

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



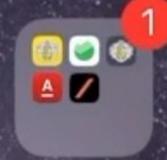
Calendar



Camera



Messages



Bank



App Store



GoogleCalendar



Photos



Clock



Settings



Zoom



Navi



Google Maps



Transport



Yandex.Taxi



Trello



Music



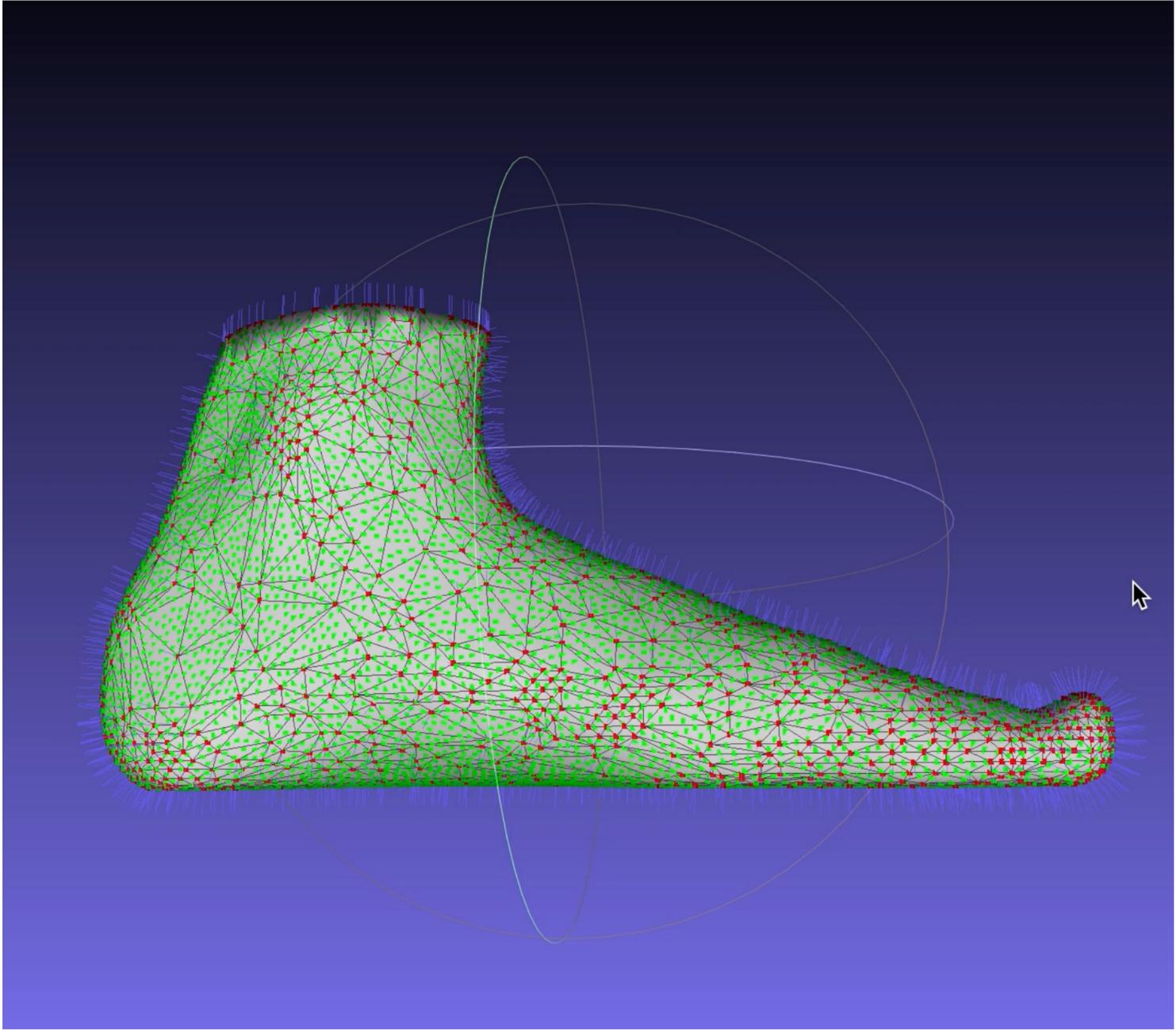
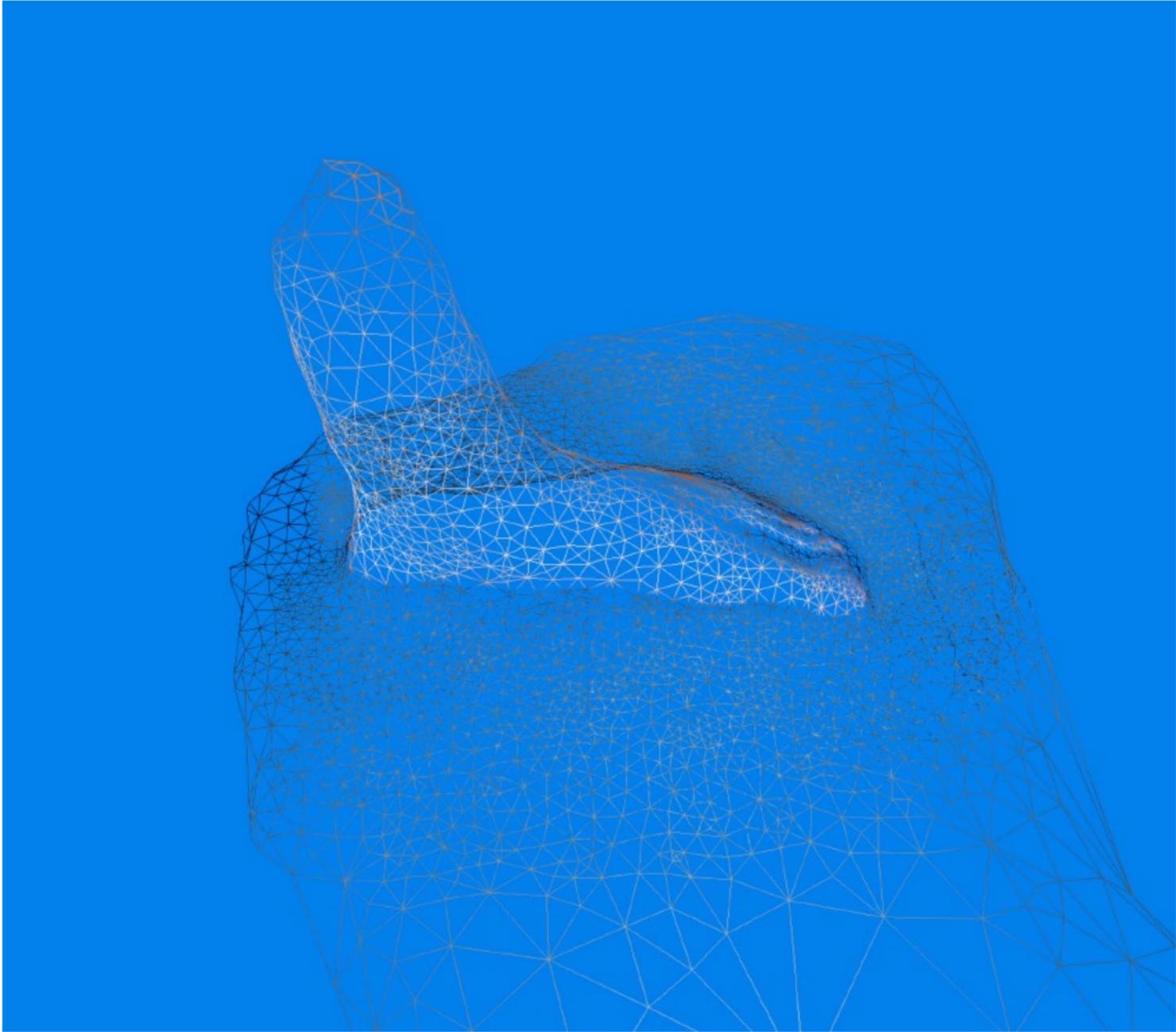
Neatsy



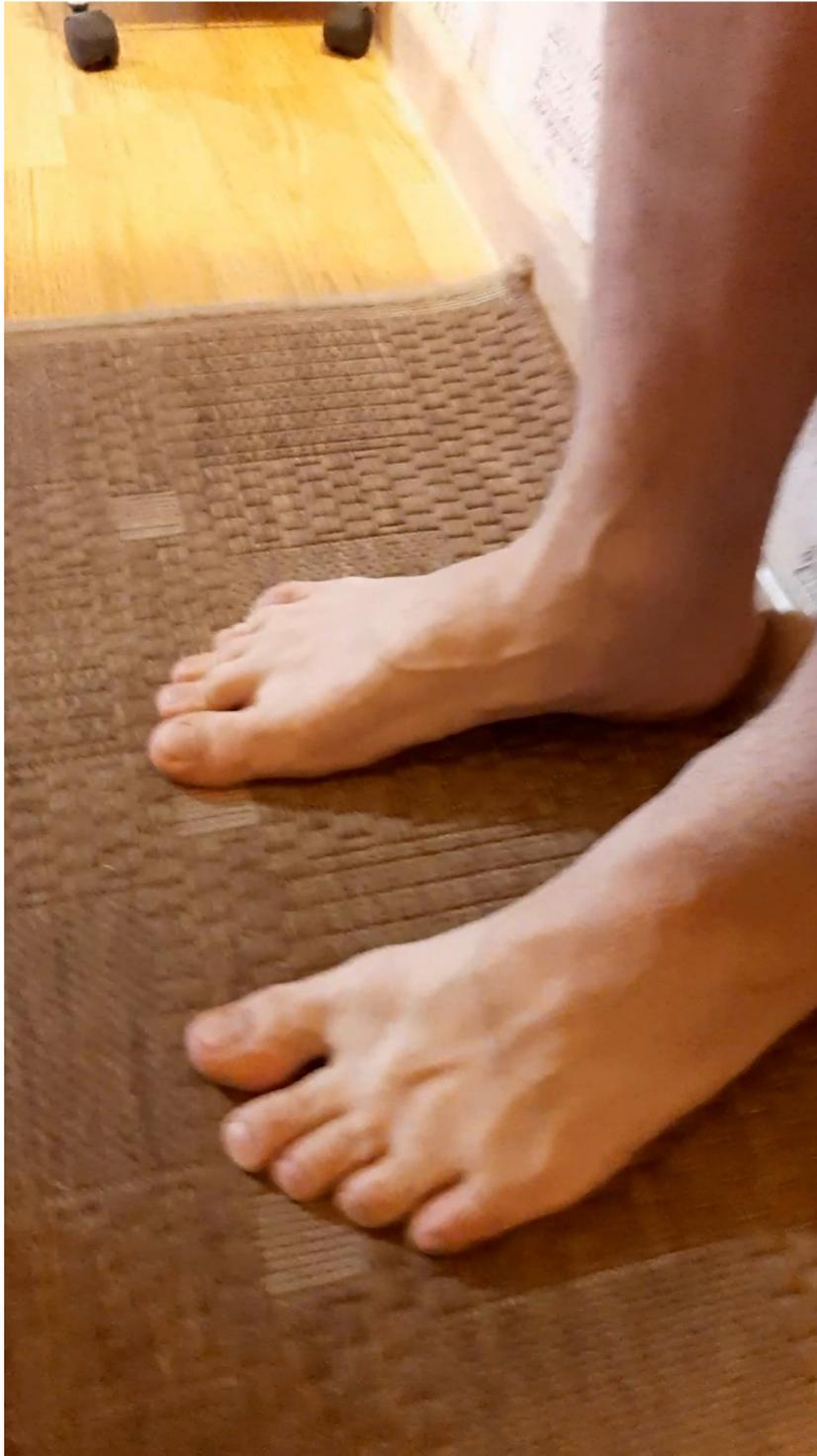
Dropbox



Neatsy



Neatsy



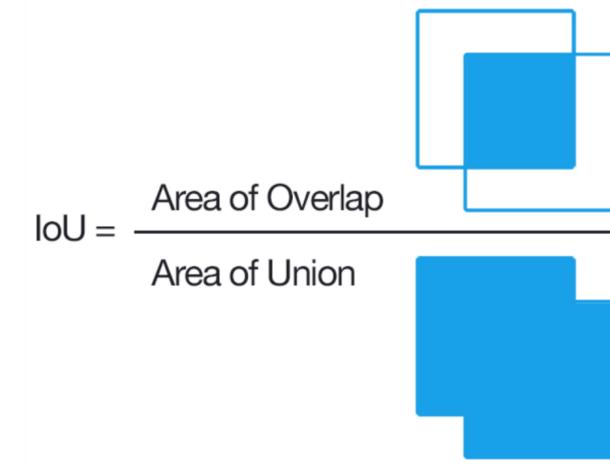
Какую выбрать метрику?
Какой порог метрики нам нужен?

Цель

Получить Intersection over Union (IoU) > 0.95

Вопросы

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

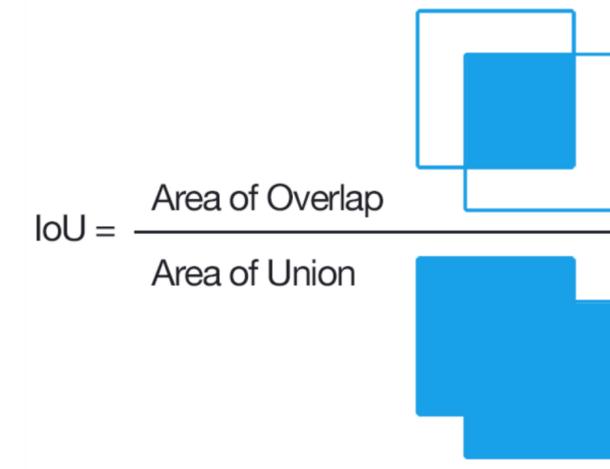


Цель

Получить Intersection over Union (IoU) > 0.95

Вопросы

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

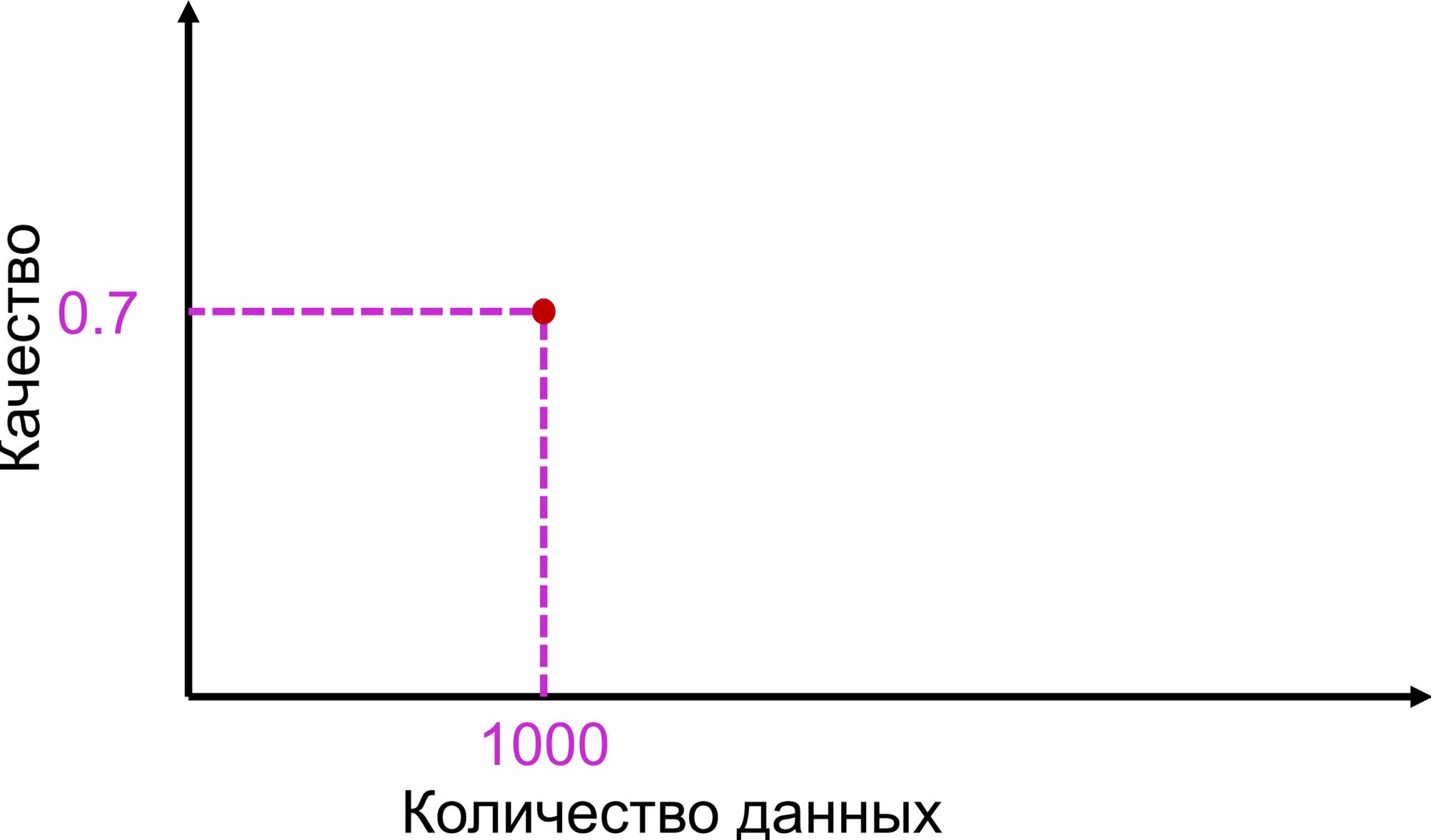


Шаг 1. Разметим 1000 изображений

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

Шаг 1. Разметим 1000 изображений

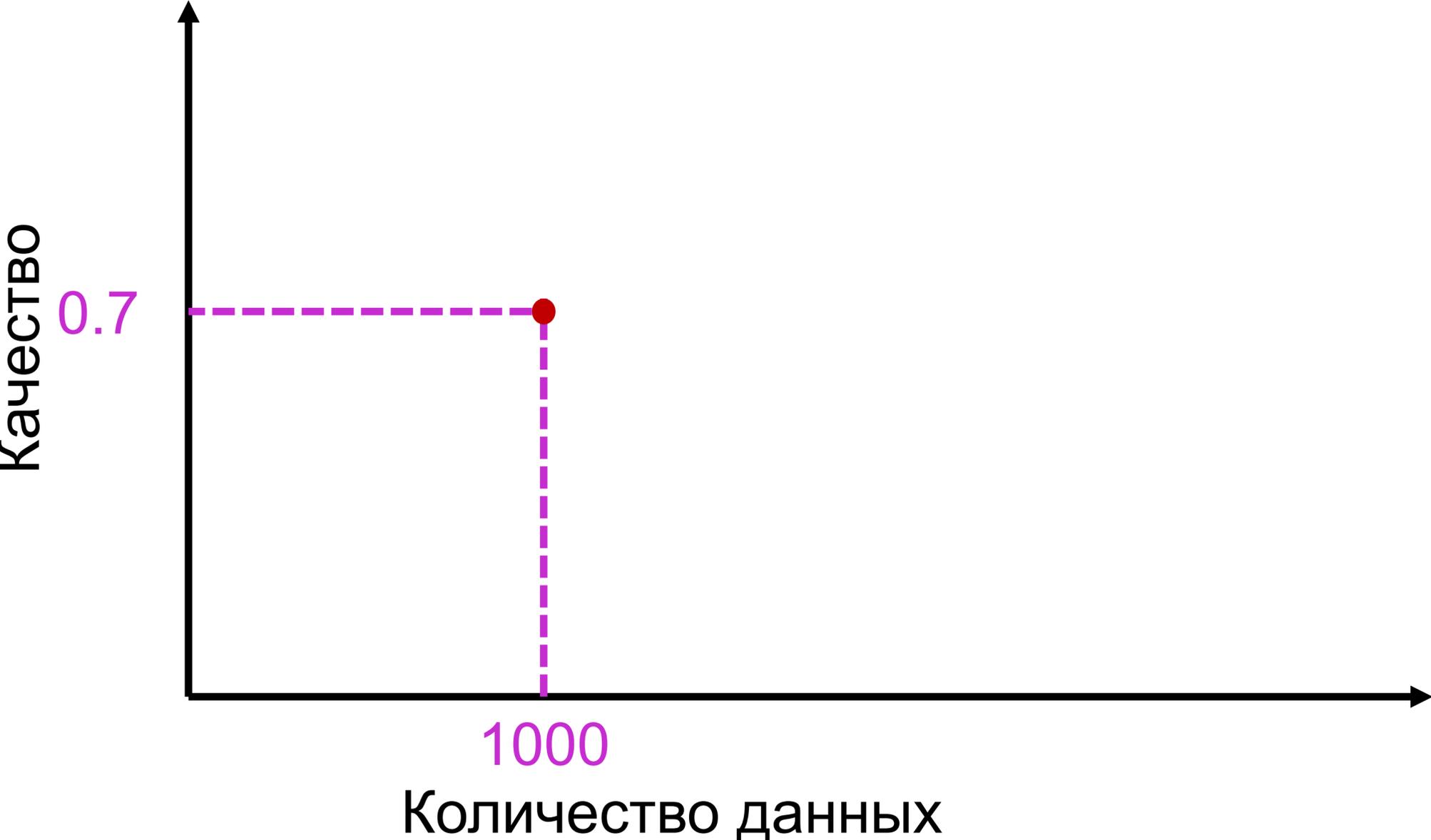
- Почему 1000? Эмпирически, по опыту в других задачах.



