

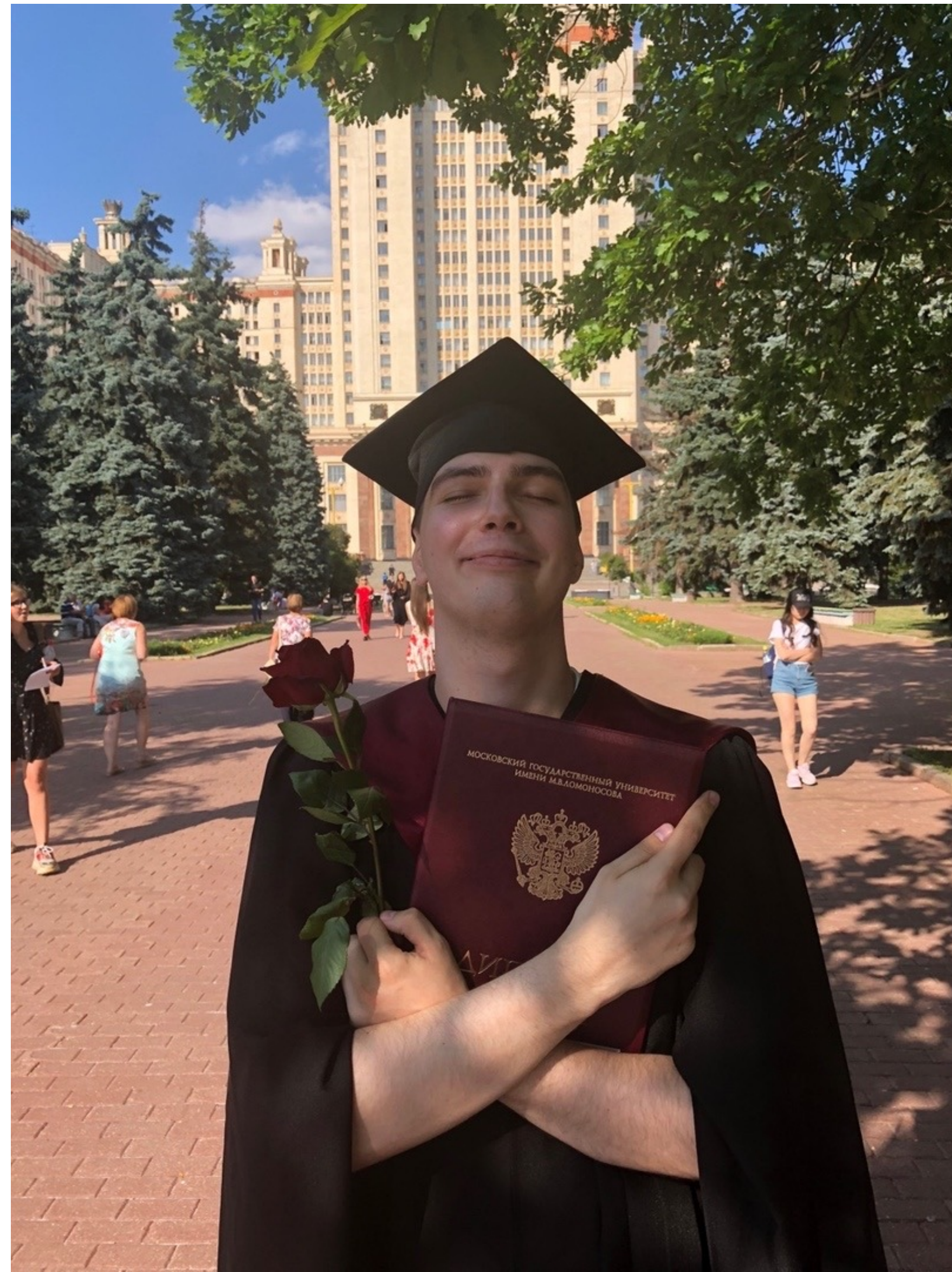
Методы оптимизации разметки для обучения нейронных сетей

Куцев Роман
CTO TrainingData.Solutions

Несколько фактов о себе

- Занимаюсь сбором и разметкой данных больше 5 лет
- Руководил штатом разметчиков в Prisma
- Являюсь сертифицированным партнером Толоки
- Реализовал более 300 проектов по разметке
- Преподаю краудсорсинг в ШАДе и ВШЭ

Несколько фактов о себе



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?

