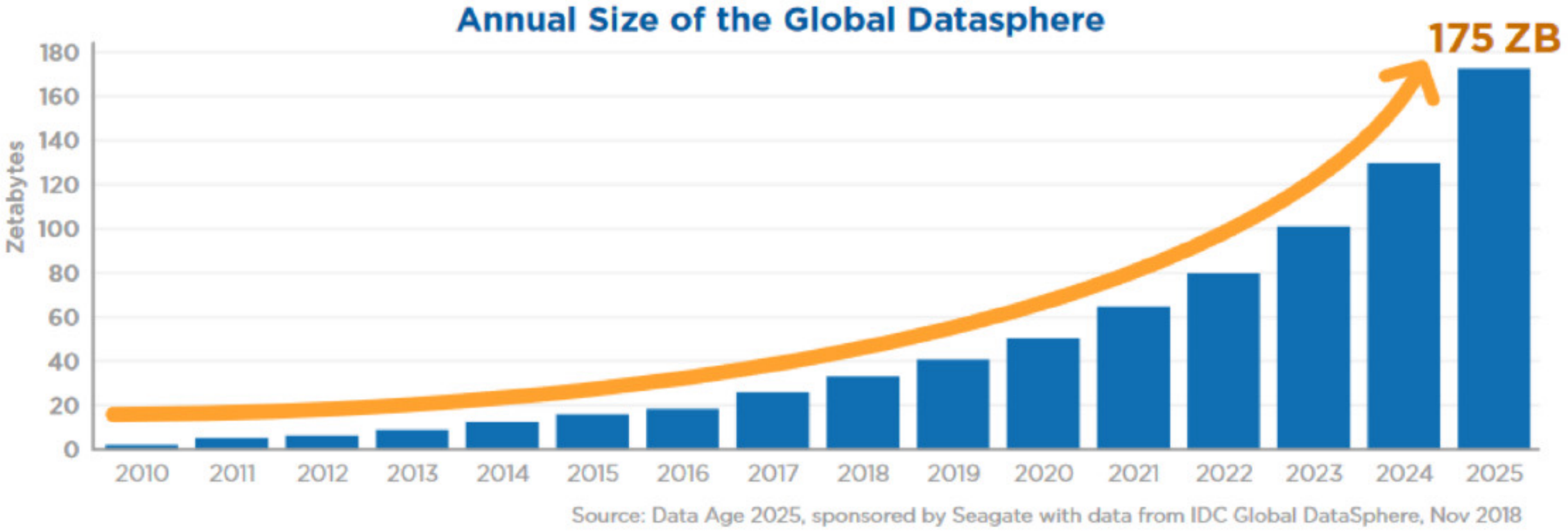
Валидация вместо разметки



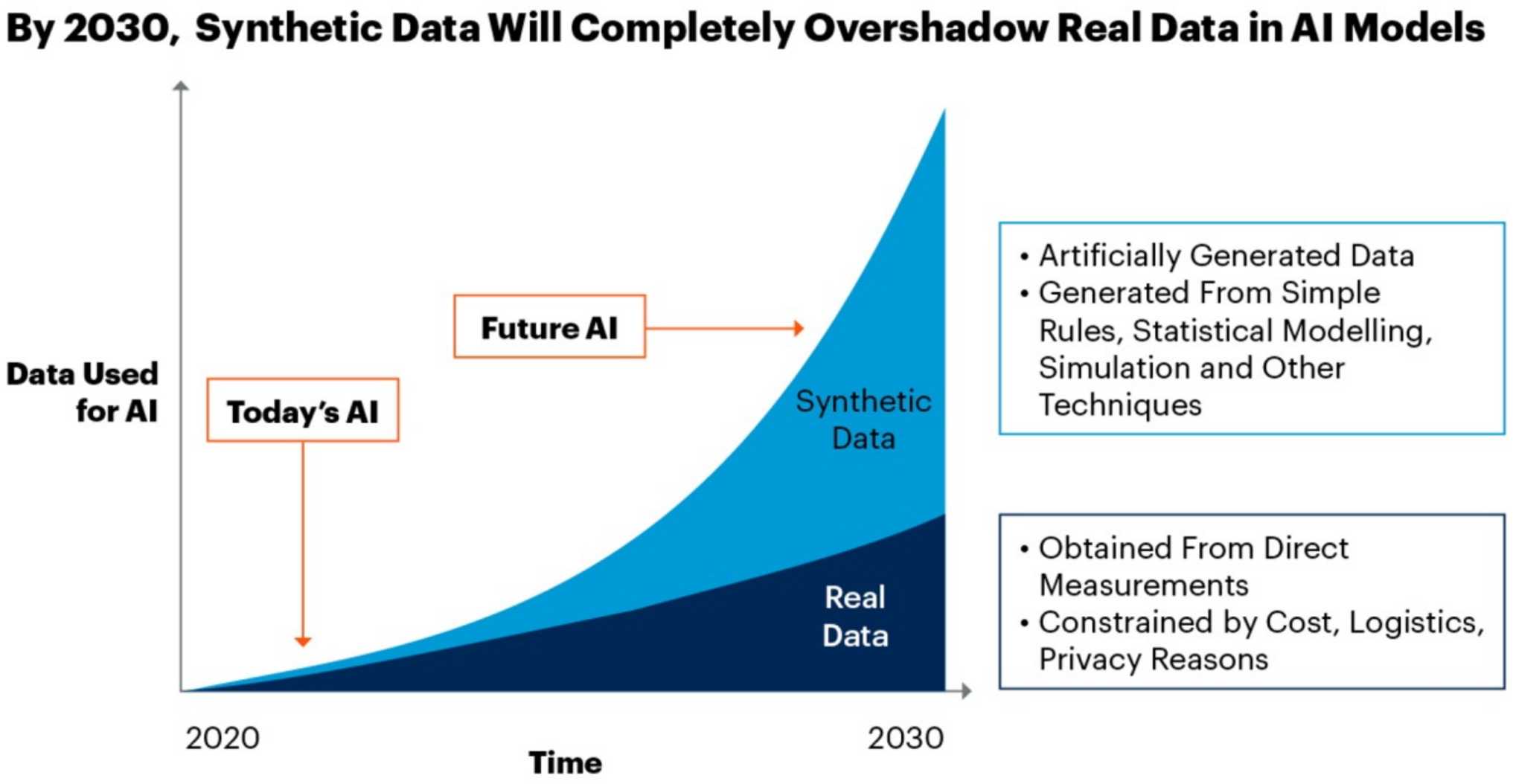
Преобладание разметки узкими специалистами