Шаг 1. Разметим 1000 изображений

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

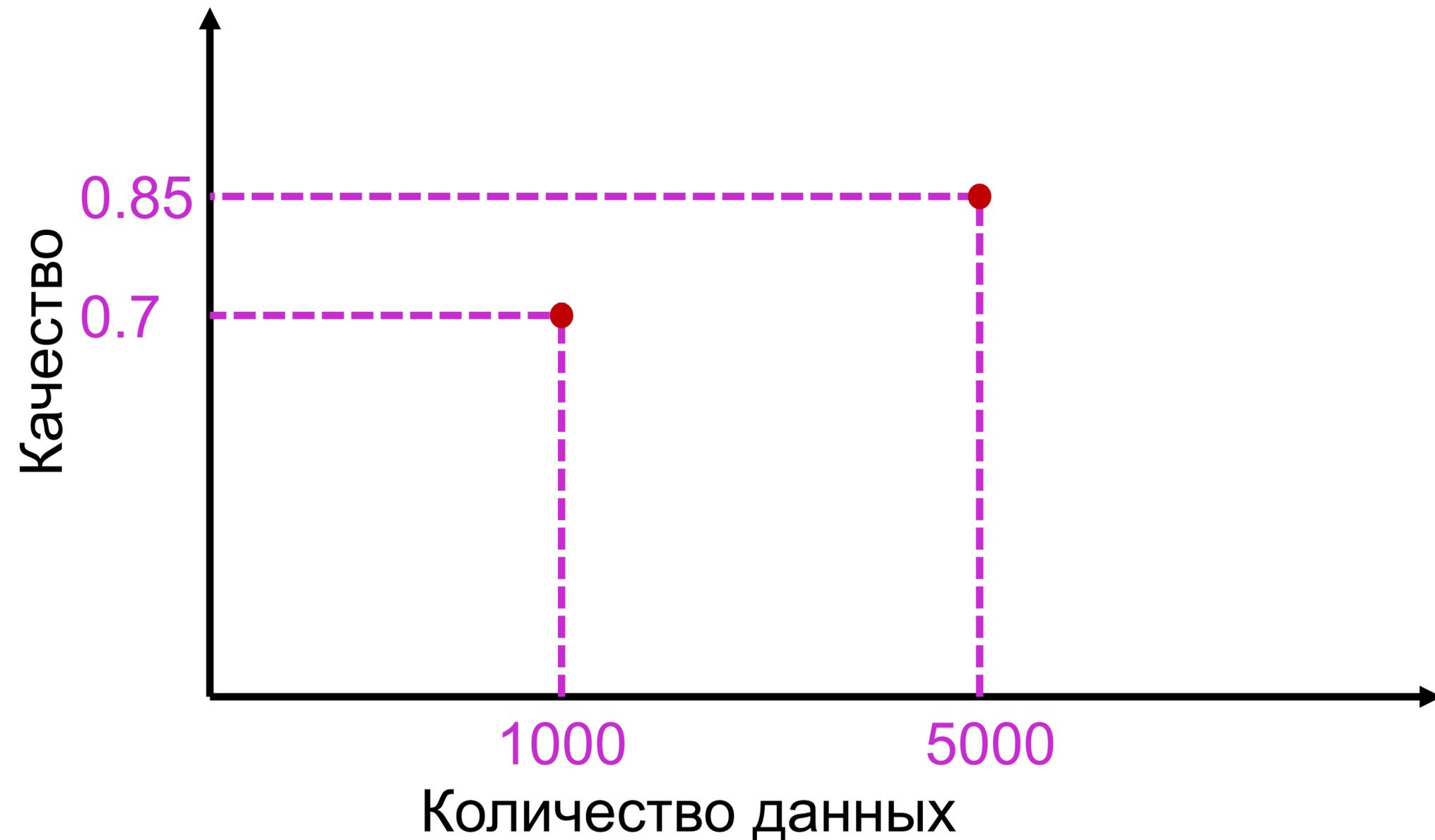
Шаг 2. Разметим 5000 изображений



Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

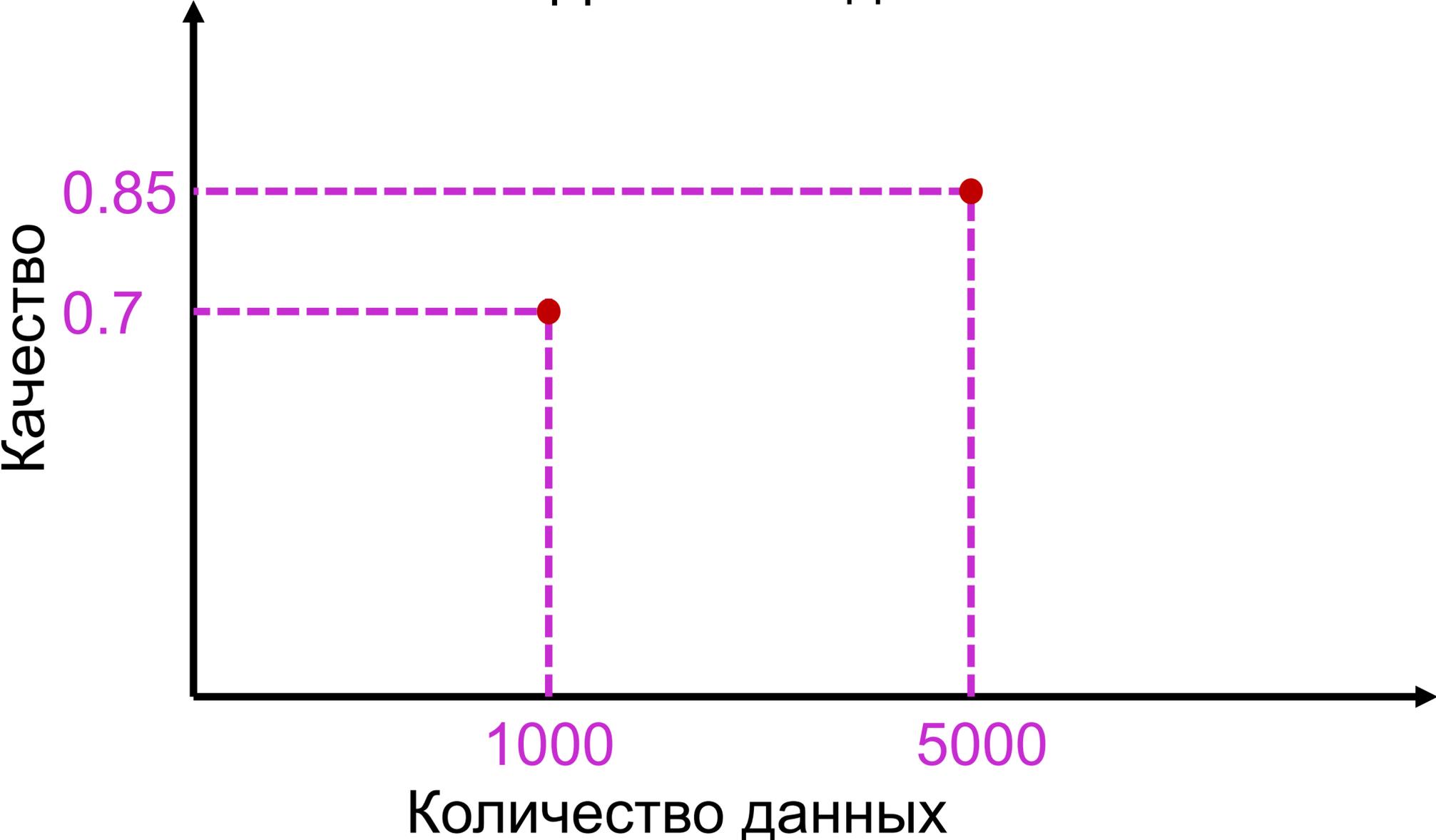


Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

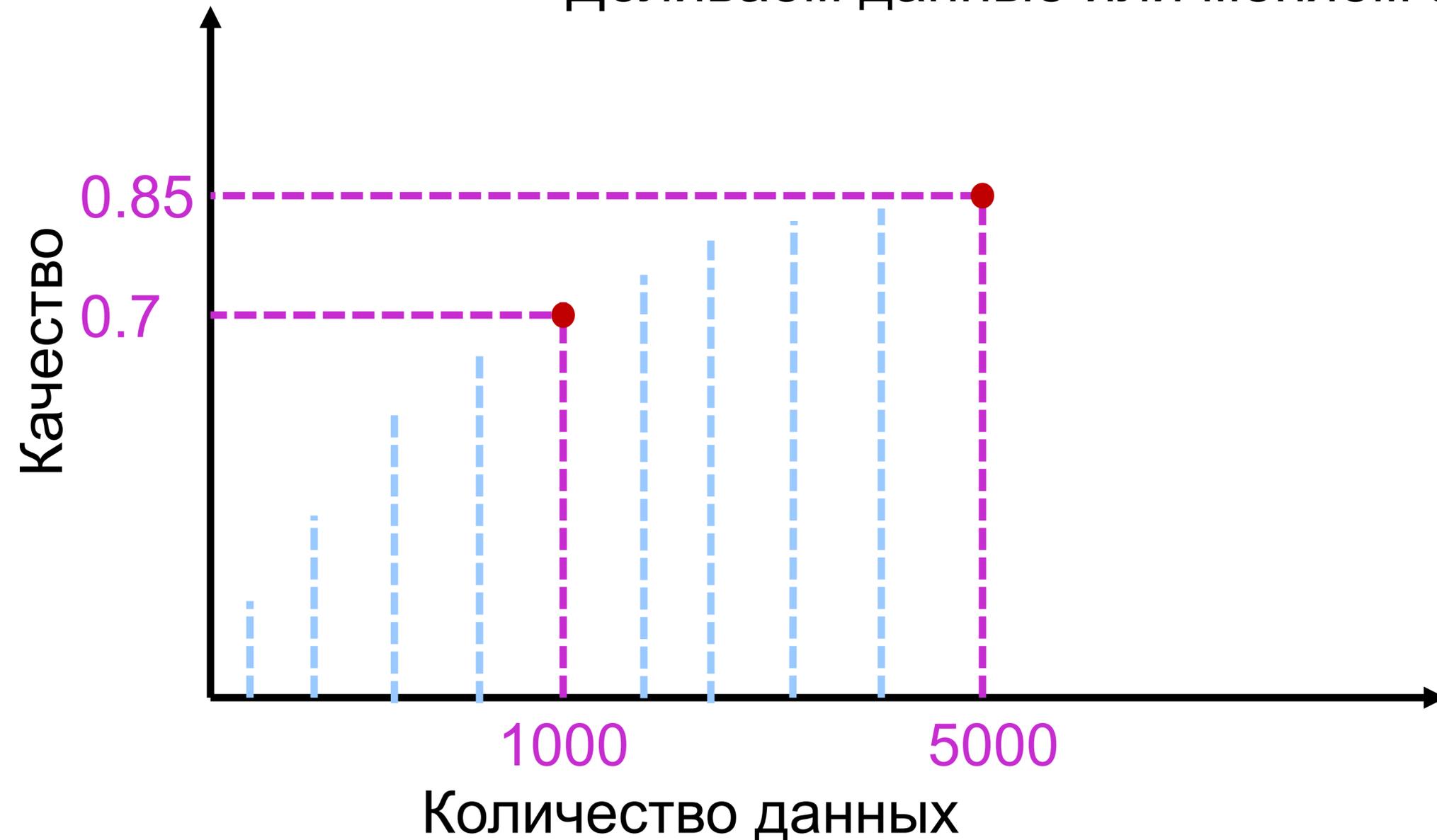


Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

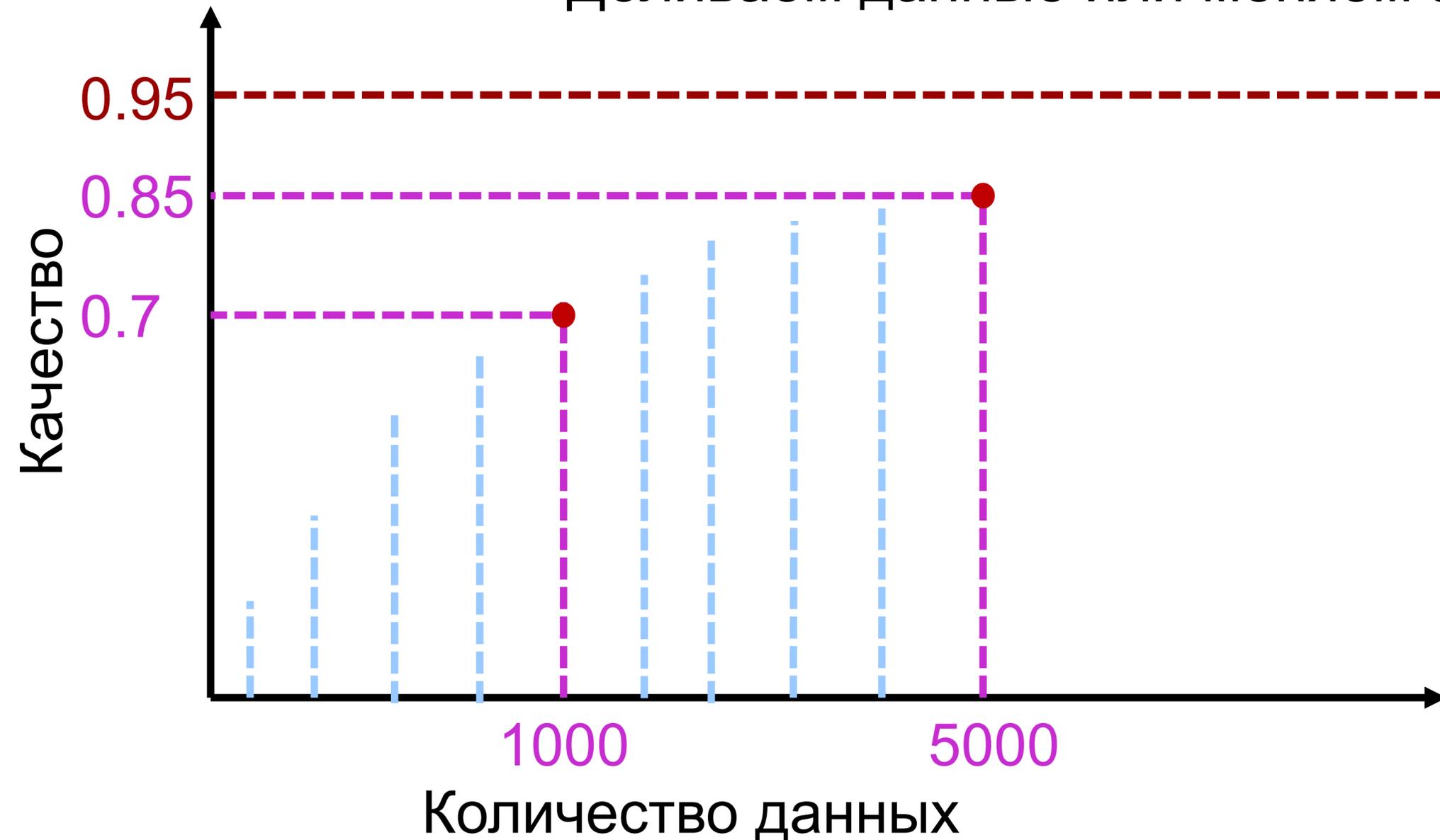


Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

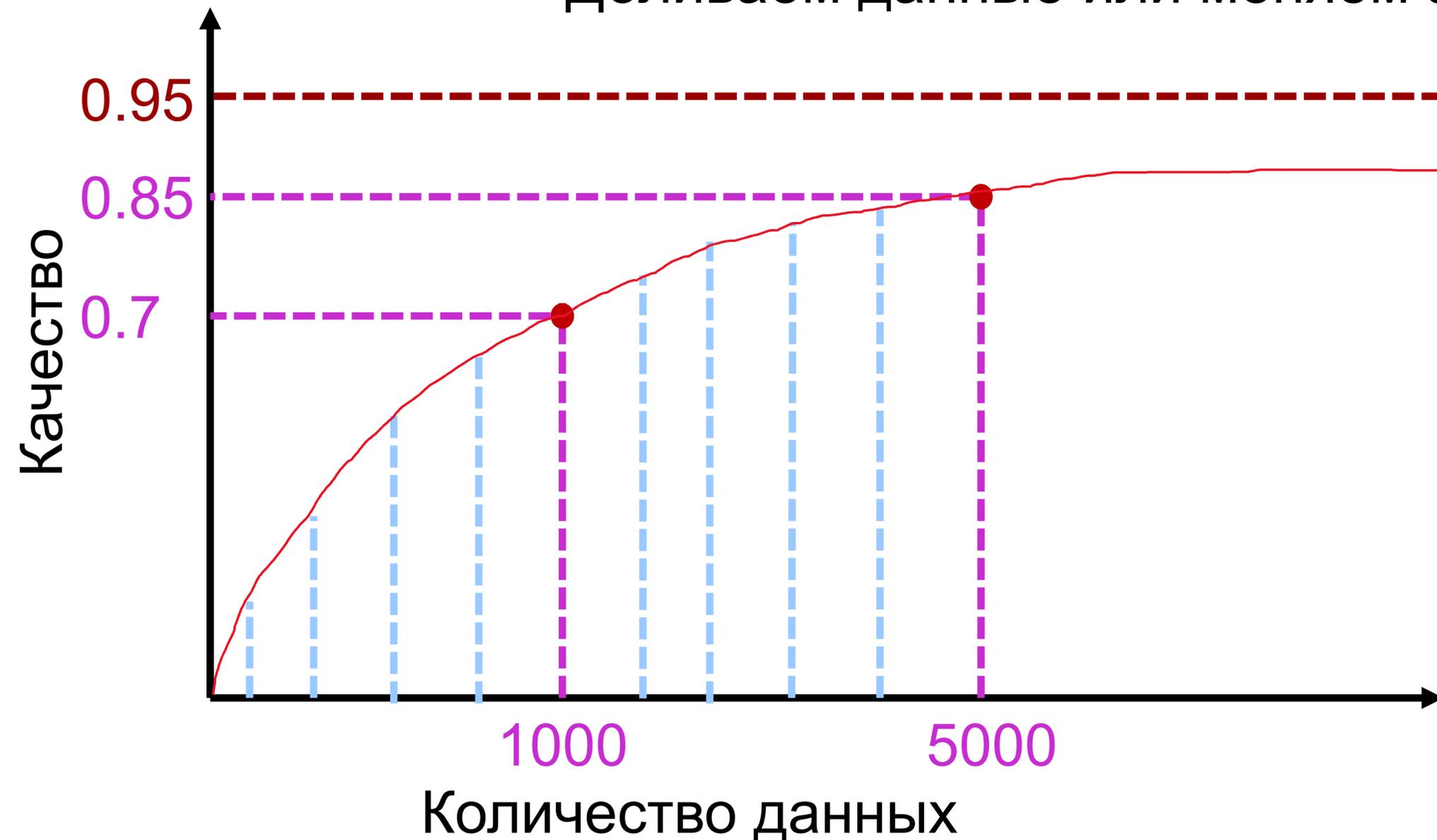


Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?