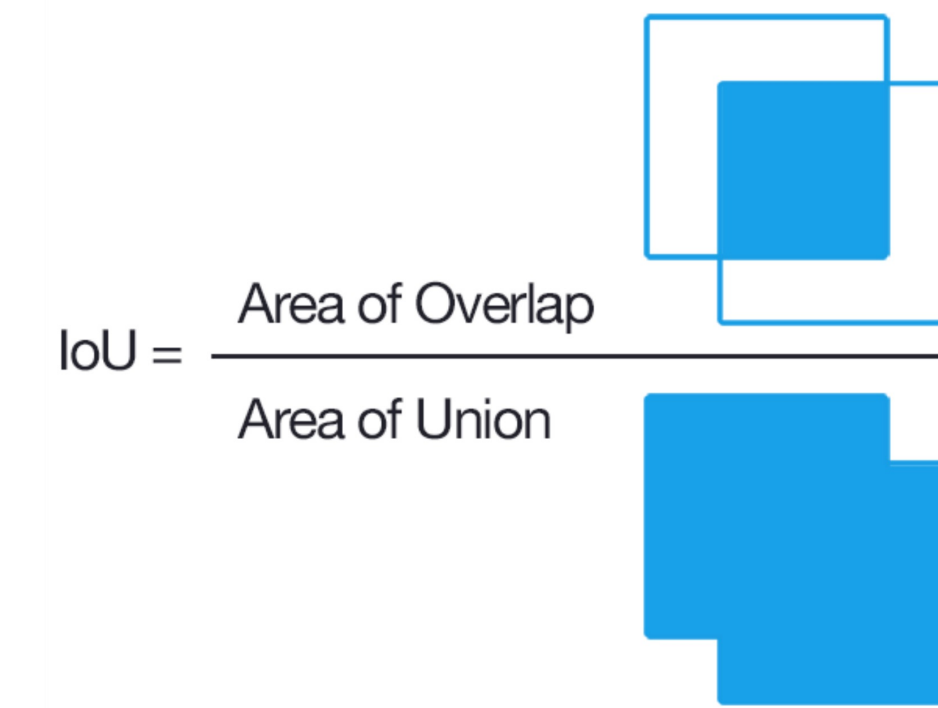
Цель

Получить Intersection over Union (IoU) > 0.95

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$


Цель

Получить Intersection over Union (IoU) > 0.95

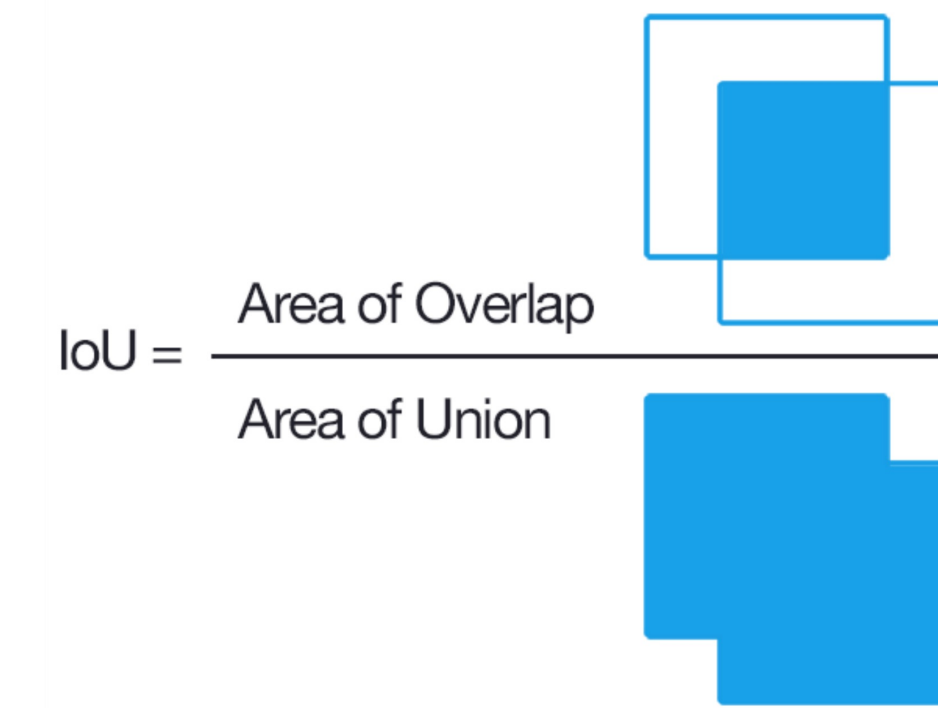


Вопросы

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

Цель

Получить Intersection over Union (IoU) > 0.95



Вопросы

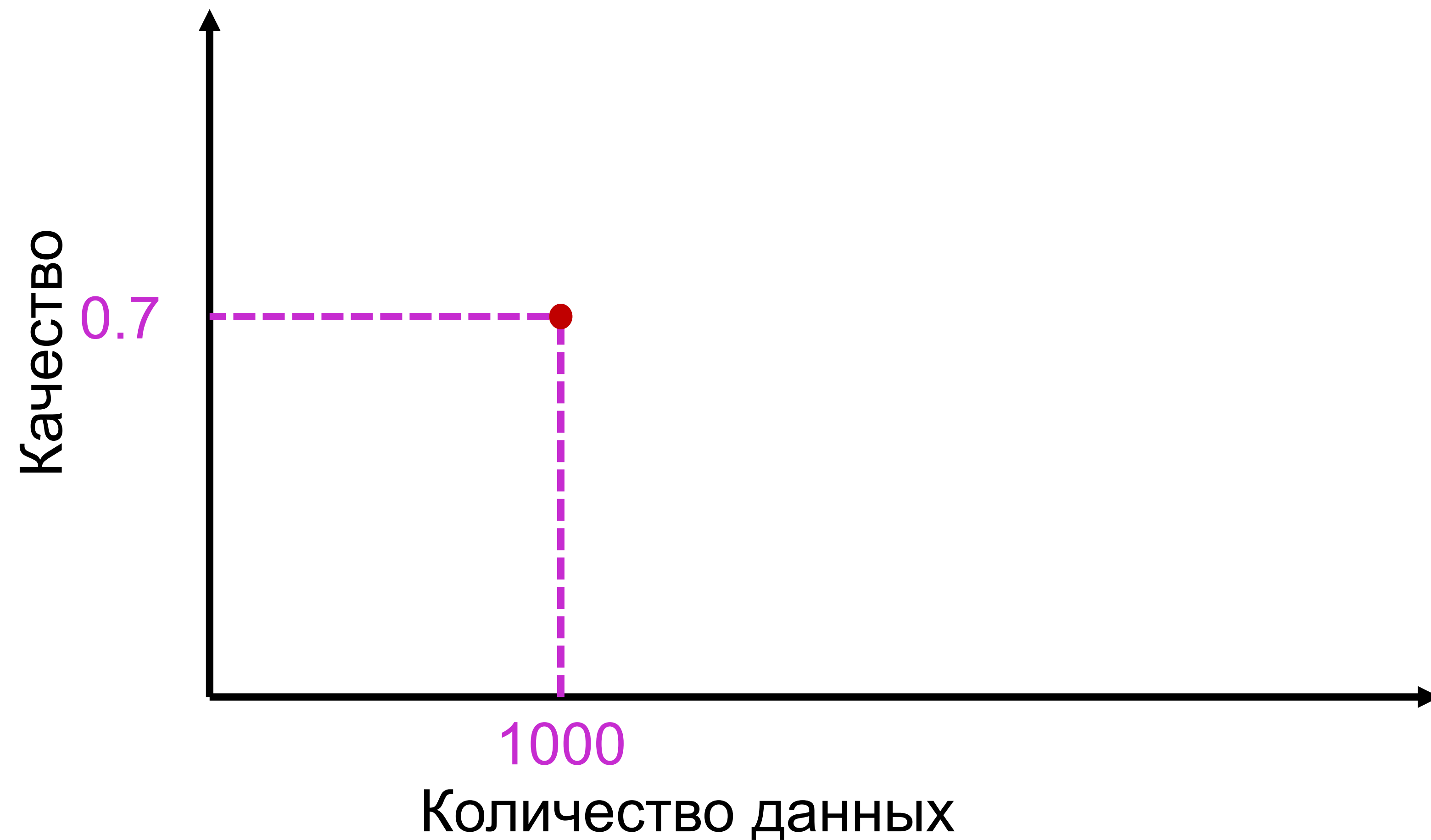
1. Как составить отражающий реальность валидационный датасет?
2. Какую архитектуру выбрать?
3. Сколько данных нужно?

Шаг 1. Разметим 1000 изображений

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

Шаг 1. Разметим 1000 изображений

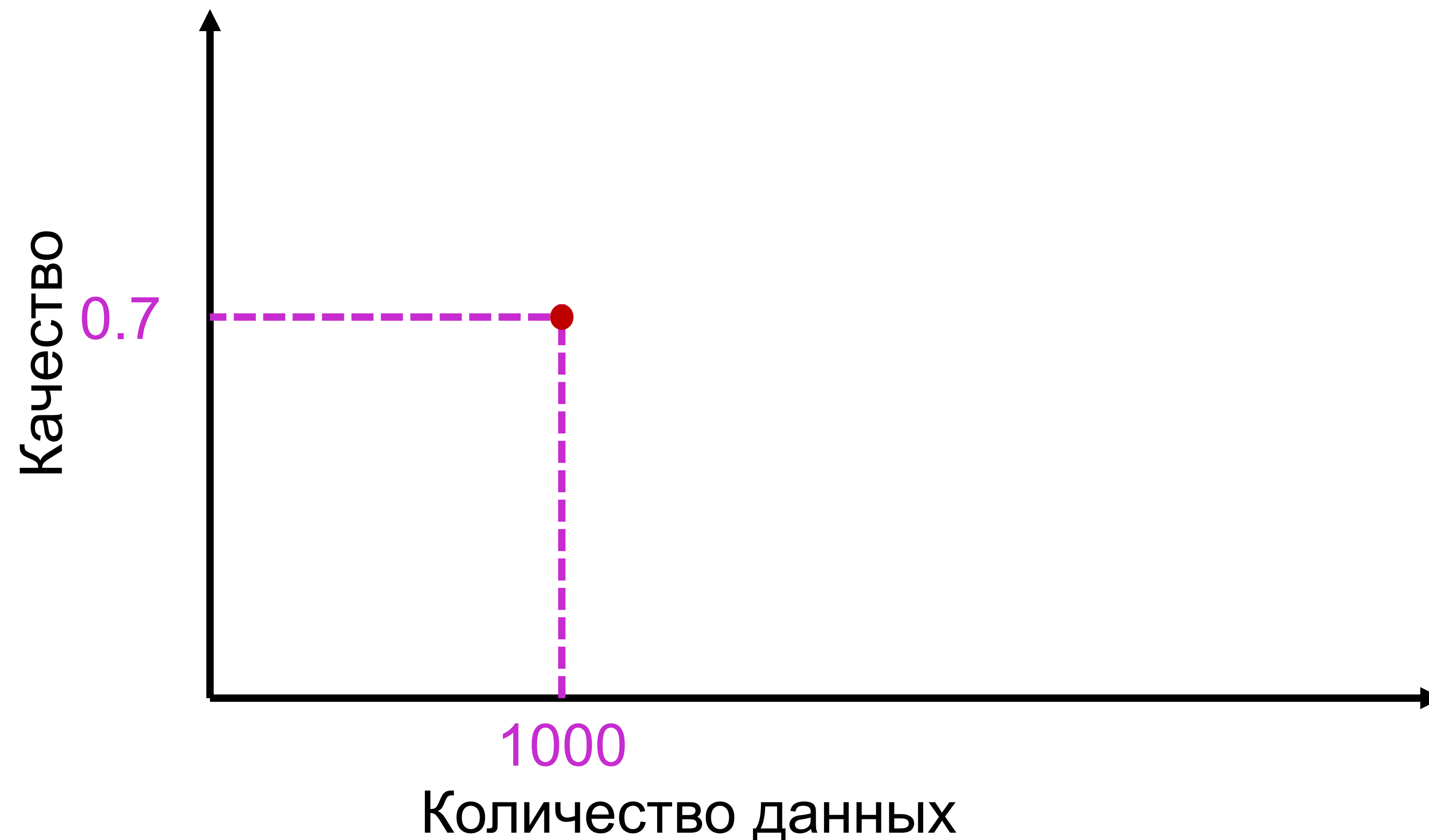
- Почему 1000? Эмпирически, по опыту в других задачах.