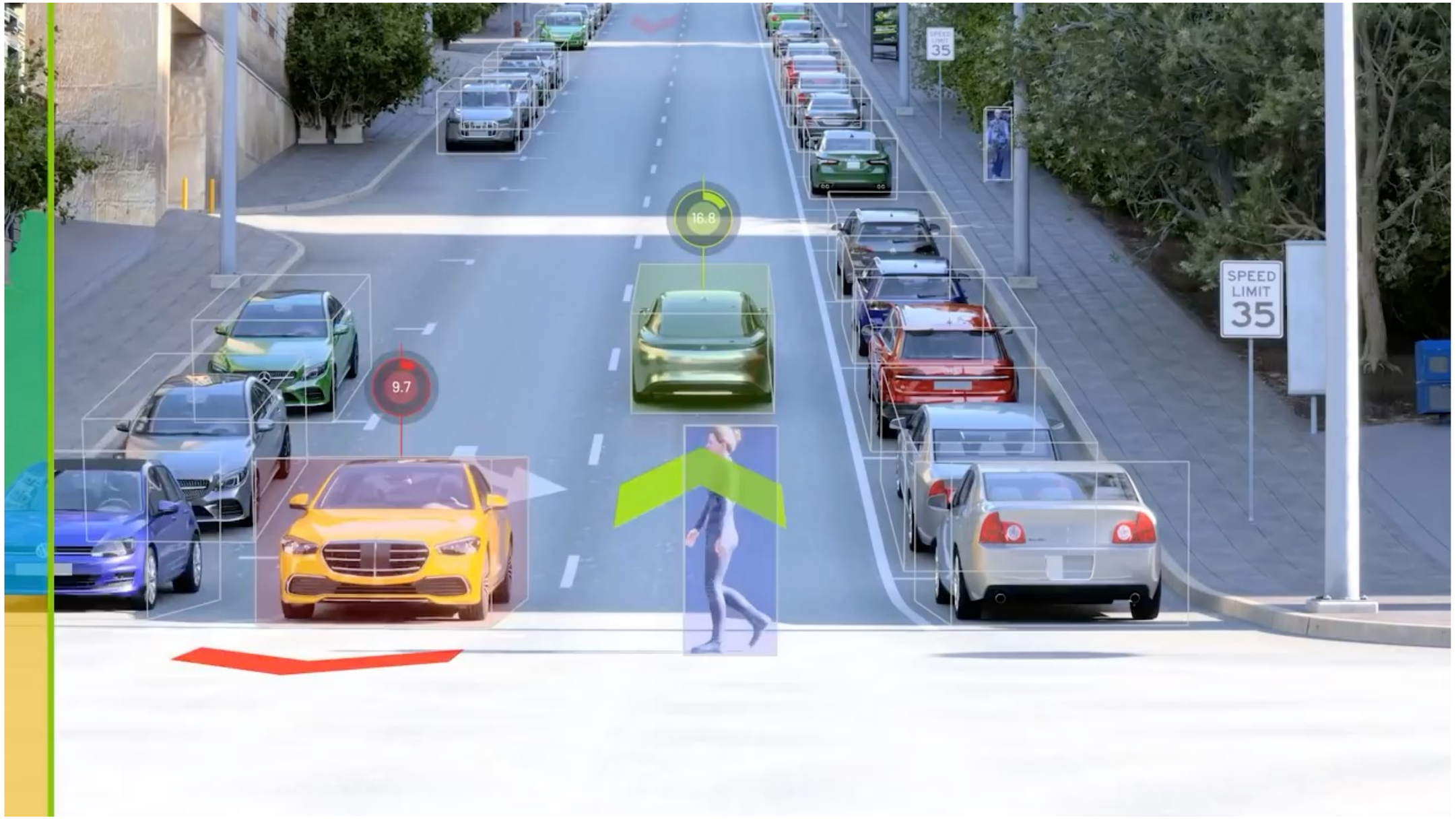
Увеличение объема размечаемых данных



Рендеринг синтетических данных



Рендеринг синтетических данных



Я: Куцев Роман

t.me/roman_kucev

ТГ канал: Рома ❤️ Толоку

t.me/roma_toloka