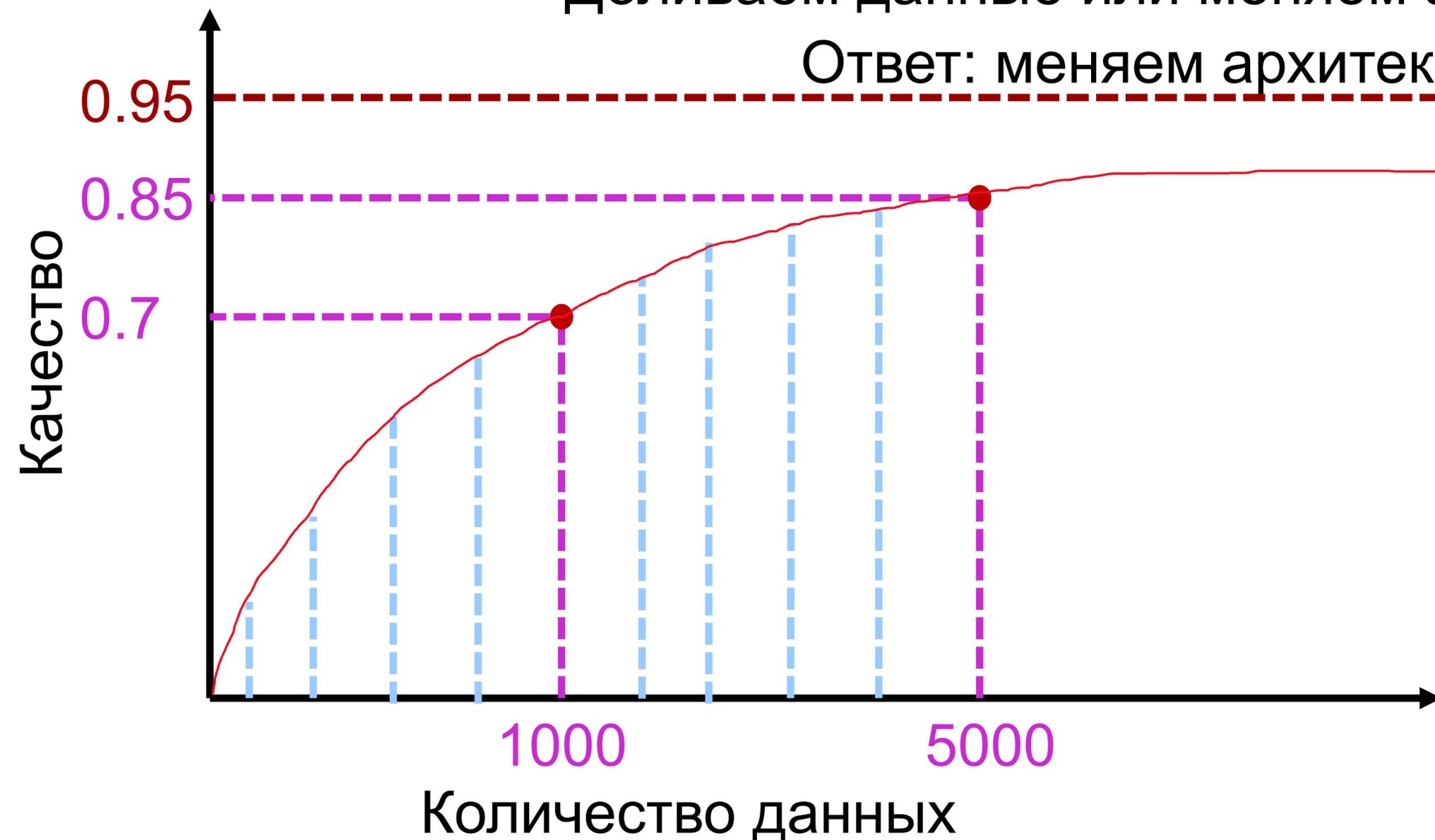
Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

Ответ: меняем архитектуру



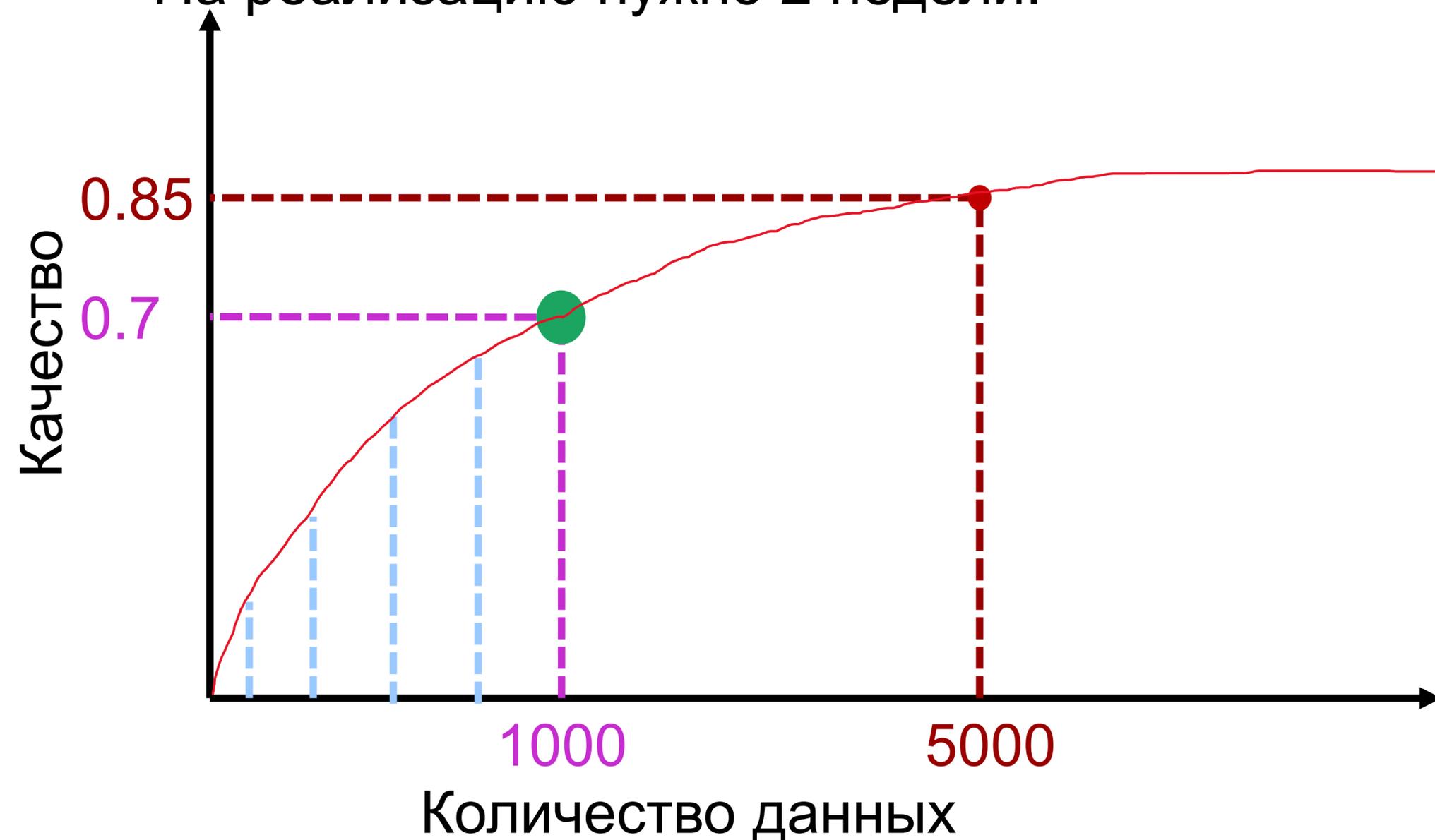
Кейс 2:

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.

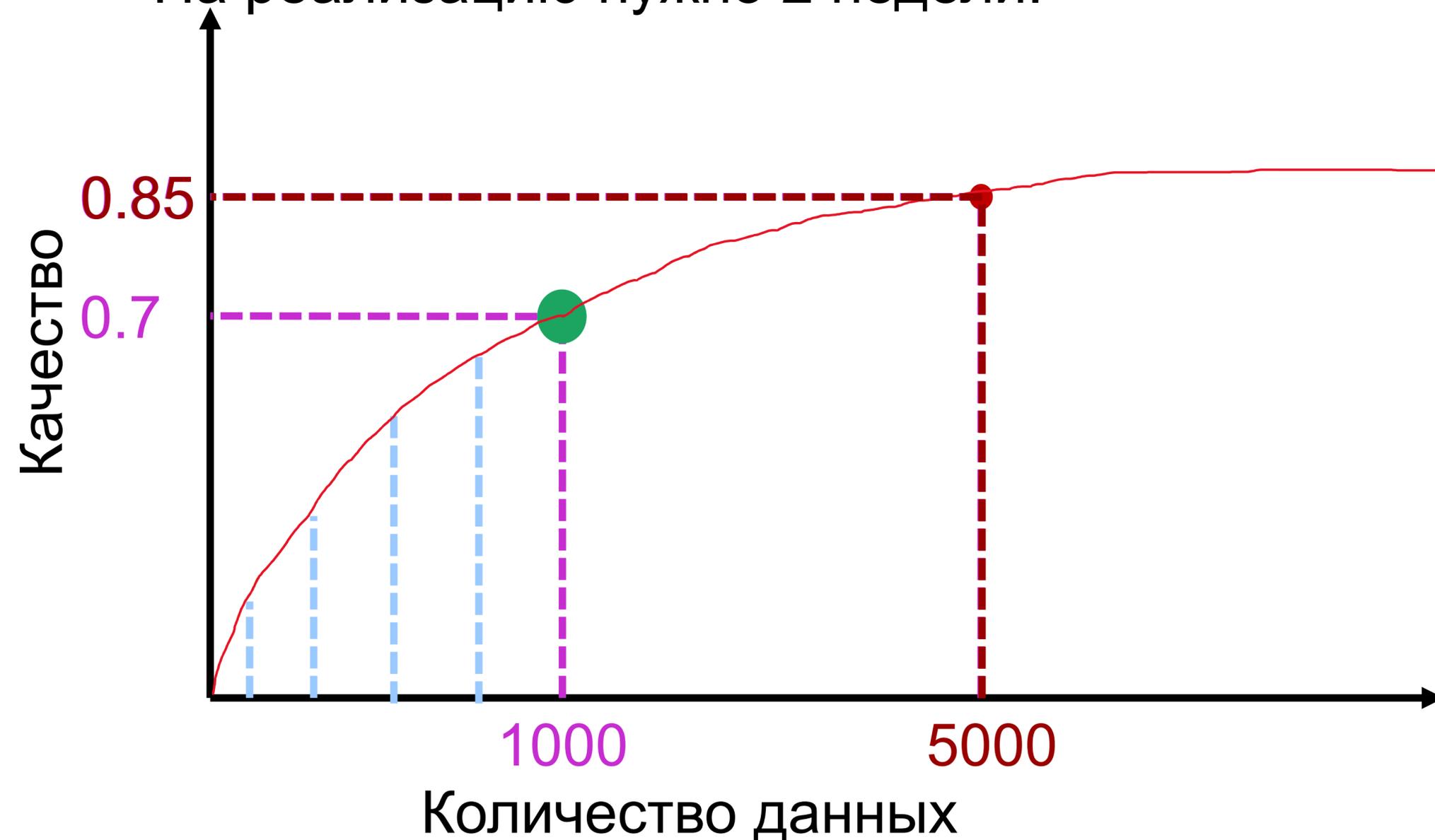
Вася прочитал статью о новом алгоритме.

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.

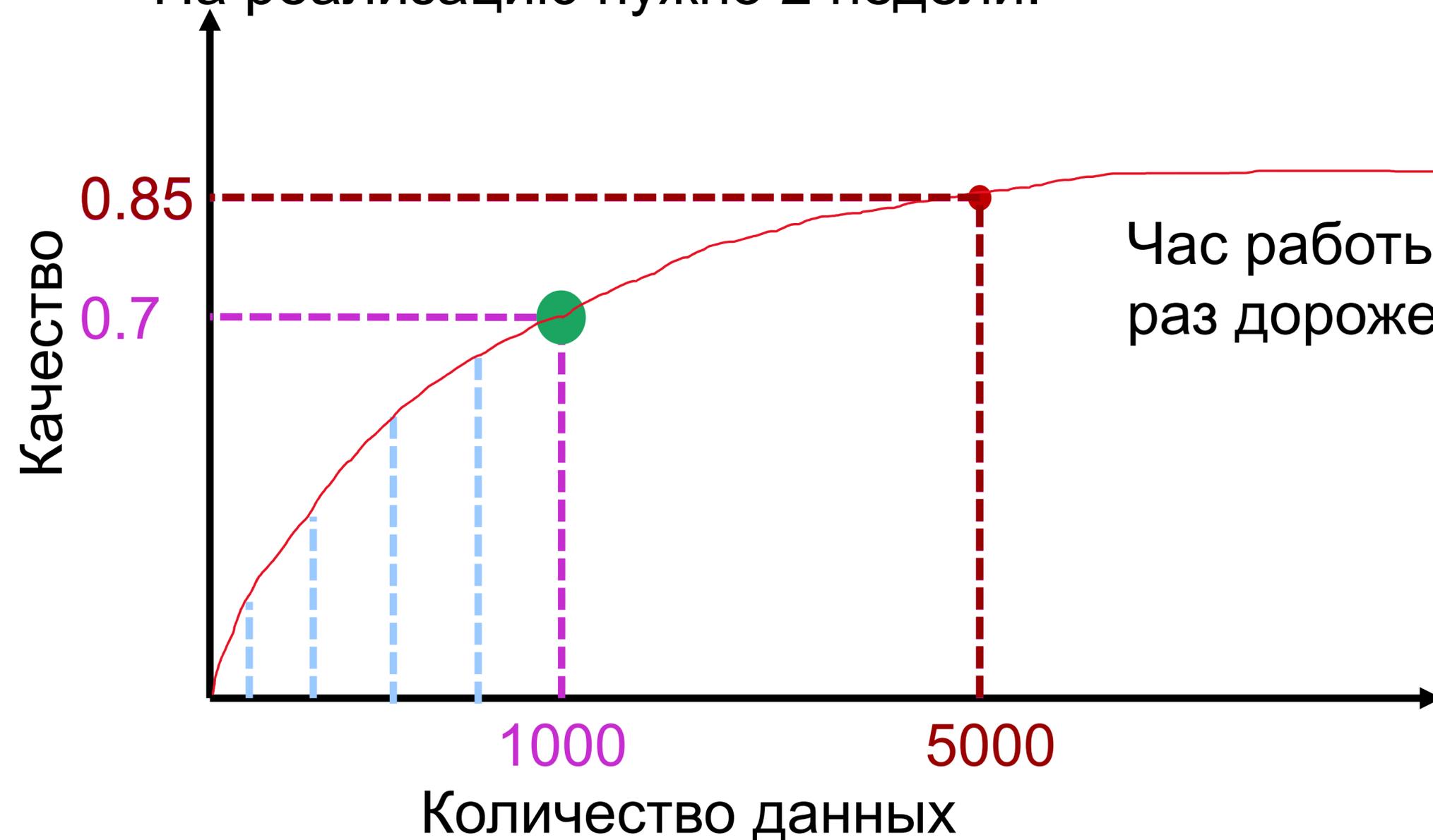
Уверен, что получится достичь 0.85.

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 ассессора

Кейс 2:

Сейчас имеем 0.7, нужно 0.85.