Шаг 1. Разметим 1000 изображений

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

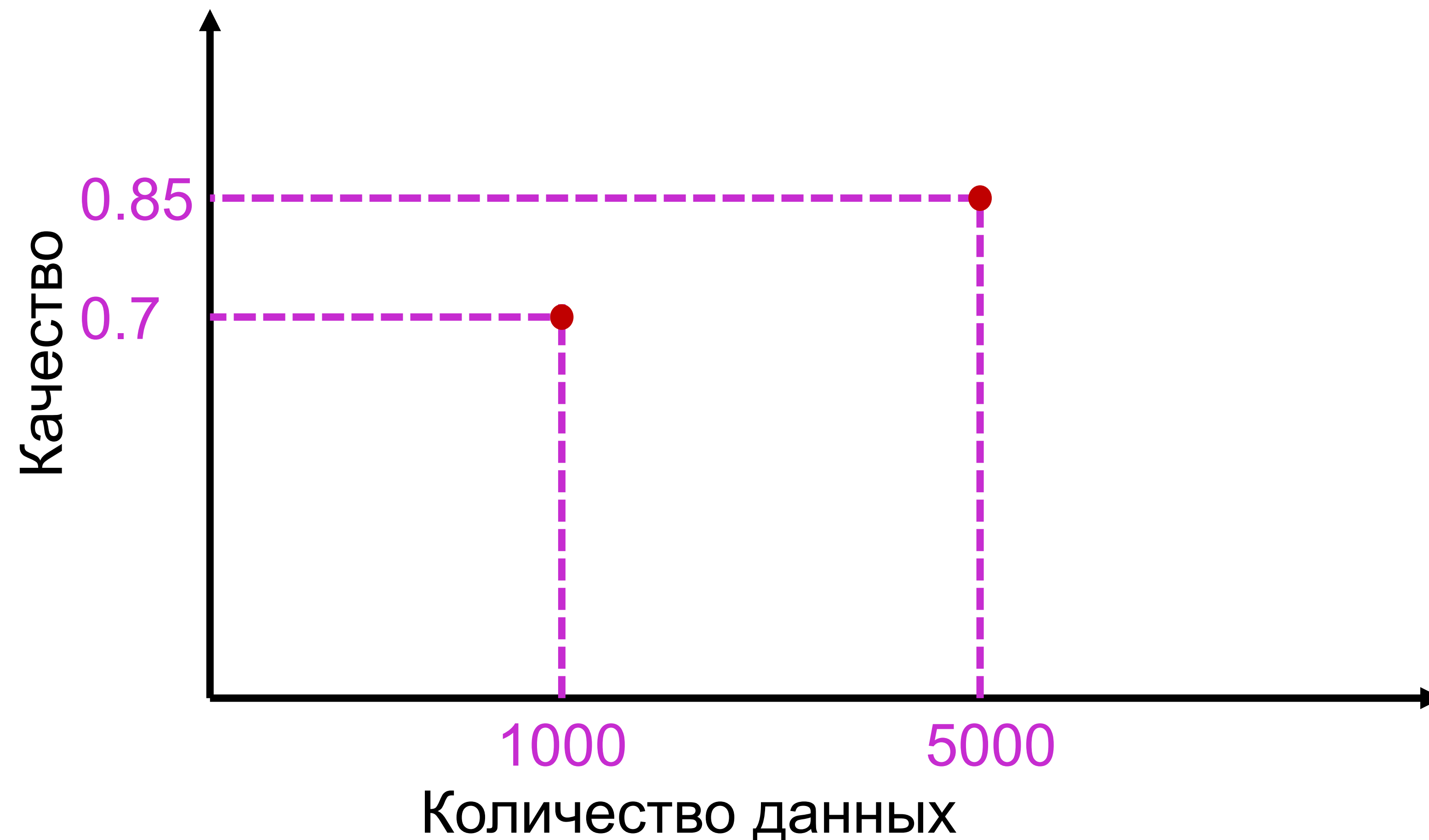
Шаг 2. Разметим 5000 изображений



Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

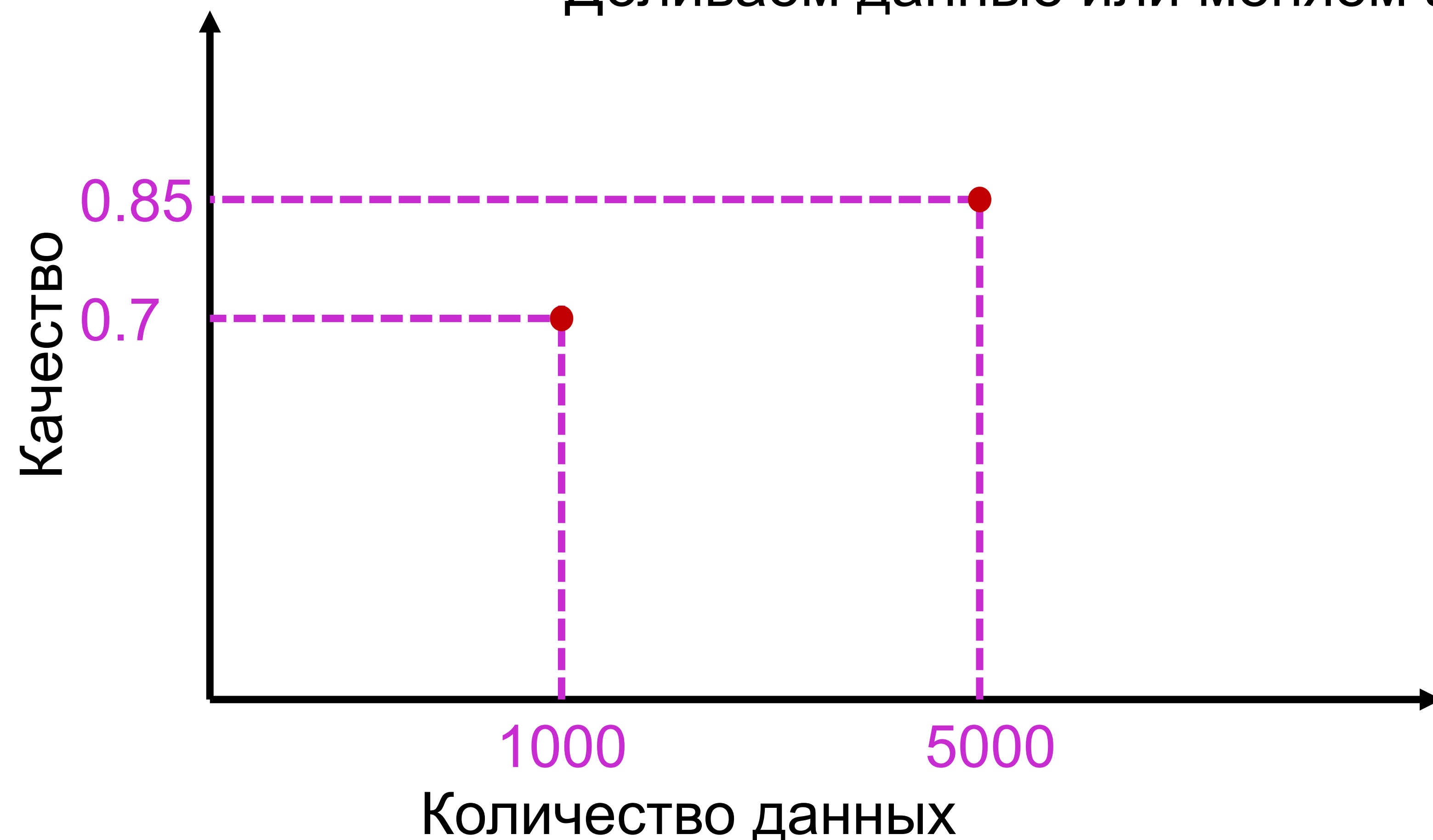


Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

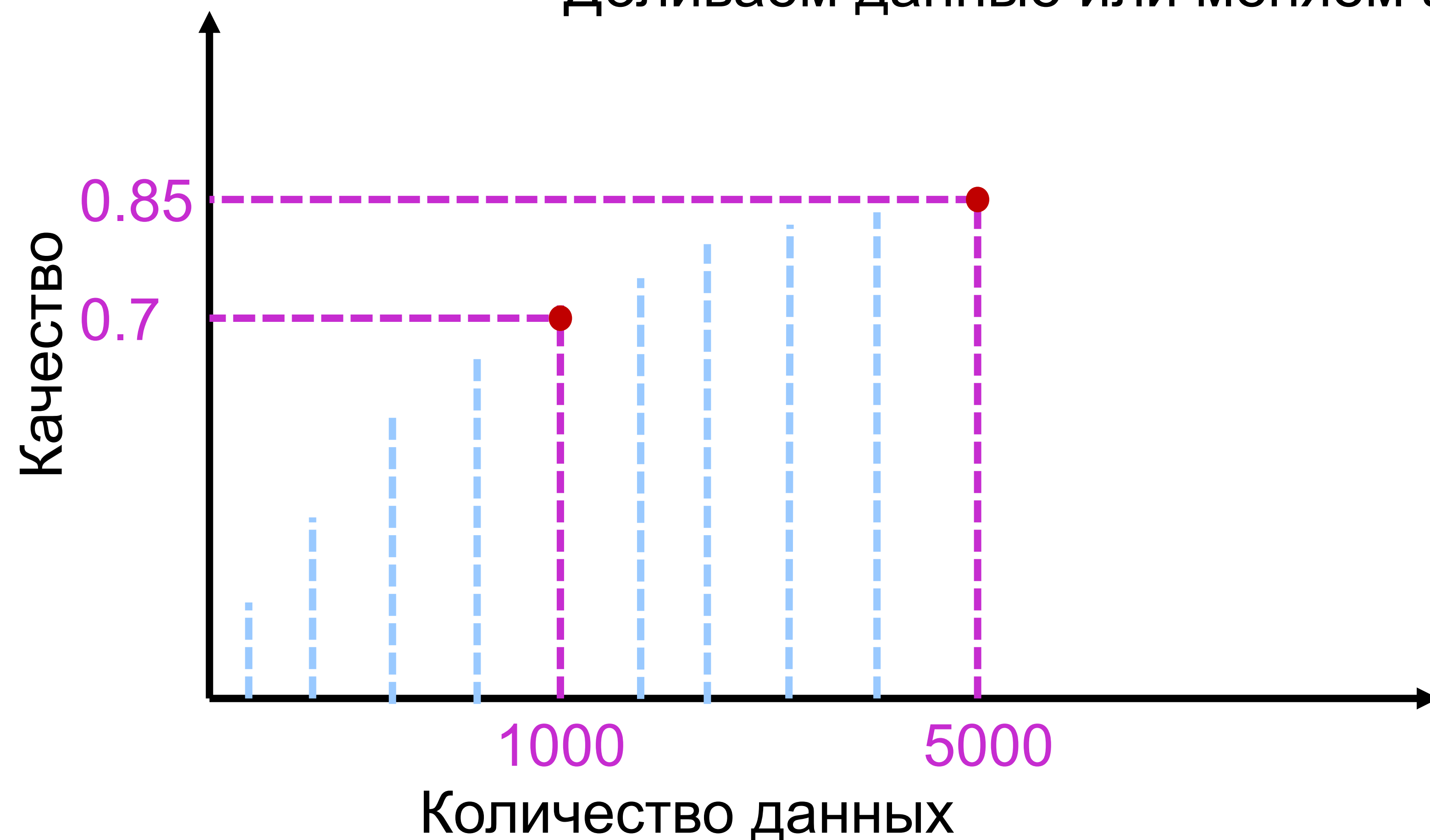


Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

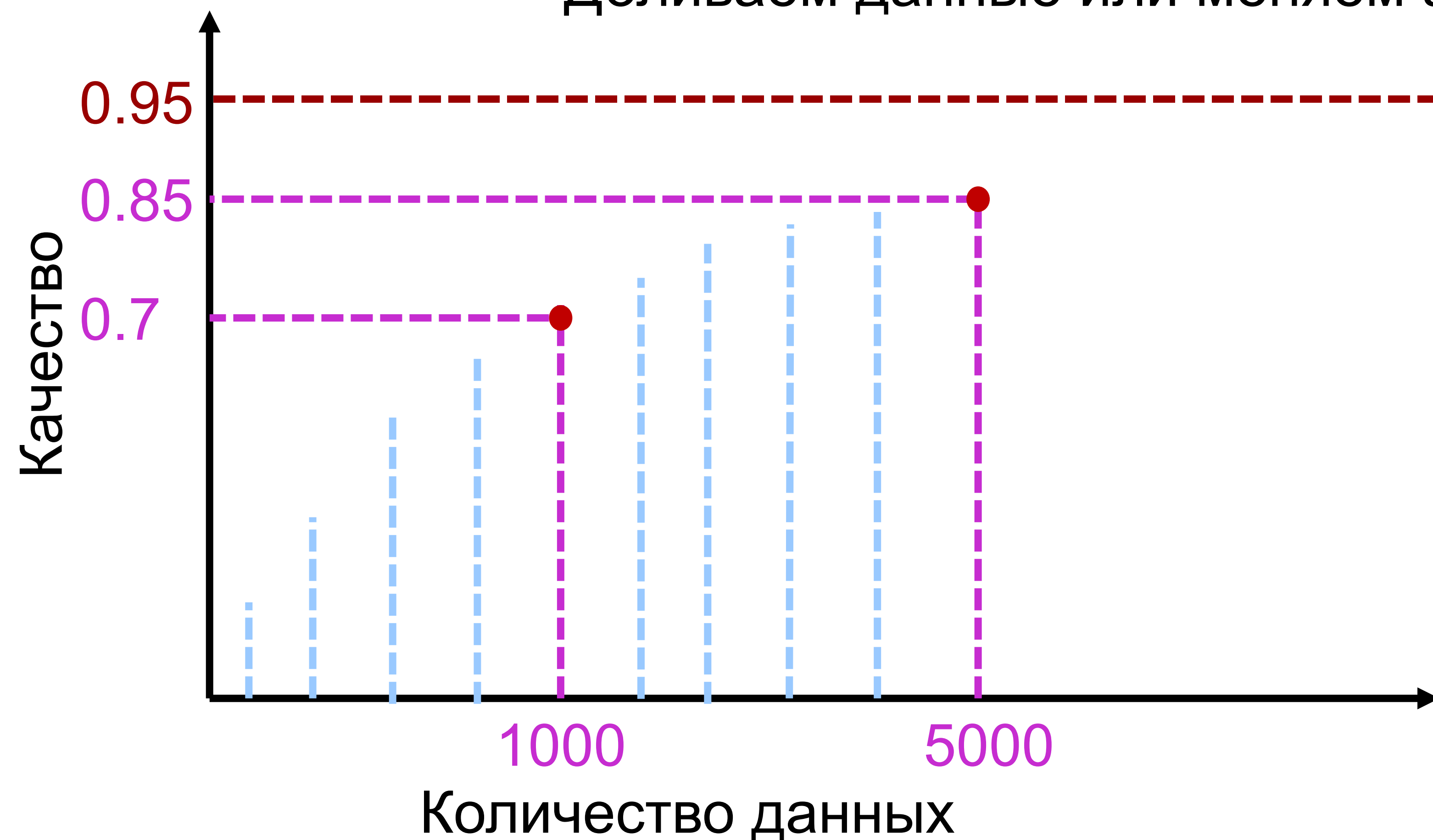


Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

Доливаем данные или меняем архитектуру?