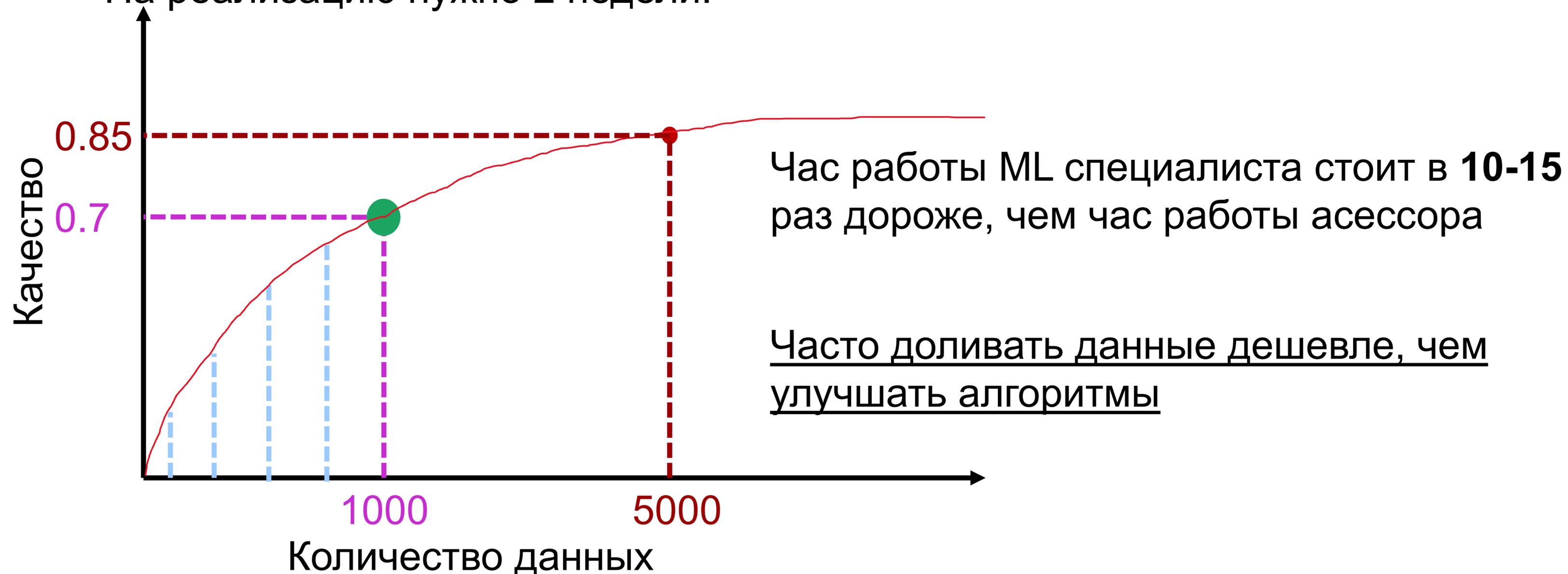
Вася прочитал статью о новом алгоритме.

Уверен, что получится достичь 0.85.

На реализацию нужно 2 недели.

Варианты:

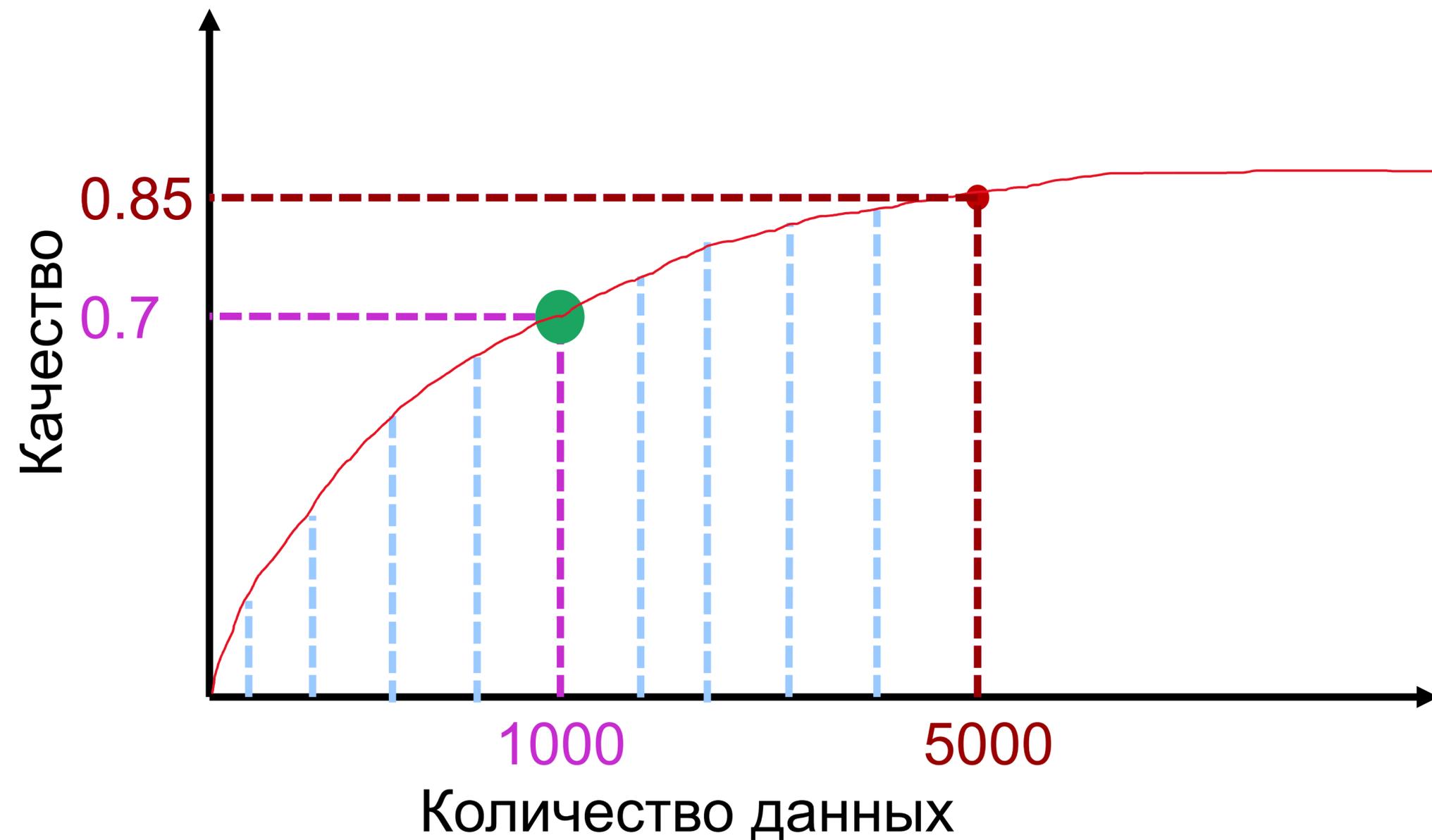
- 1) Вася реализует новый алгоритм
- 2) Размечаем больше данных



Кейс 3:

Сейчас имеем 0.7, нужно 0.85.

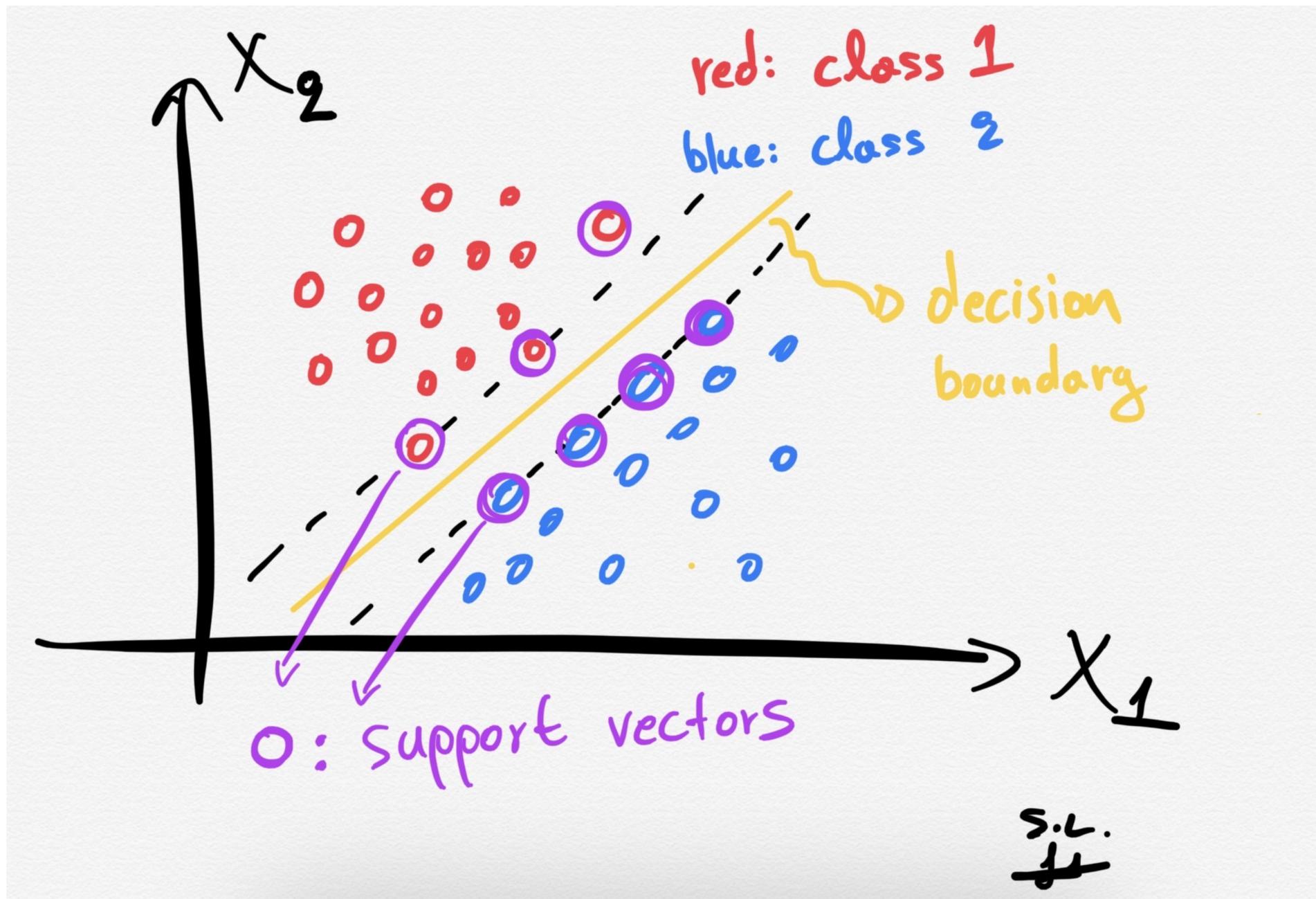
Можем ли мы меньше, чем за 5000 картинок достичь качества 0.85?



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, используя как можно меньше обучающих примеров.

Подробнее тут: <https://youtu.be/M-geBIVIUXY>

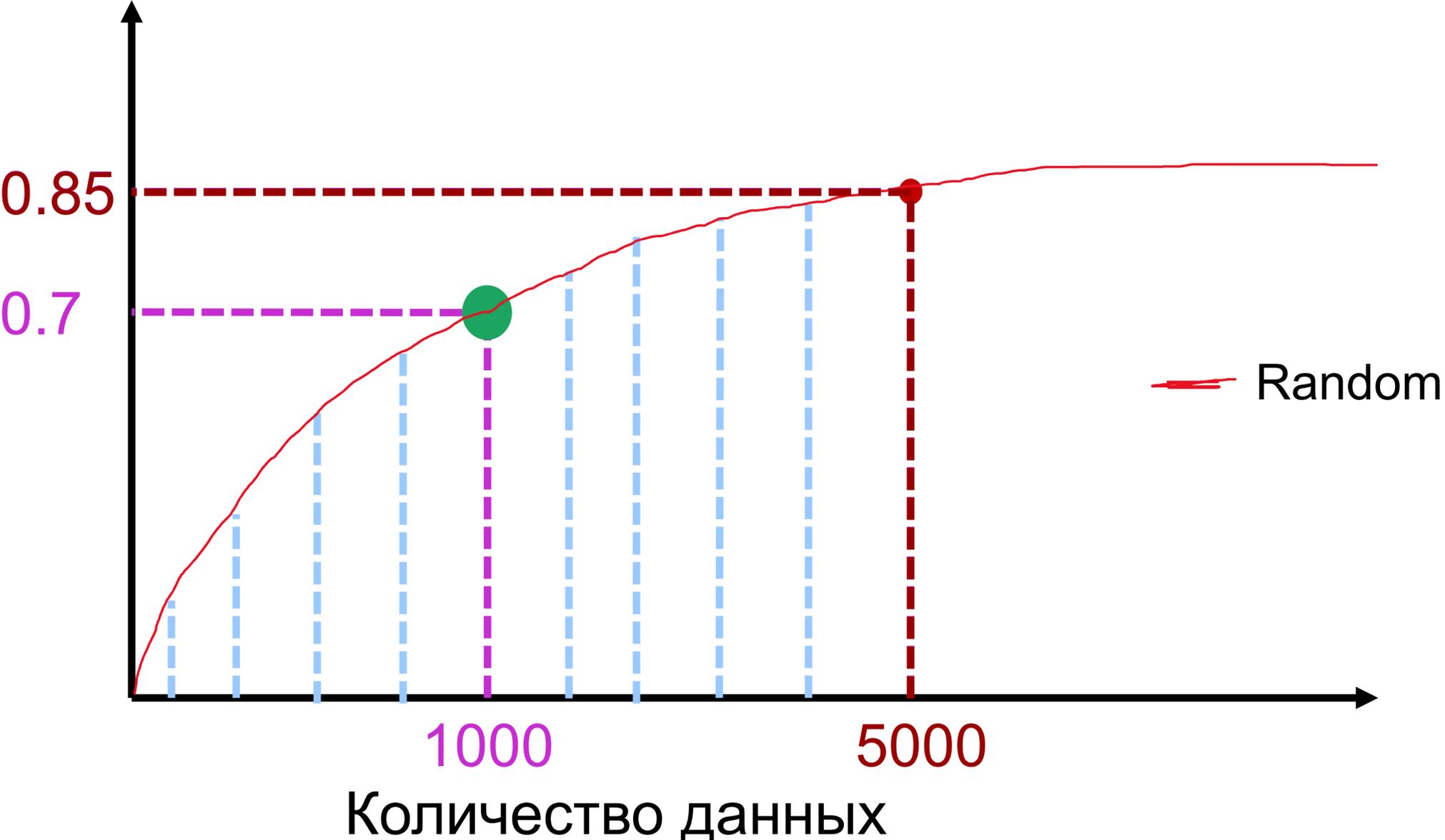
Есть различные методики:

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

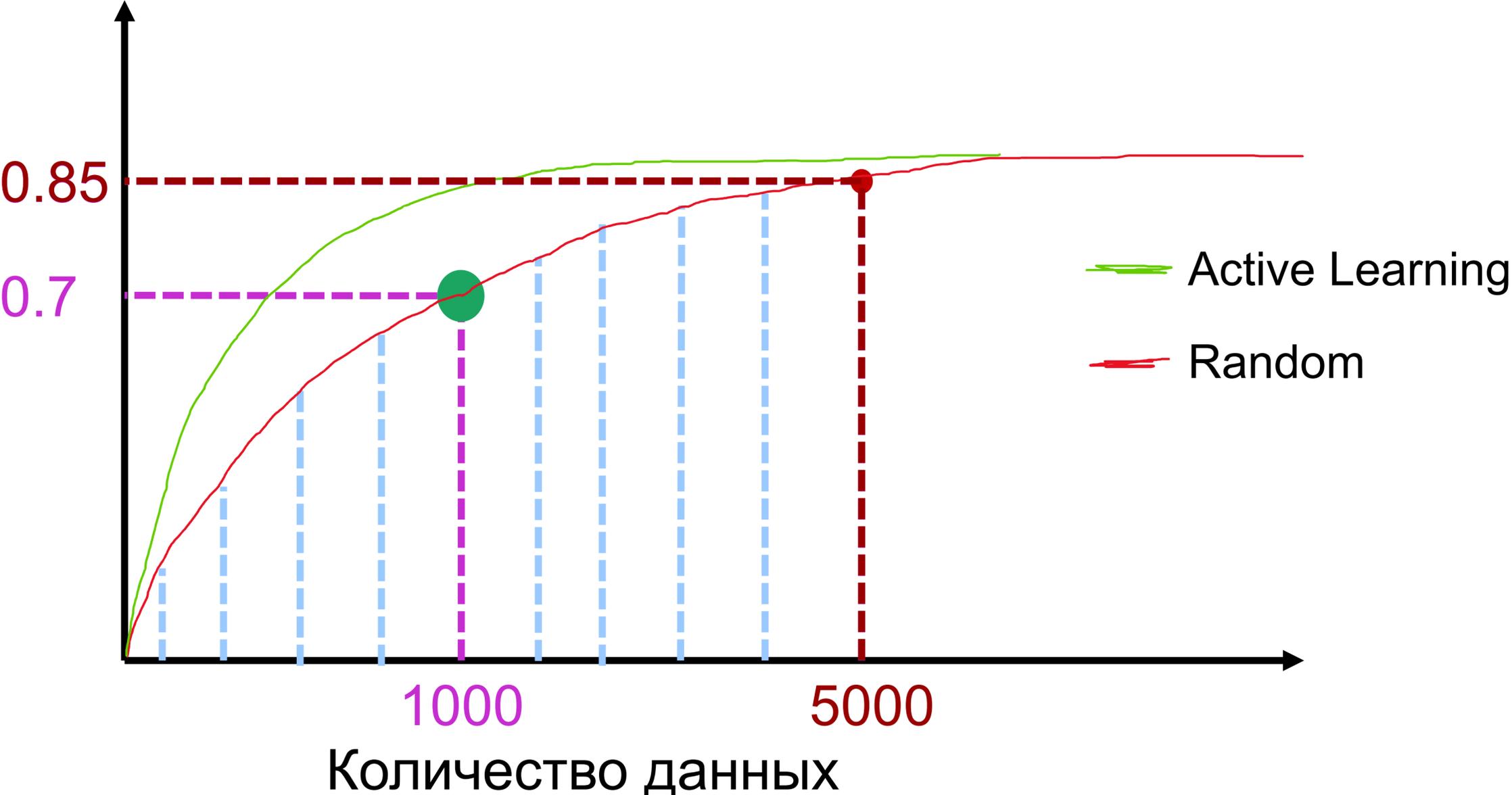
Active Learning

- 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

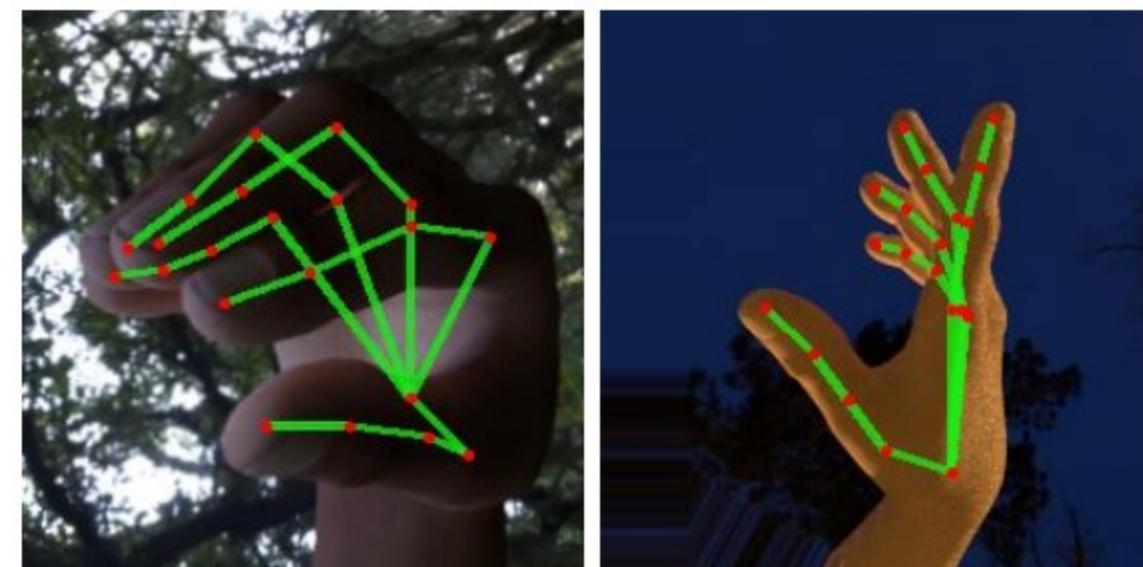
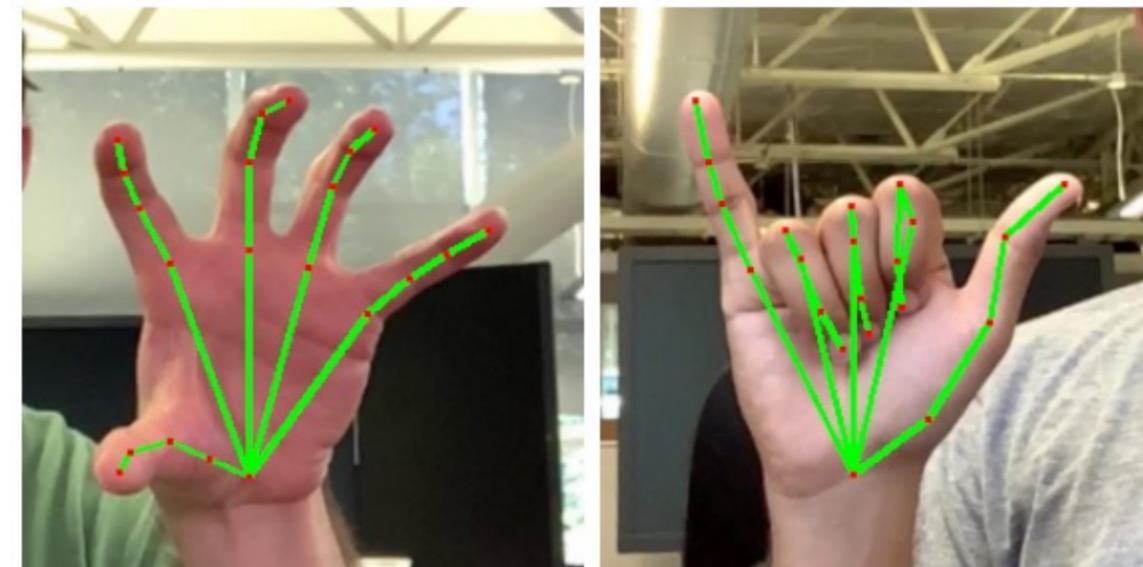
Active Learning



Active Learning



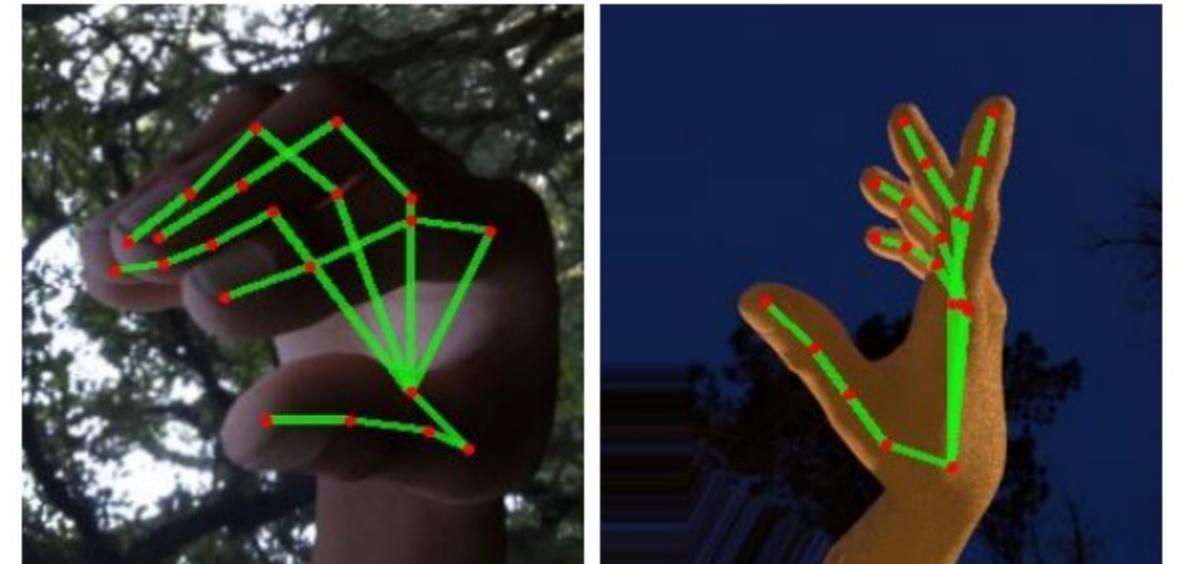
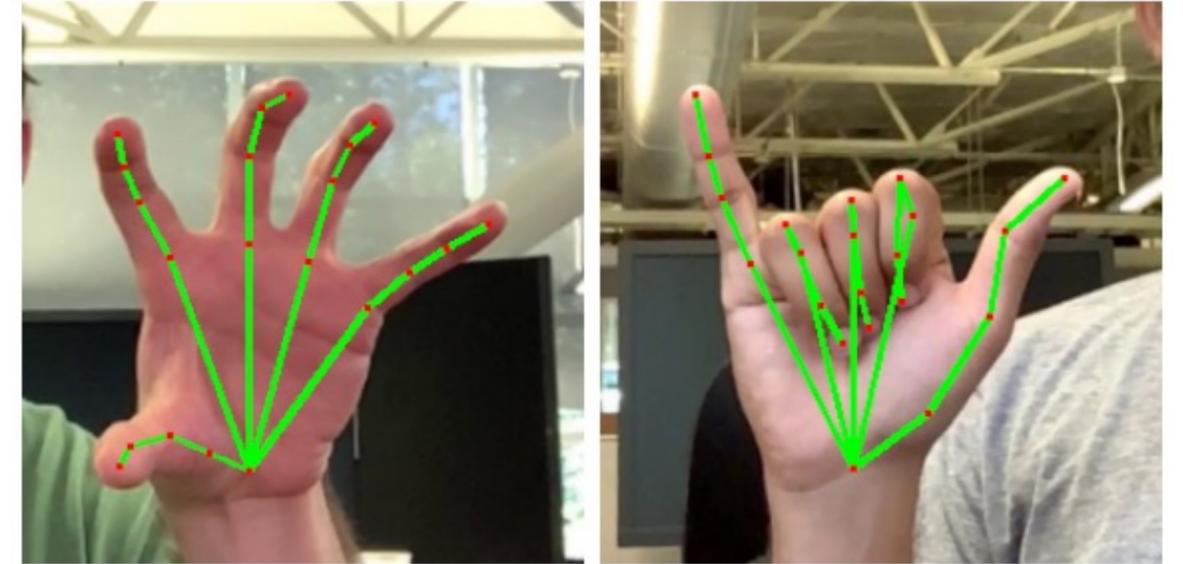
Synthetic datasets



Synthetic datasets

Используется, когда:

- Реальные данные собрать невозможно
- Сбор реальных данных очень дорогой



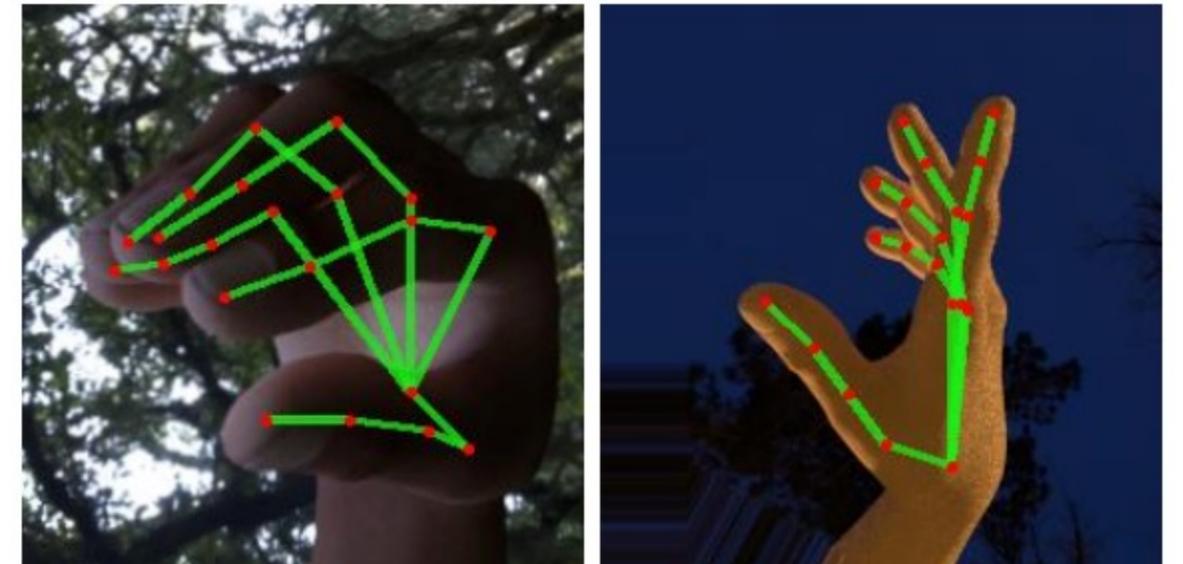
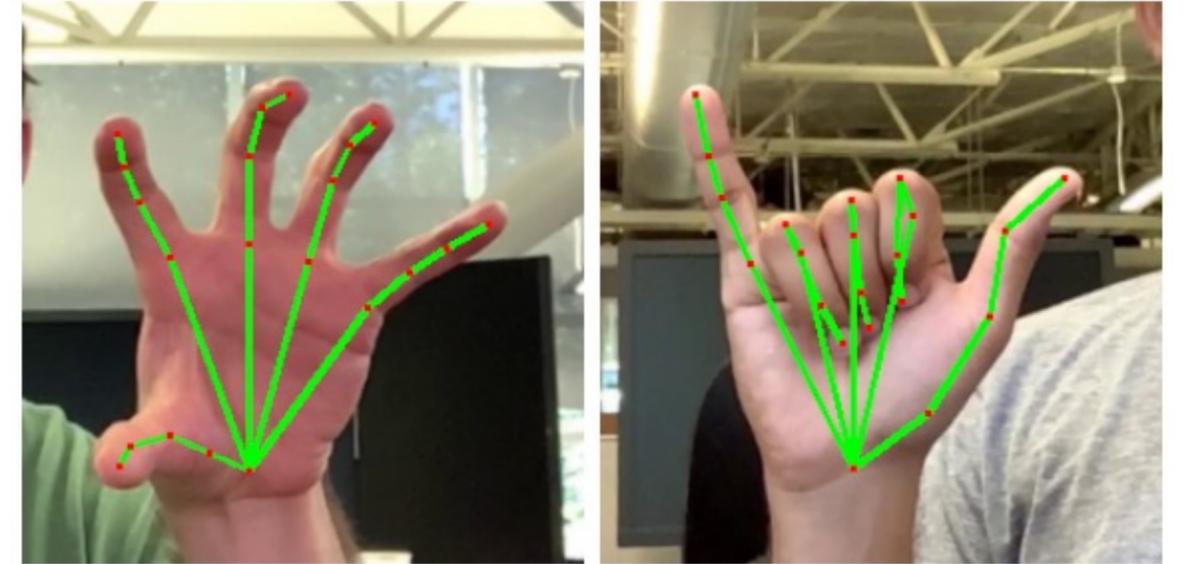
Synthetic datasets

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

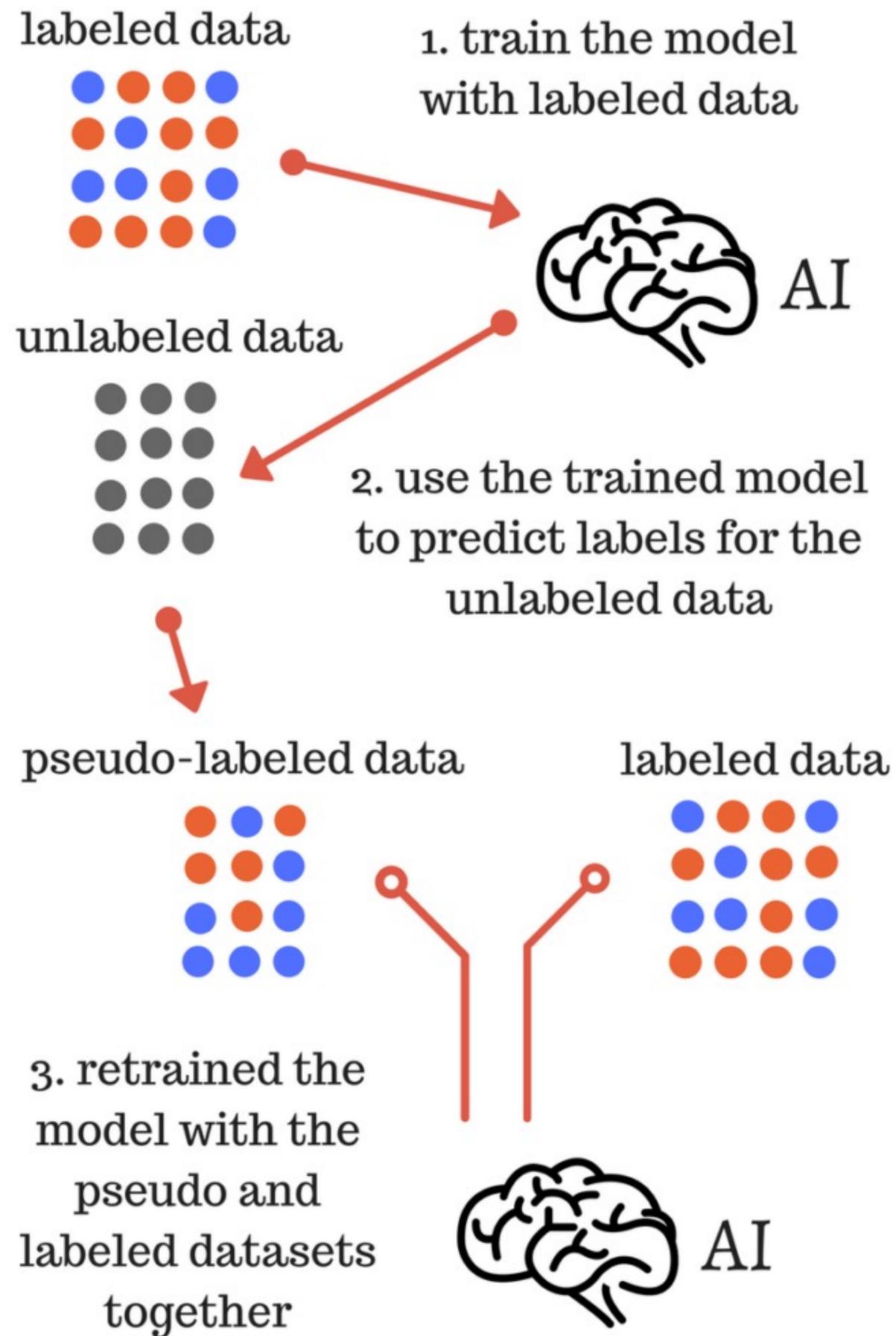


Synthetic datasets



Pseudo Labeling

Pseudo Labeling

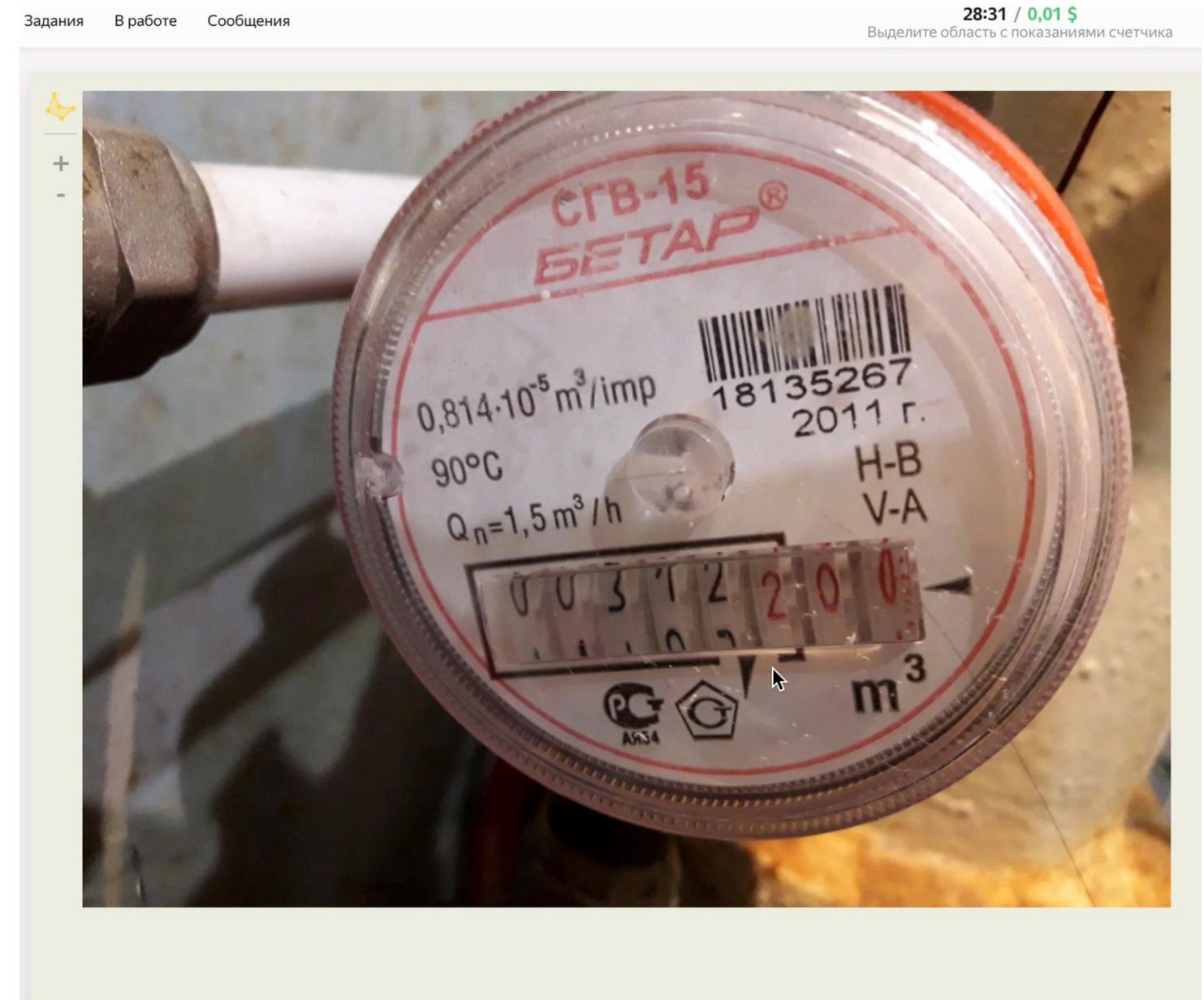


towardsdatascience.com/pseudo-labeling-to-deal-with-small-datasets-what-why-how-fd6f903213af

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 $\times 10$ раз
- Необходимо иметь хоть как-то работающую модель
- Немного смещаются ответы в разметке

Pre-labelling



Какой номер у автомобиля?