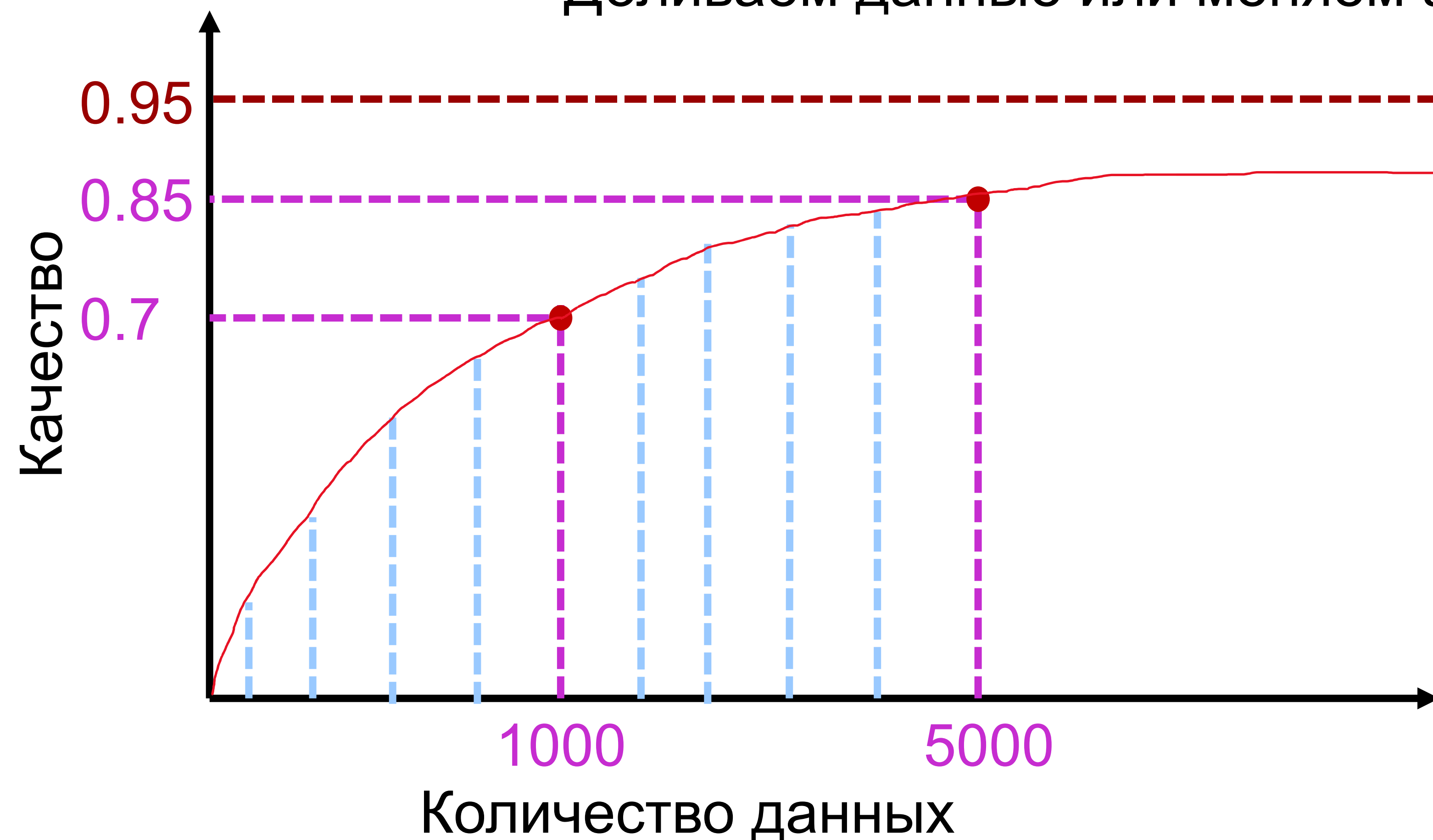


Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений

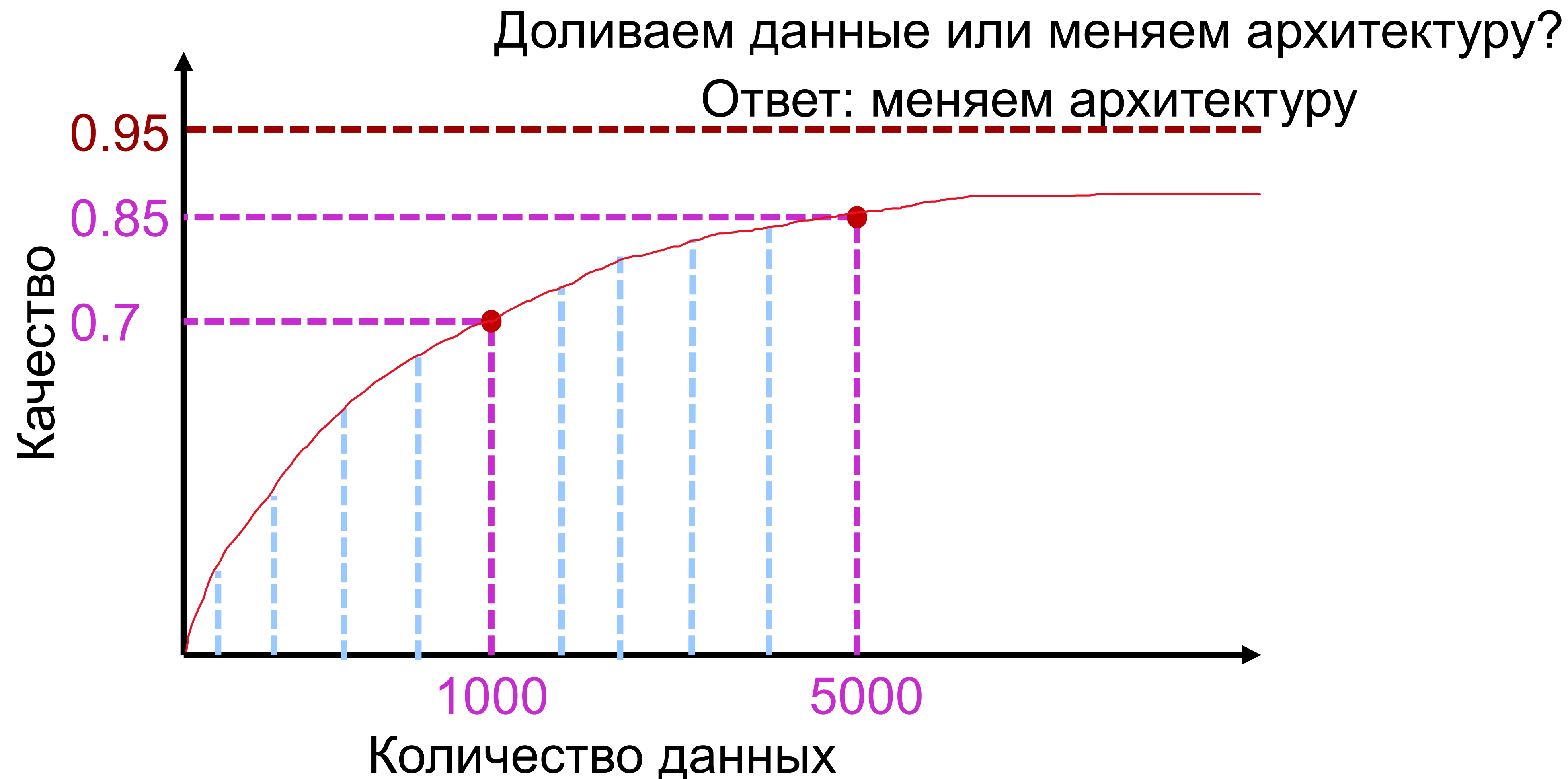
Доливаем данные или меняем архитектуру?



Шаг 1. Разметим 1000 изображений

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

Шаг 2. Разметим 5000 изображений



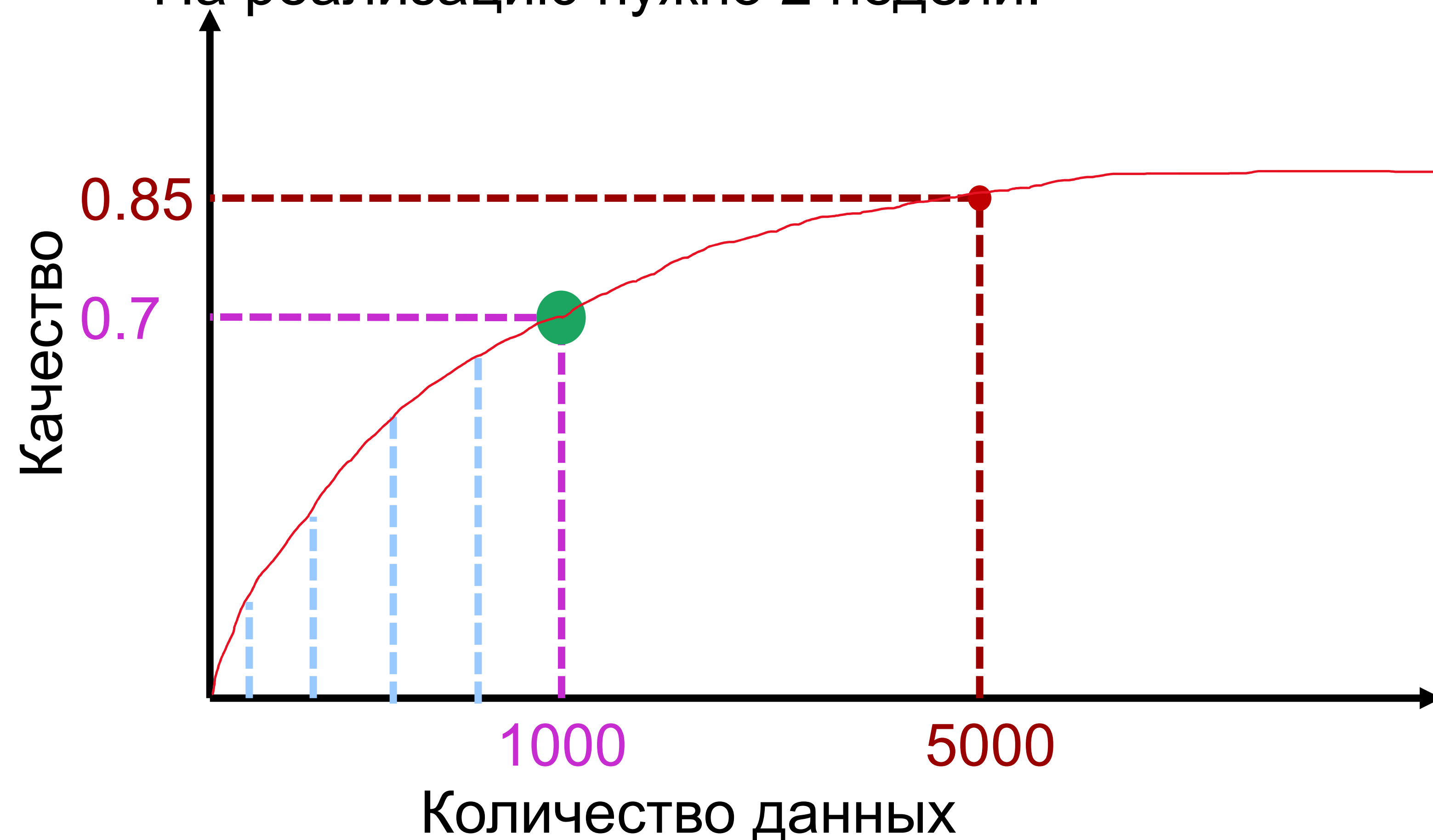
Кейс 2:

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.

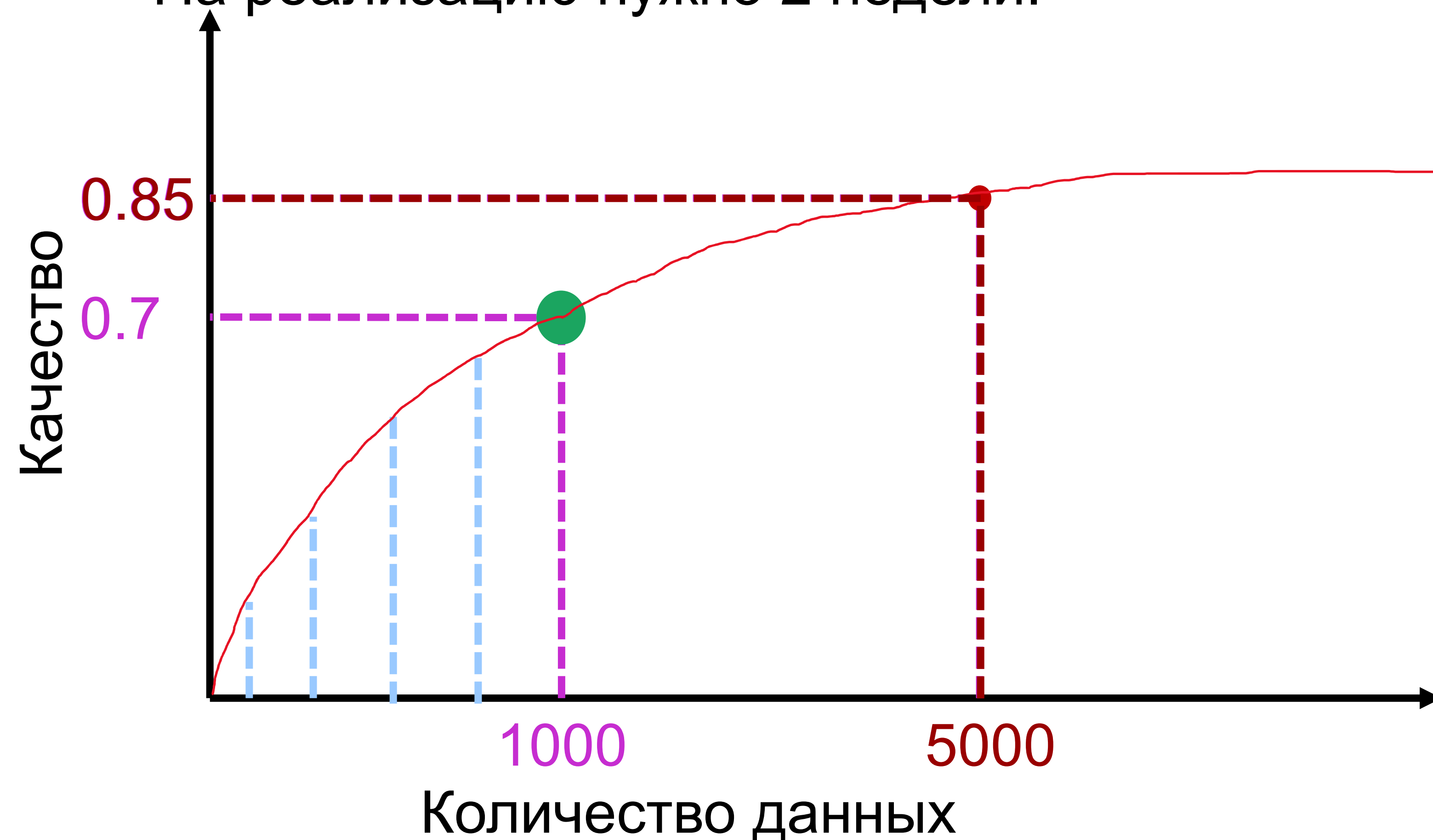
Уверен, что получится достичь 0.85.

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.

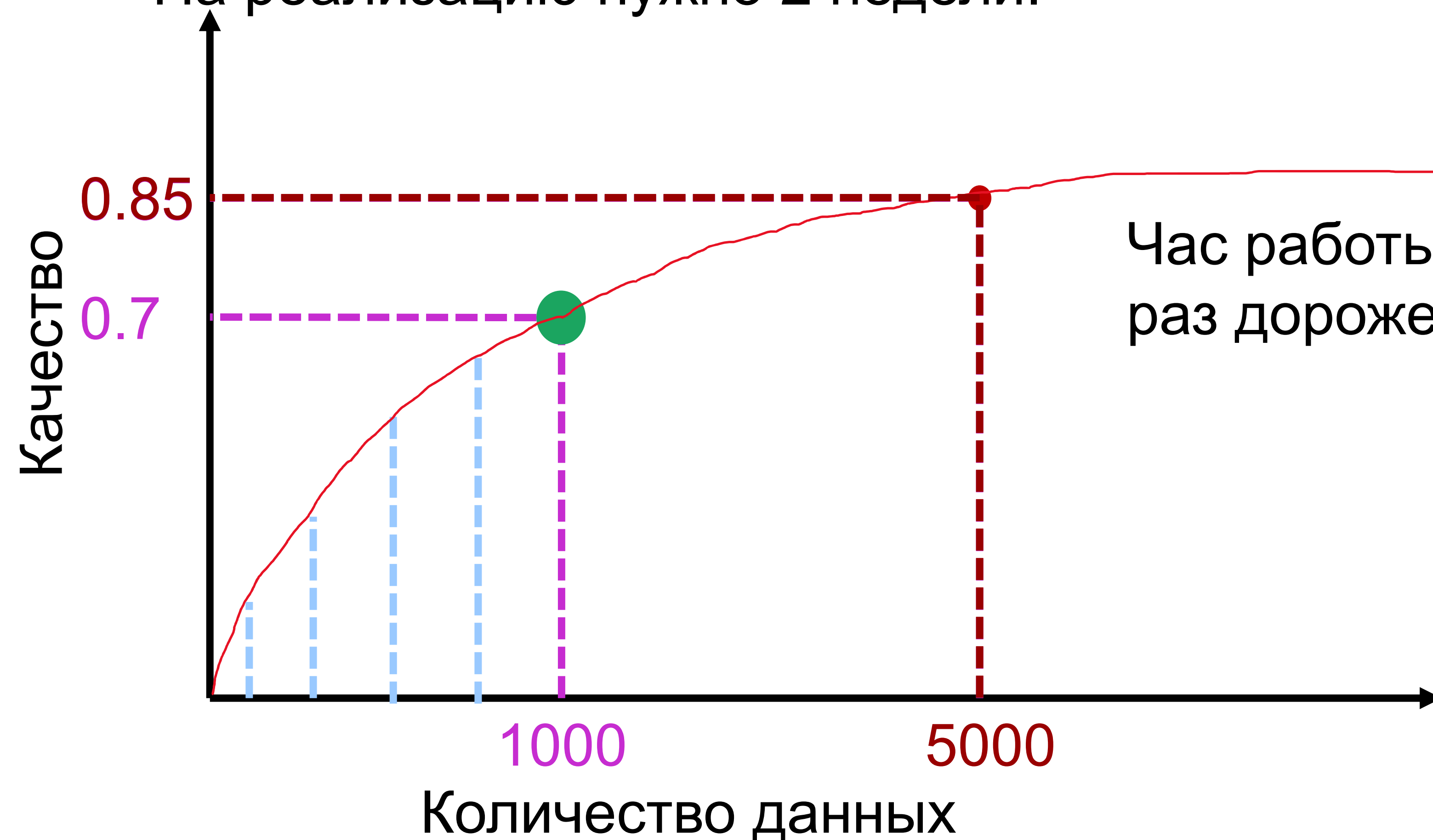
Уверен, что получится достичь 0.85.