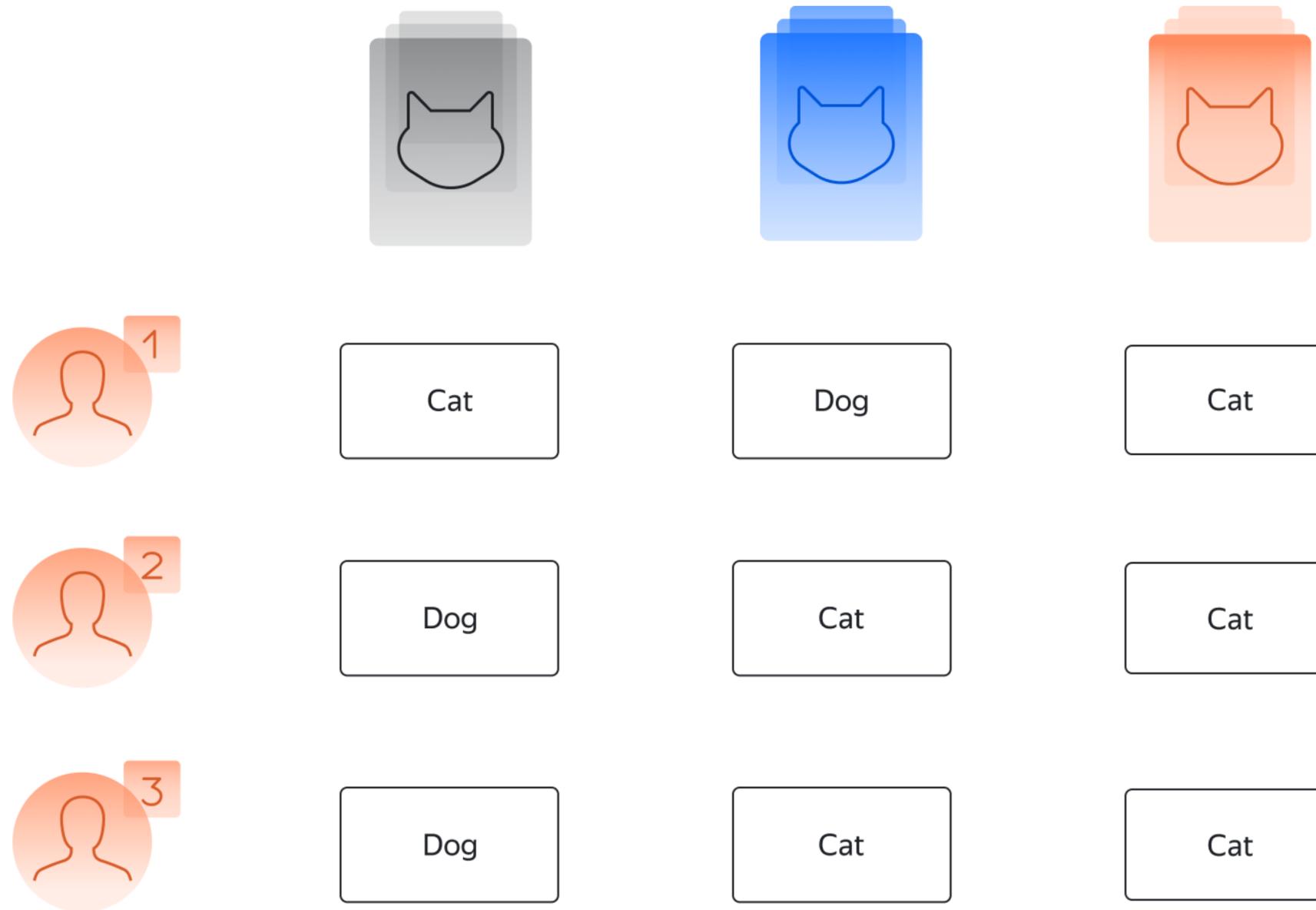
Pre-labelling



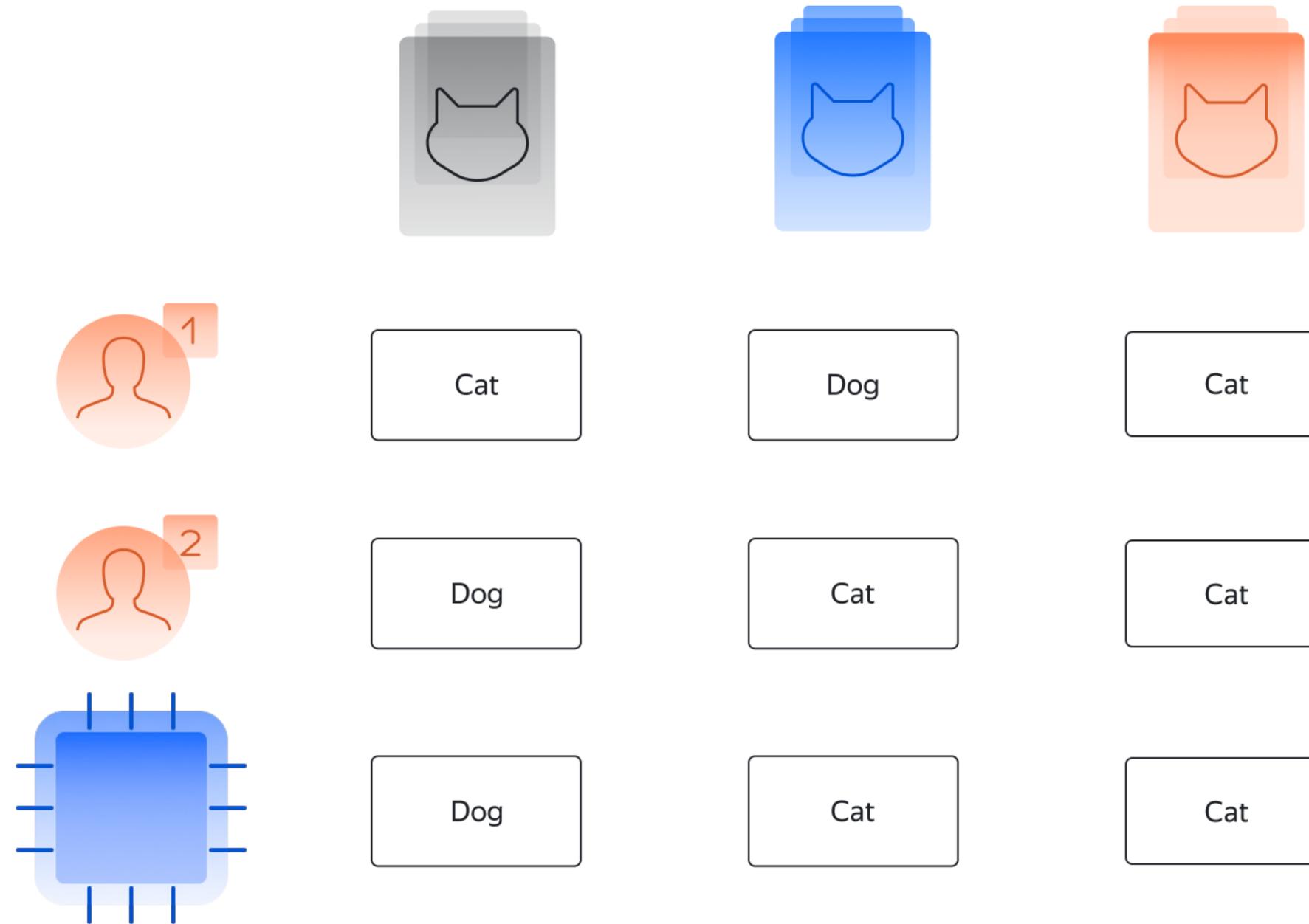
Какой номер у автомобиля?

Если бы нейронная сеть предзаполнила «В999ВХ74», вы бы стали исправлять?

Добавление ответов ML модели в консенсус



Добавление ответов ML модели в консенсус



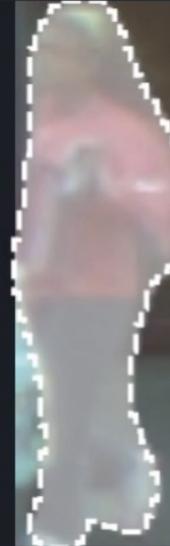
Поиск ошибок в разметке с помощью ML модели



Error type: Low Iou
Iou: 88.53%



Error type: Missing label
Predicted: Pedestrian (99.89%)



Error type: Missing label
Predicted: Pedestrian (97.23%)



Error type: Missing label
Predicted: Pedestrian (96.76%)



Error type: Missing label
Predicted: Pedestrian (90.05%)



Error type: Missing label
Predicted: Pedestrian (99.84%)



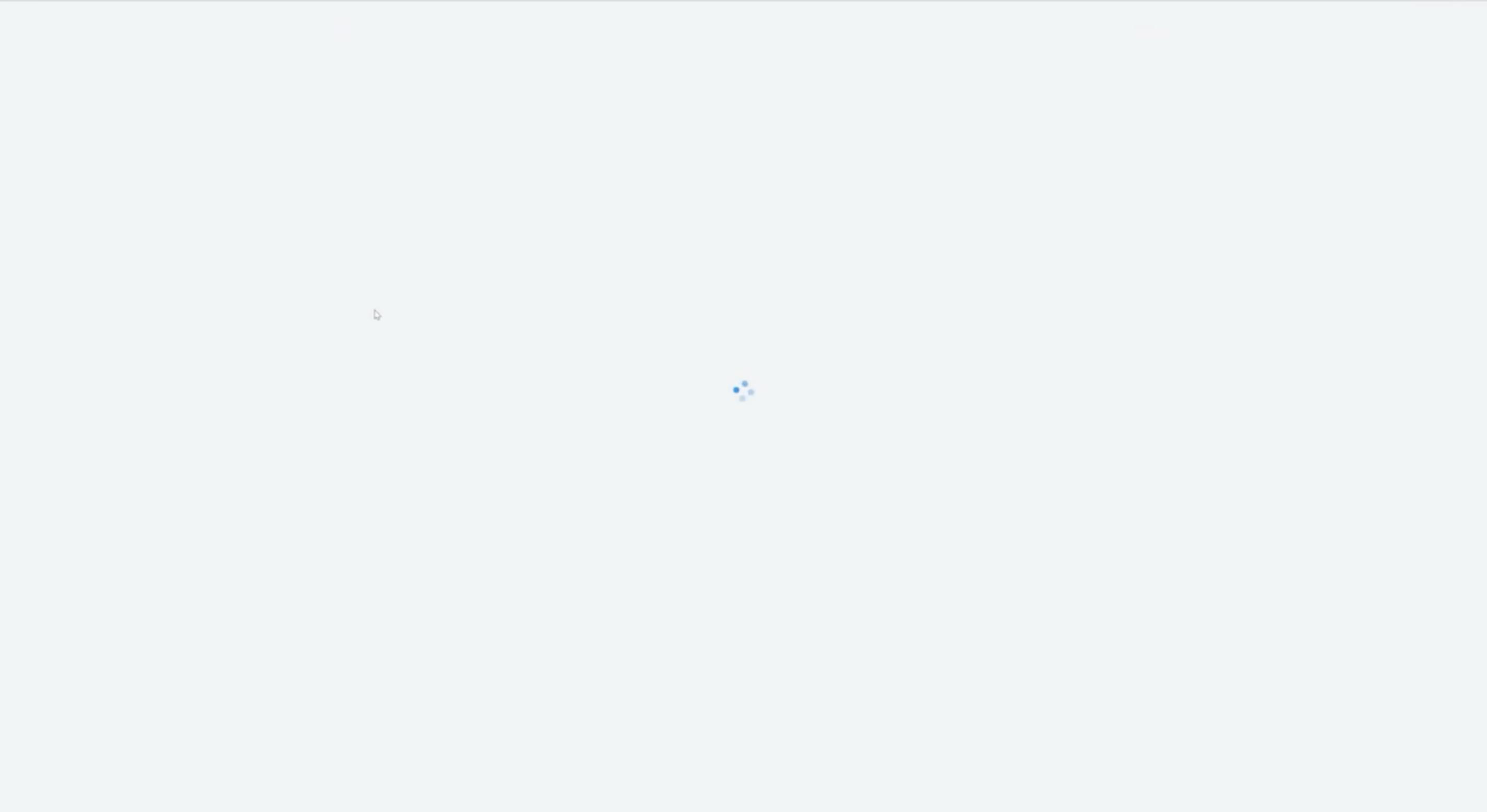
Error type: Low Iou
Iou: 87.28%



Error type: Low Iou
Iou: 86.96%



Интерактивная сегментация SAM и HRnet



Semi-automatic annotation: Point2Label

Label Studio Projects / OpenMMLabPlayground: / Labeling

SAM: Point2Label

#2 2 of 2

Outliner Manual Grouping Ordered by Time ↑

Regions not added

Auto accept annotation suggestions



cat 1 | person 2

cat 3 | person 4

OpenMMLab Playground

Semi-automatic annotation: Bbox2Label

Outliner

Manual Grouping Ordered by Time ↑

Regions not added

Auto accept annotation suggestions

SAM: Bbox2Label

OpenMMLab
Playground

cat 1 | person 2

cat 3 | person 4

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

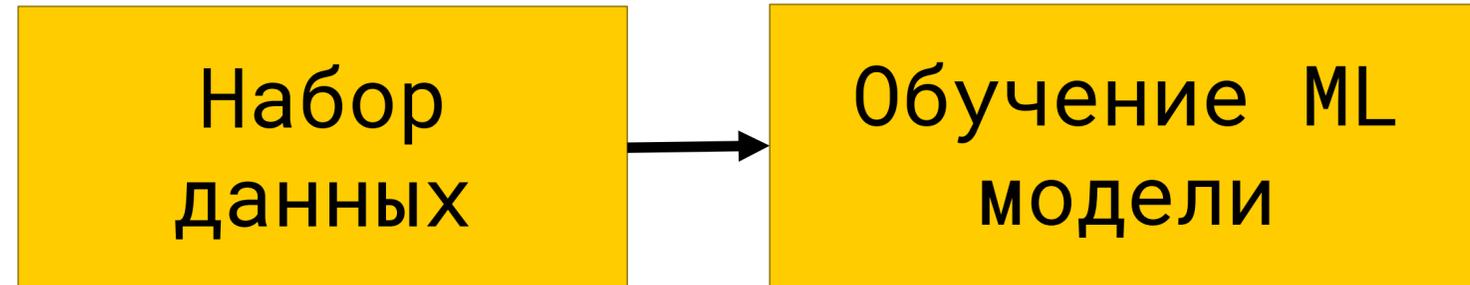


Система фиксации
нарушений

Human in the Loop

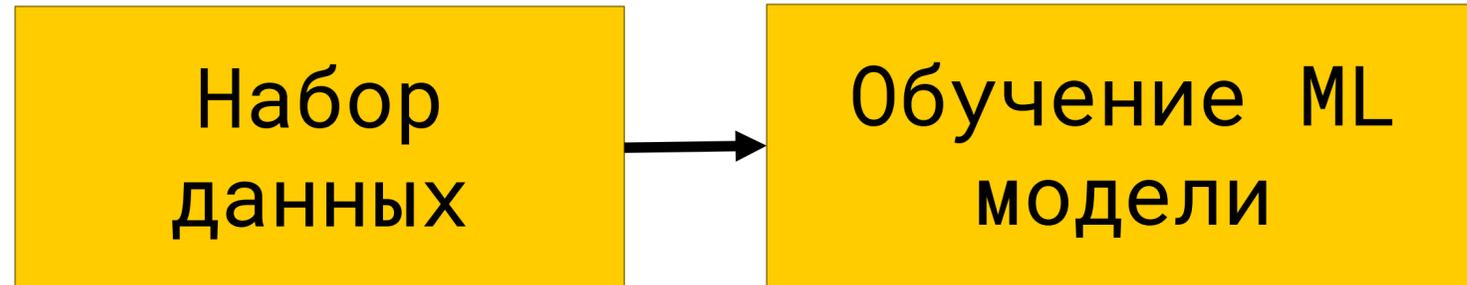
Набор
данных

Human in the Loop

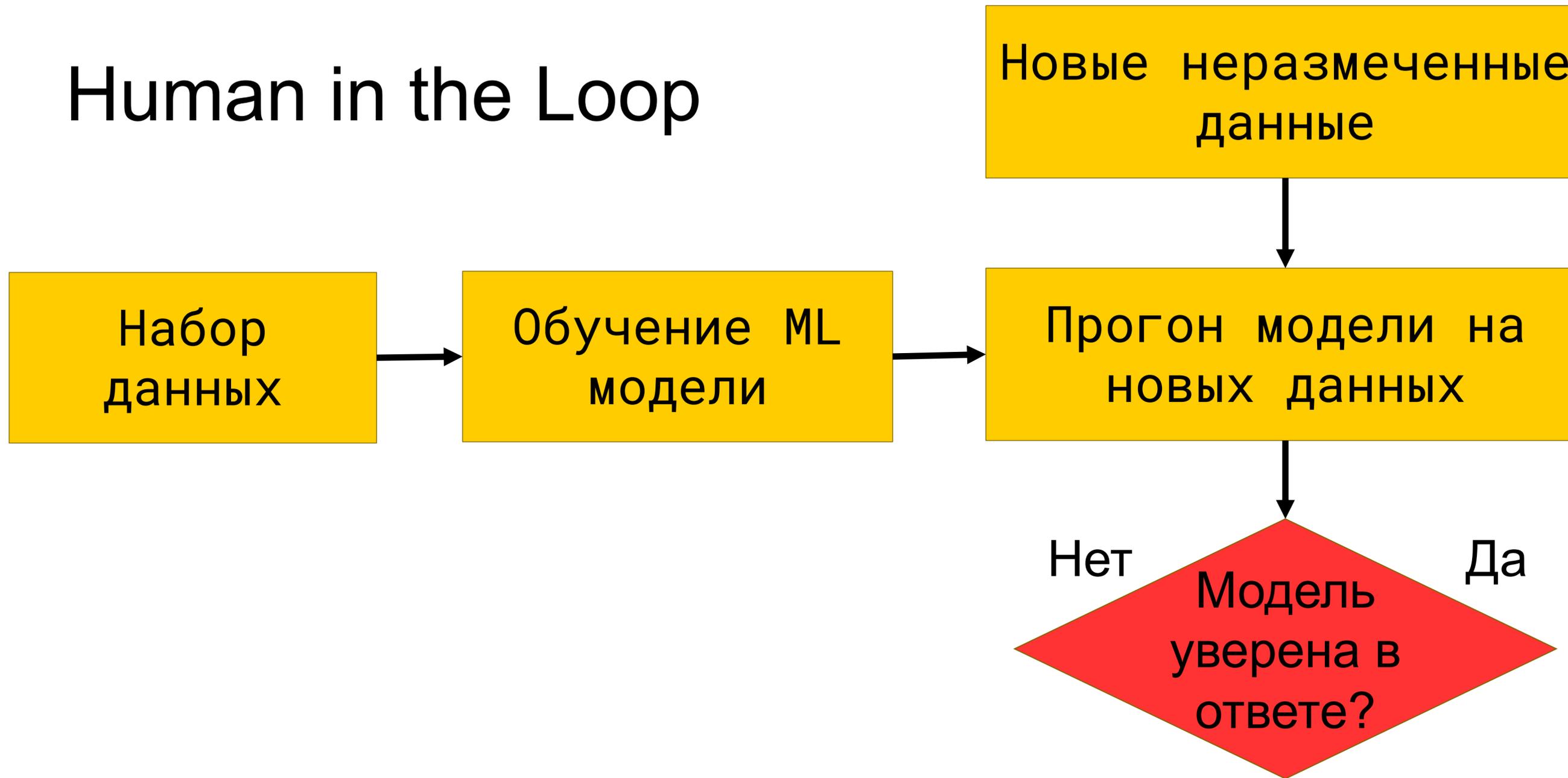


Human in the Loop

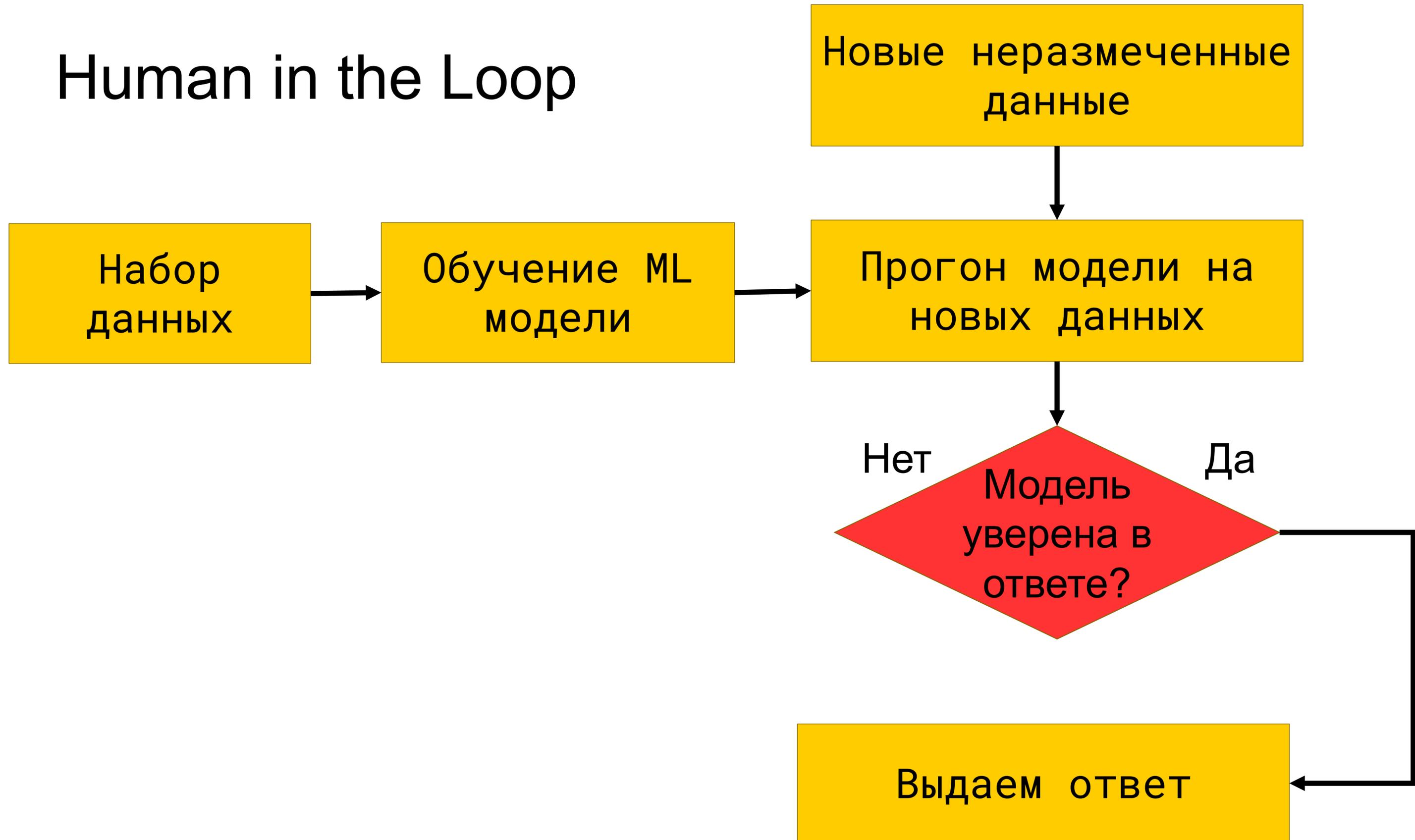
Новые неразмеченные
данные



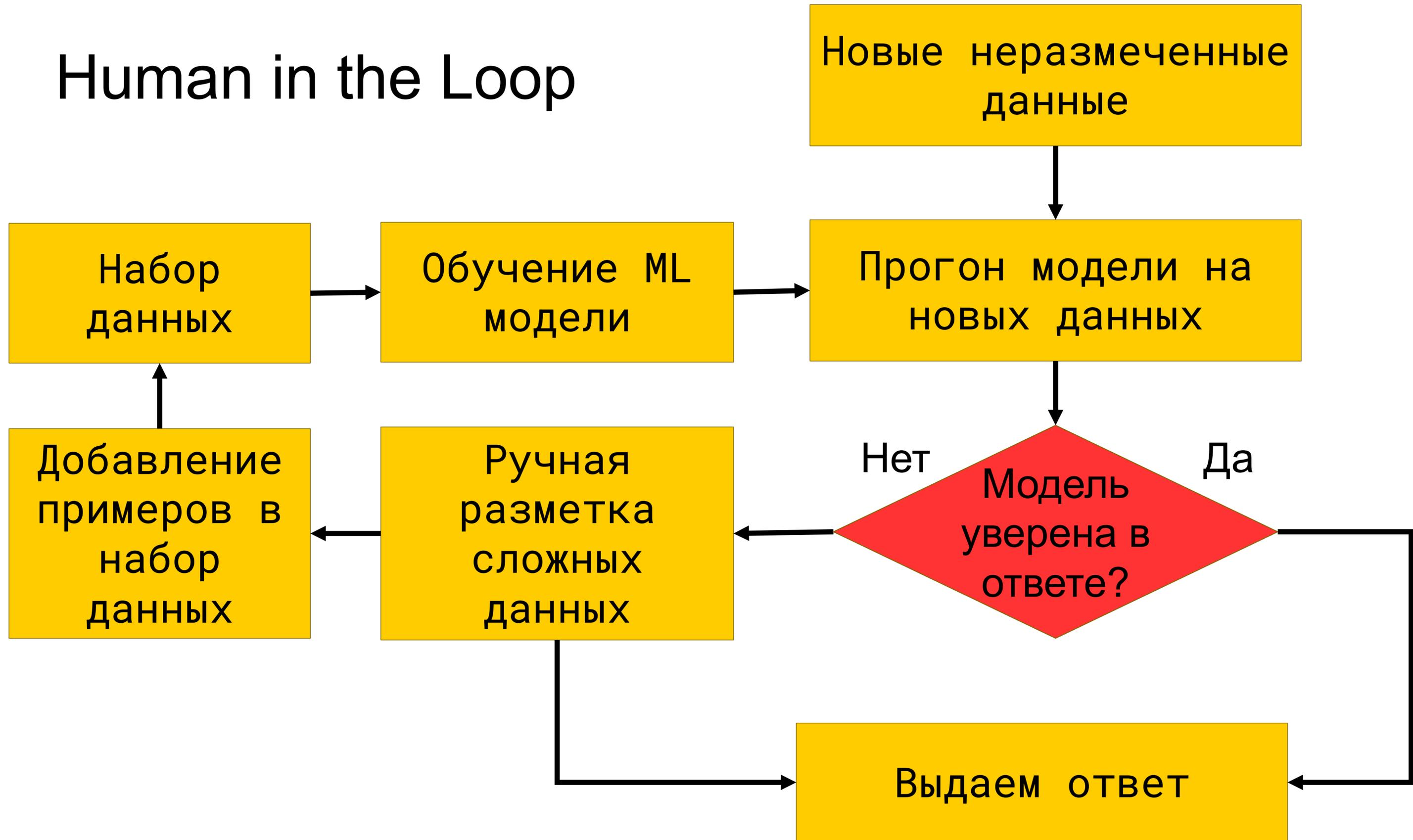
Human in the Loop



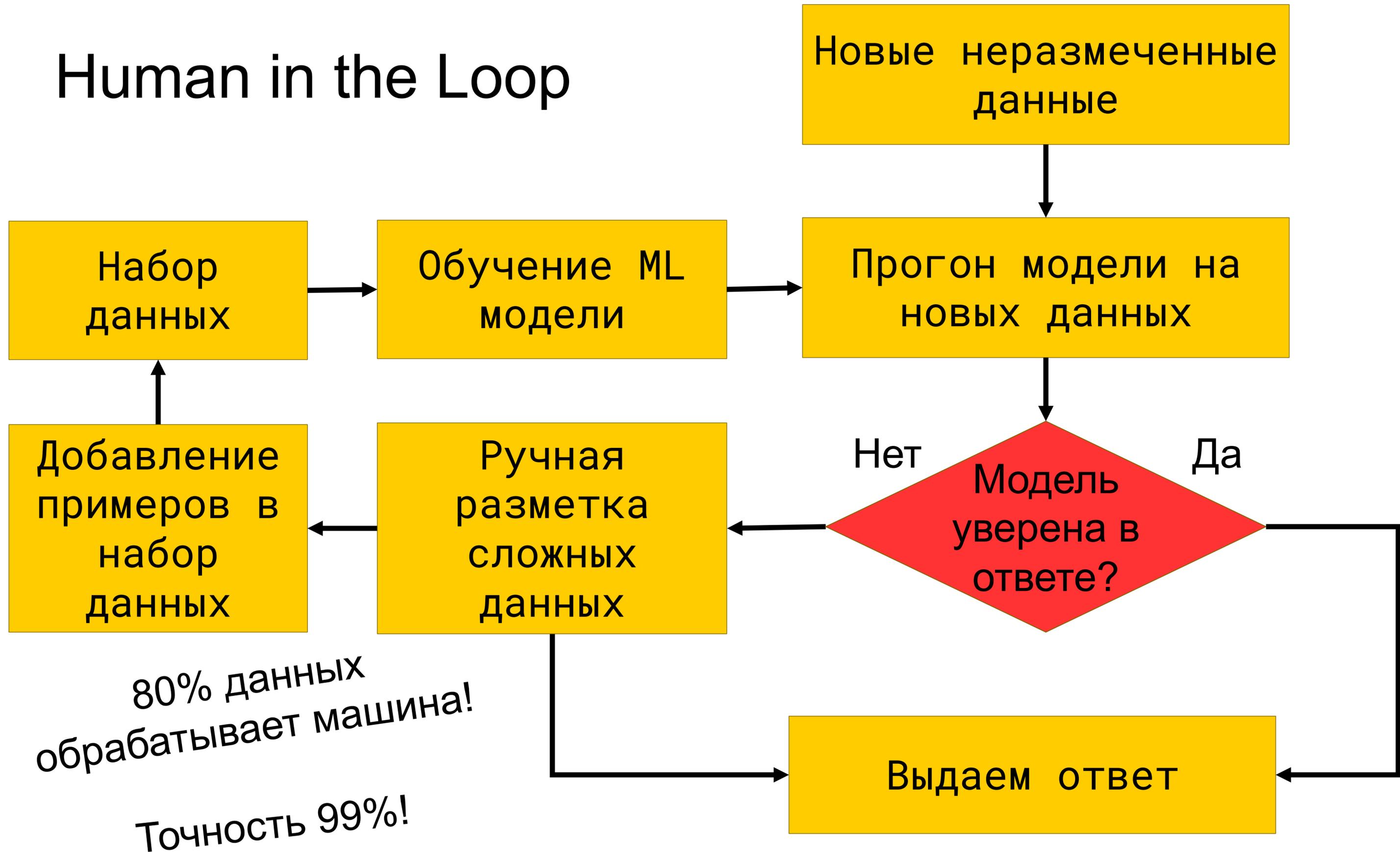
Human in the Loop



Human in the Loop



Human in the Loop



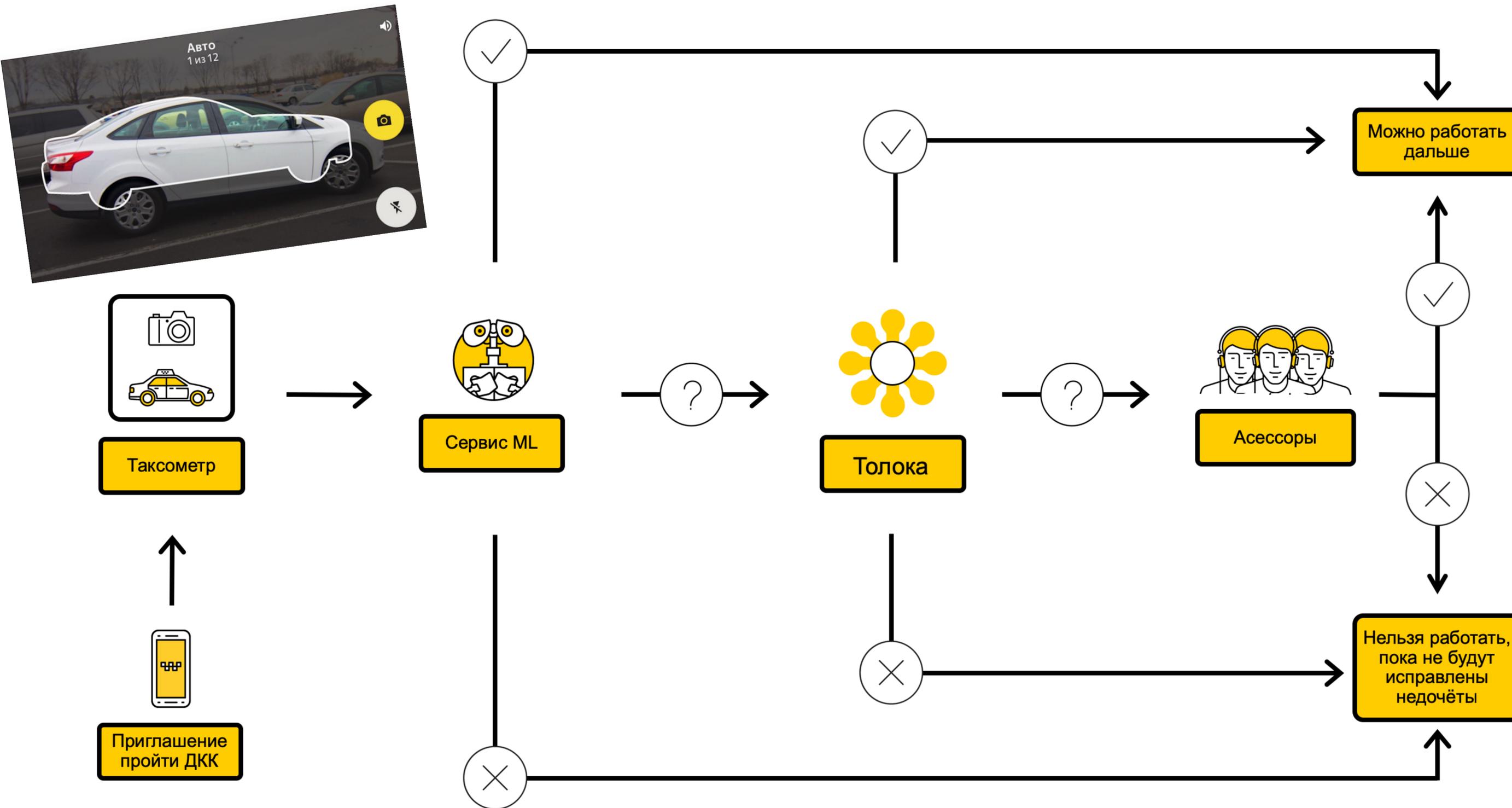
Human in the Loop в Dbrain



Отдельные сервисы для решения любой задачи



Human in the Loop в Яндекс.Такси



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

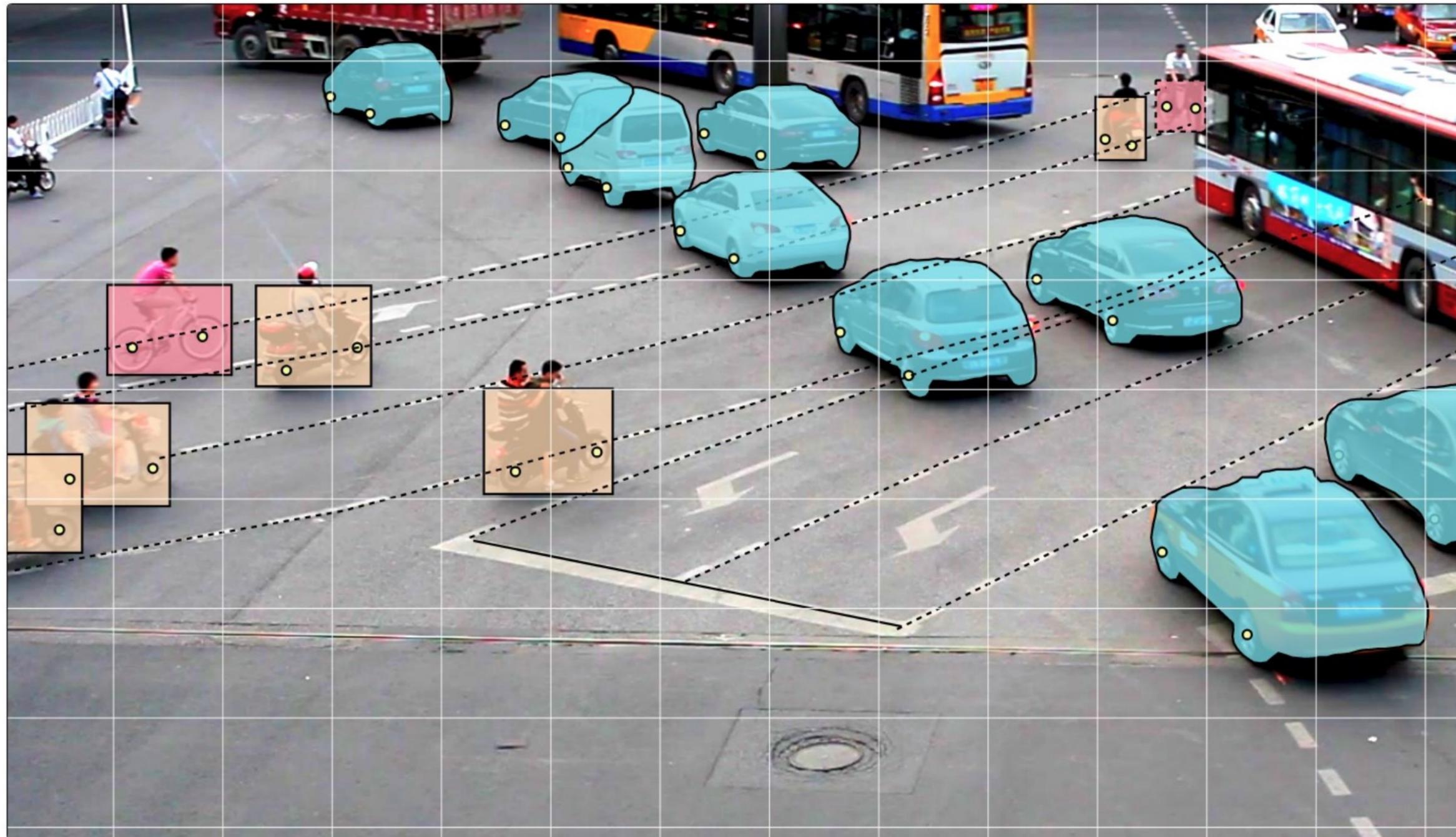
Автоматизация

The image shows the CVAT web interface. At the top, there is a navigation bar with 'CVAT' logo and menu items: Projects, Tasks, Jobs, Cloud Storages, Models, Analytics. On the right of the navigation bar, there are icons for help, user profile (kucev), and a dropdown menu. Below the navigation bar is a toolbar with icons for Menu, Save, Undo, Redo, Done, Block, and a progress bar. The main workspace displays a photograph of four people standing in a room with wooden beams. A bounding box is drawn around the person on the far right, and a 'Points' control is visible above it. On the right side, there is a panel with tabs for 'Objects', 'Labels', and 'Issues'. Below these tabs, there are icons for lock, visibility, and a 'Sort by' dropdown menu set to 'ID - as...'. At the bottom right, there is an 'Appearance' panel with options for 'Color by' (Label, Instance, Group), 'Opacity' (with a slider), 'Outlined borders', 'Show bitmap', and 'Show projections'.