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.

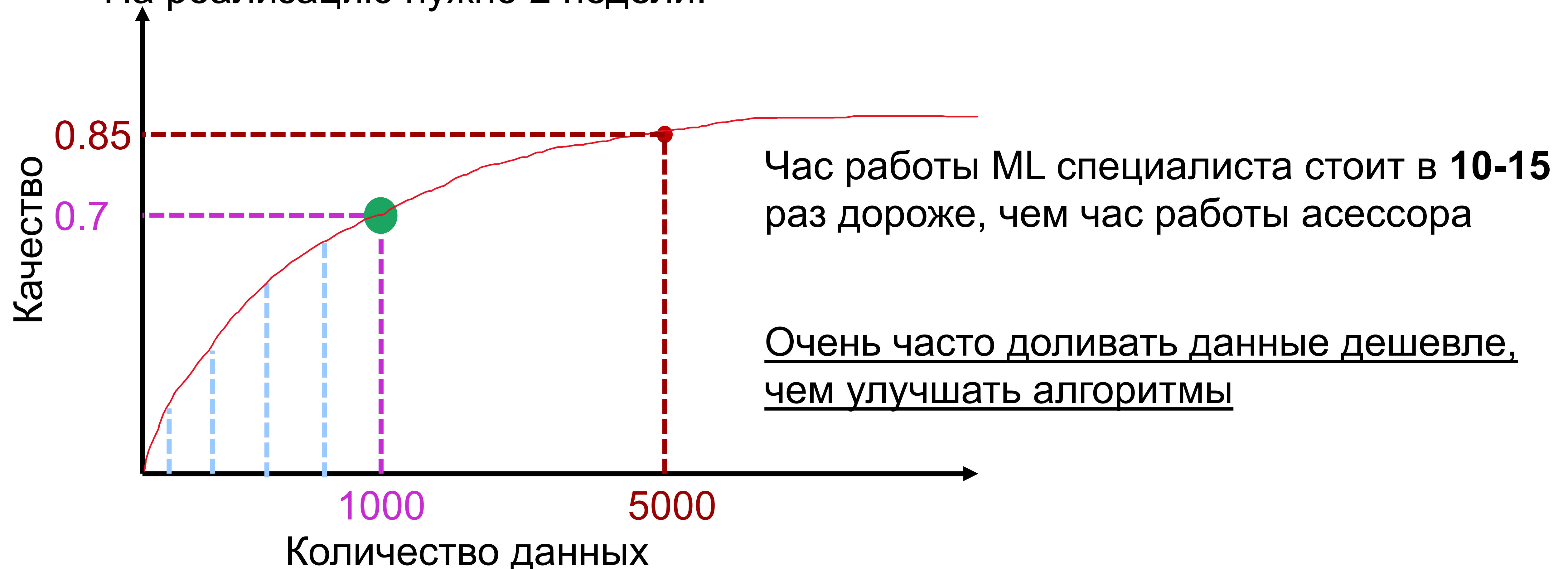
Уверен, что получится достичь 0.85.

На реализацию нужно 2 недели.

Варианты:

1) Вася реализует новый алгоритм

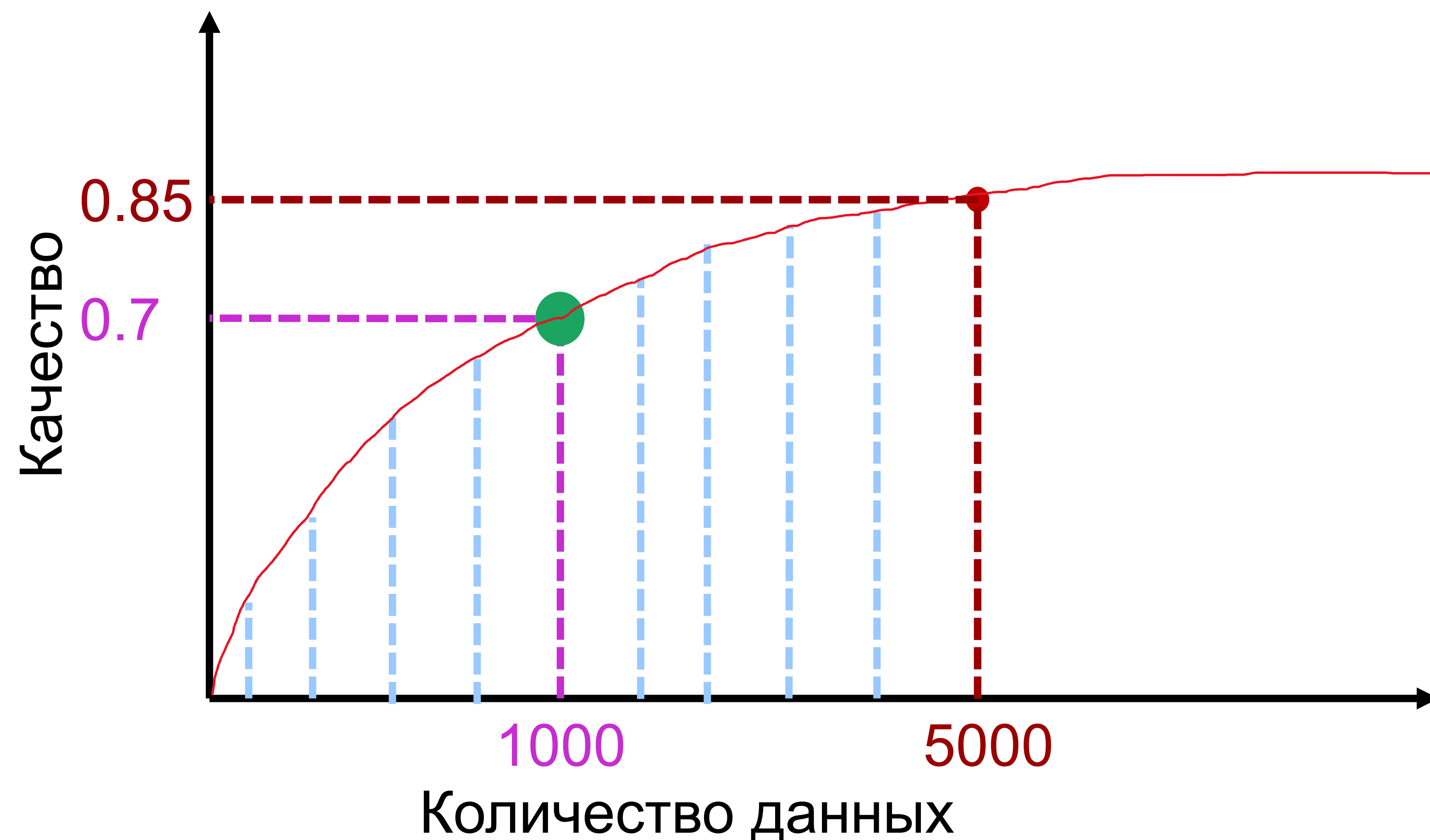
2) Размечаем больше данных



Кейс 3:

Сейчас имеем 0.7, нужно 0.85.