<https://youtu.be/bBqersGW0ic>

Повышение квалификации разметчика

РАЗМЕТЧИК ДАННЫХ: КАК СТАТЬ СПЕЦИАЛИСТОМ В РАСТУЩЕЙ ОБЛАСТИ
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ



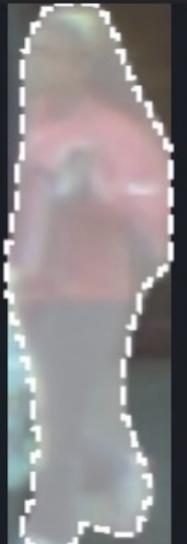
Валидация вместо разметки



Error type: Low iou
iou: 88.53%



Error type: Missing label
Predicted: Pedestrian (99.89%)



Error type: Missing label
Predicted: Pedestrian (97.23%)



Error type: Missing label
Predicted: Pedestrian (96.76%)



Error type: Missing label
Predicted: Pedestrian (90.05%)



Error type: Missing label
Predicted: Pedestrian (99.84%)



Error type: Low iou
iou: 87.28%



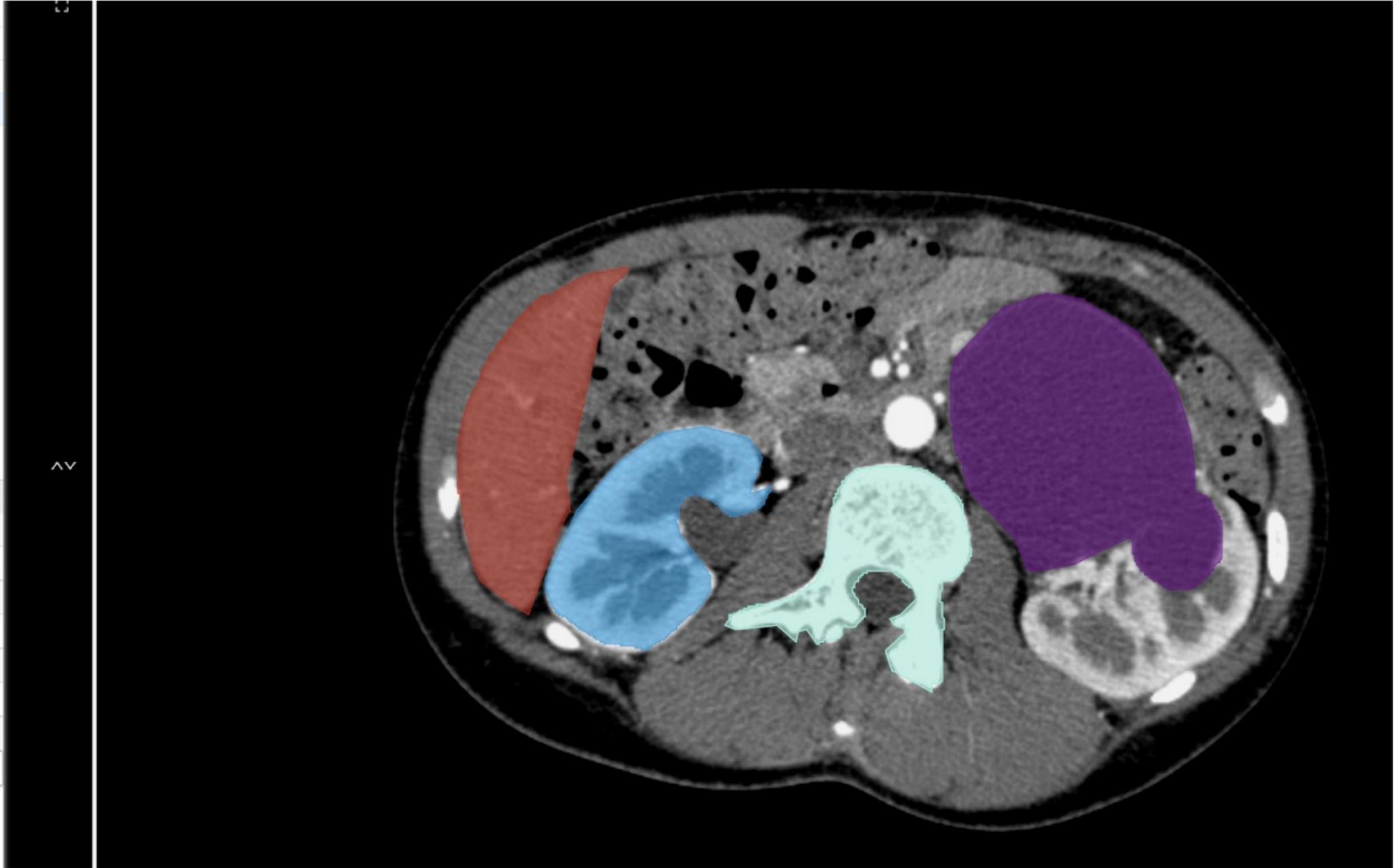
Error type: Low iou
iou: 86.96%



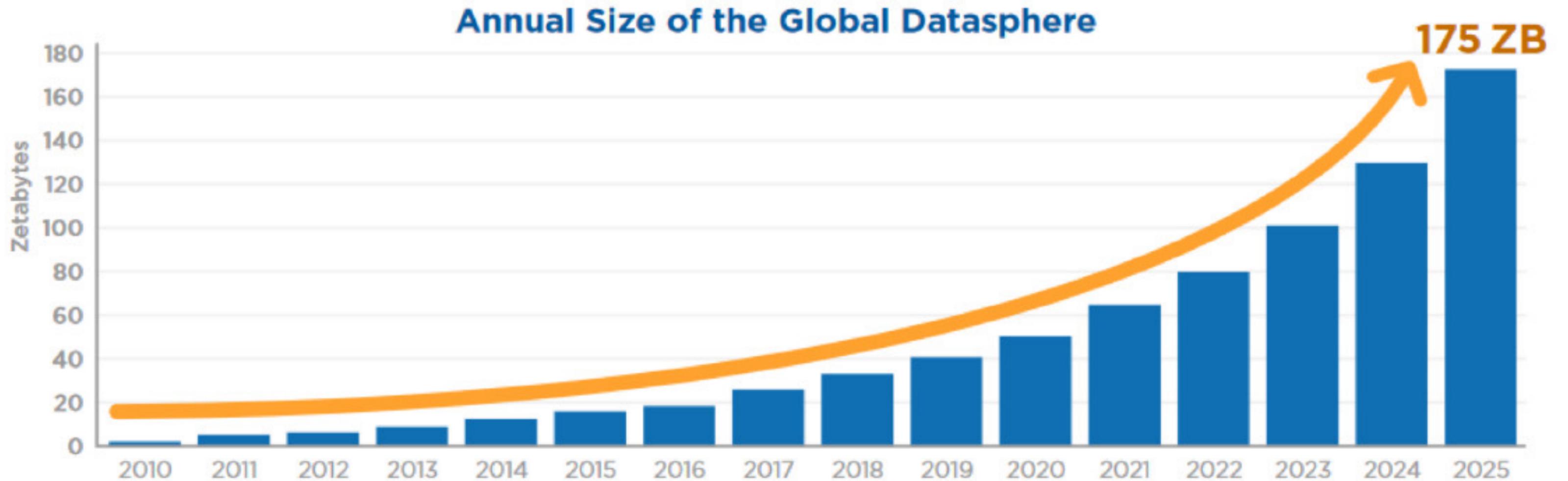
Преобладание разметки узкими специалистами

Liver	1
Kidney	2
Organ	3
Bone	4

Objects	Total: 4	
▼ Kidney (1)		⊙
> Kidney		⊙ 🗑️ ⋮
▼ Liver (1)		⊙
> Liver		⊙ 🗑️ ⋮
▼ Organ (1)		⊙
> Organ		⊙ 🗑️ ⋮
▼ Bone (1)		⊙
> Bone		⊙ 🗑️ ⋮



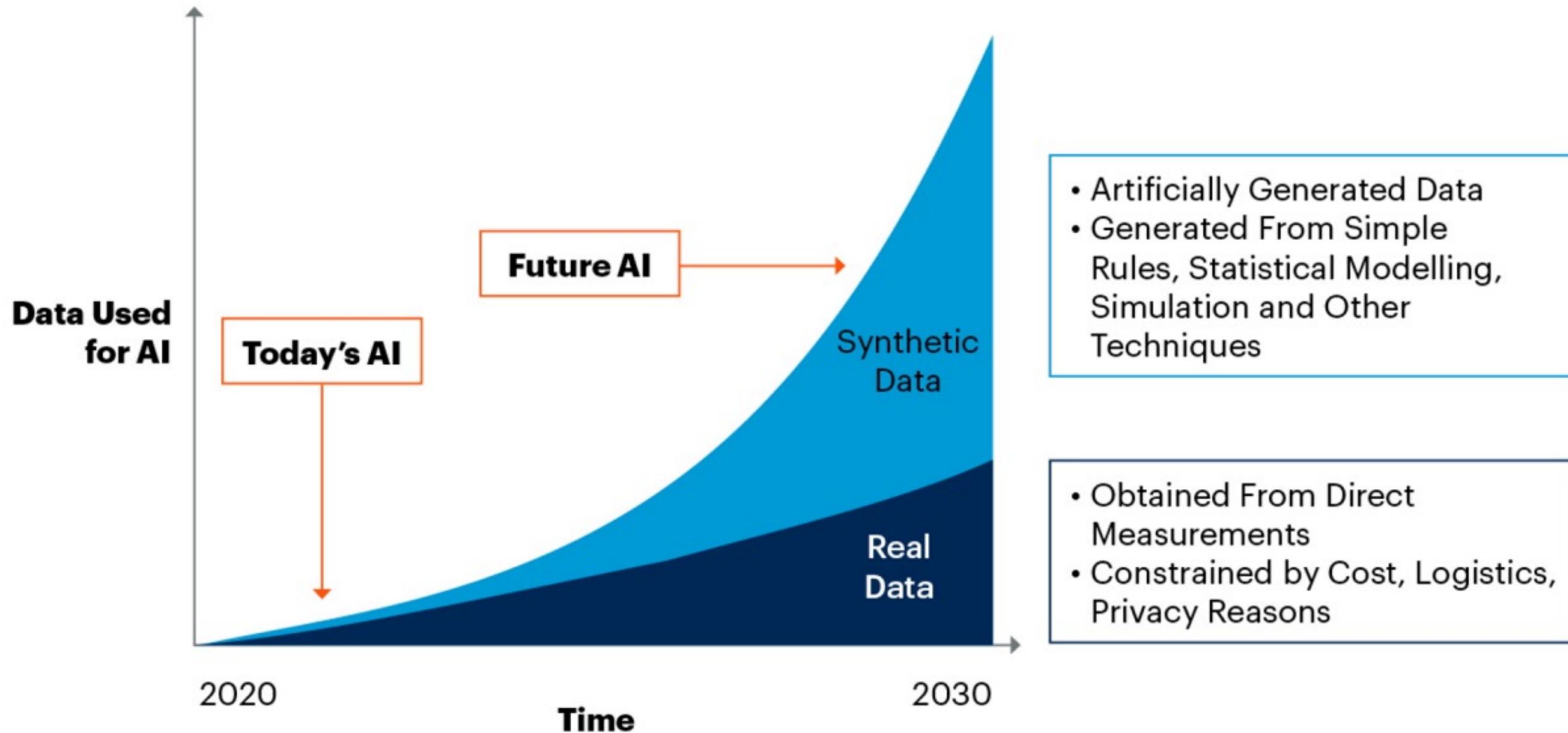
Увеличение объема размечаемых данных



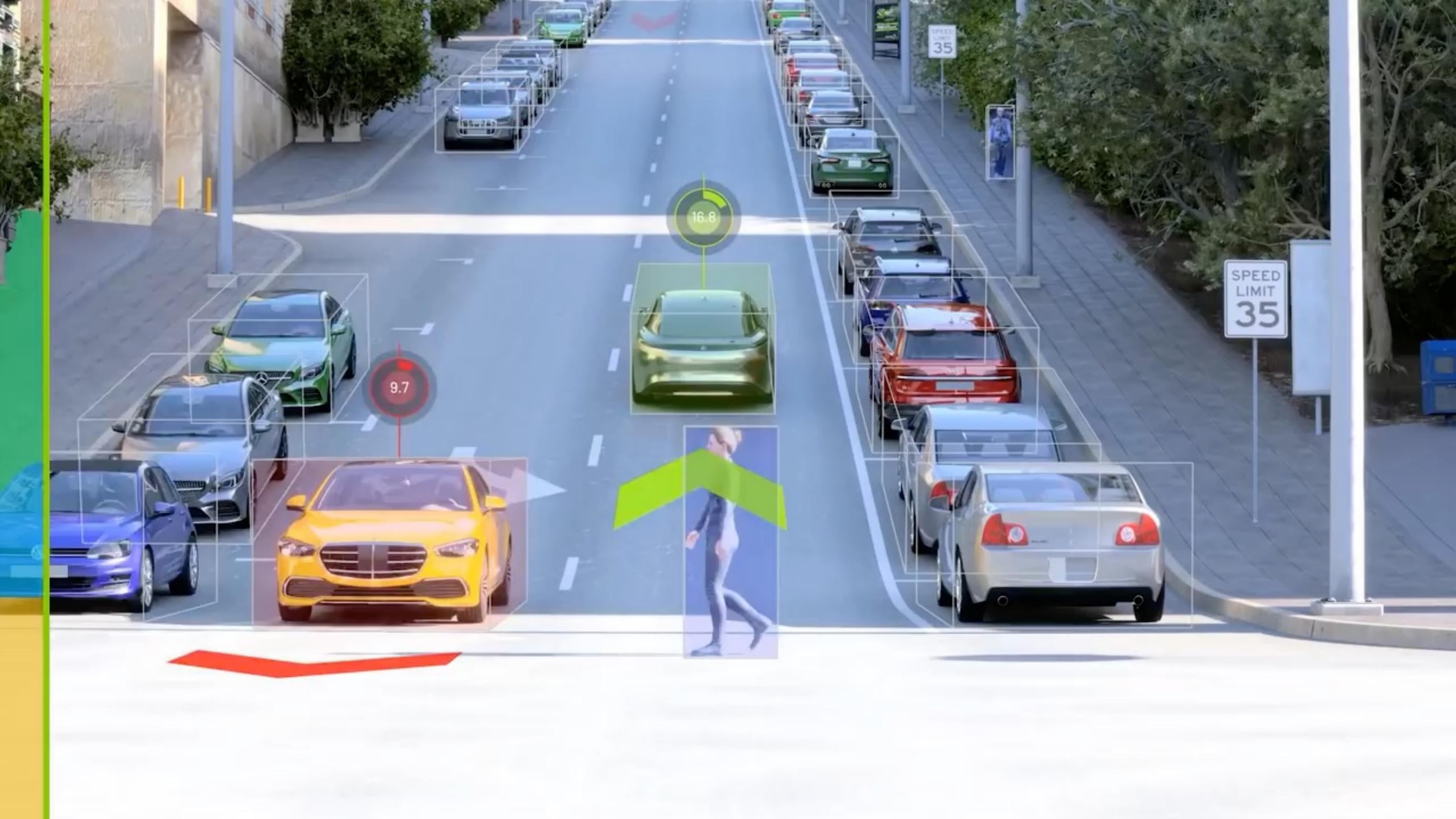
Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Рендеринг синтетических данных

By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models



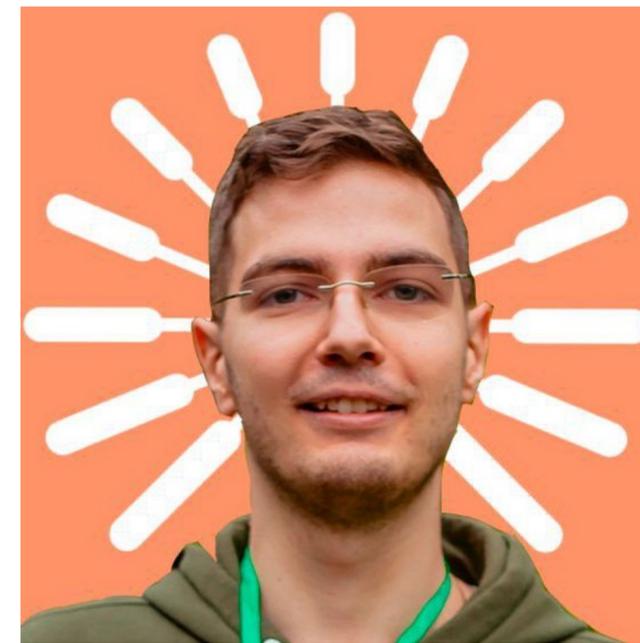
Рендеринг синтетических данных





Я: Куцев Роман

t.me/roman_kucev



ТГ канал: Рома ❤️ Толоку

t.me/roma_toloka