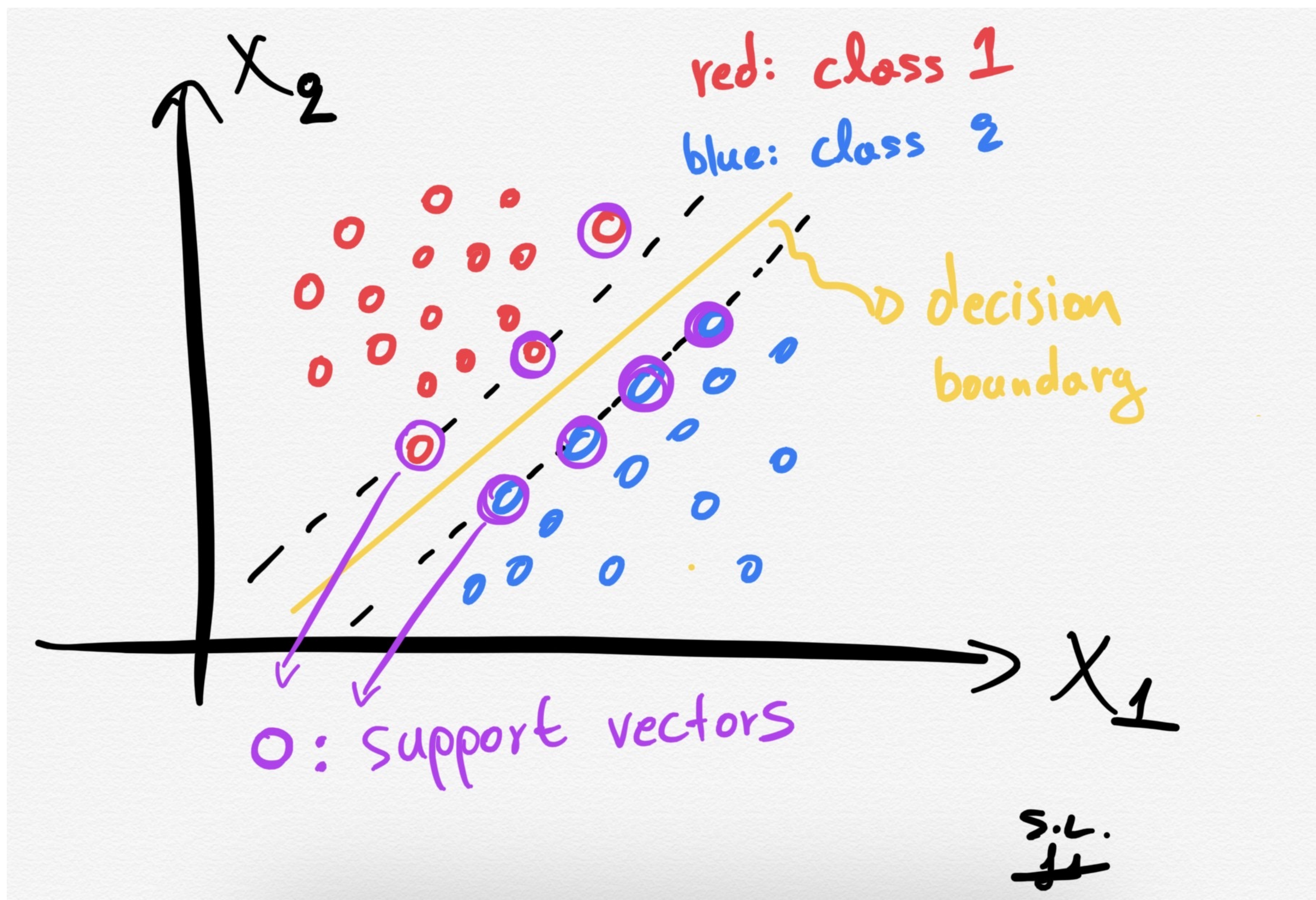
Можем ли мы меньше, чем за 5000 картинок достичь качества 0.85?



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: [http://neerc.ifmo.ru/wiki/index.php?title=Активное обучение](http://neerc.ifmo.ru/wiki/index.php?title=Активное_обучение)

Есть различные методики:

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

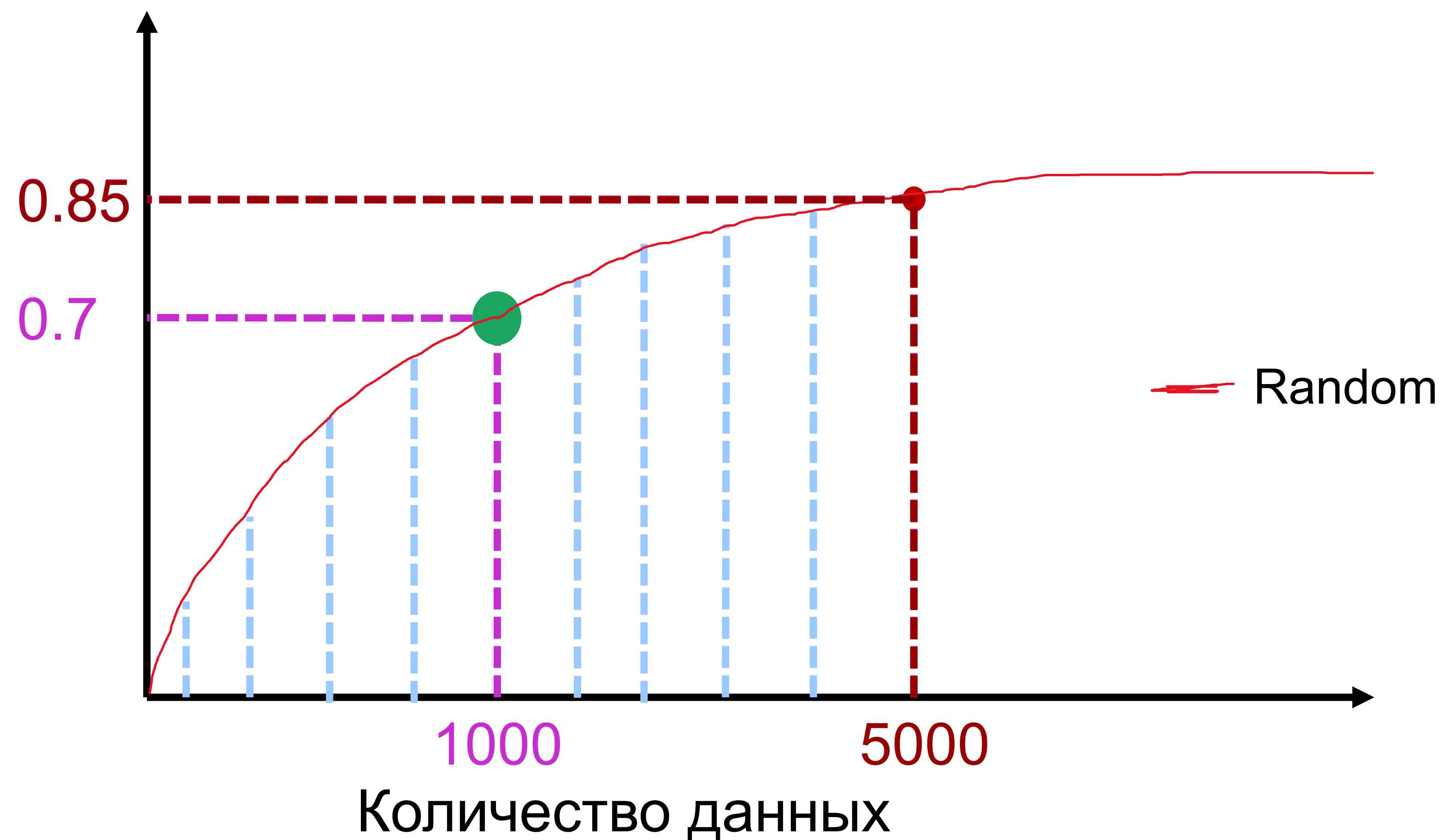
Active Learning

Неразмечанные данные прогоняются через алгоритм машинного обучения.

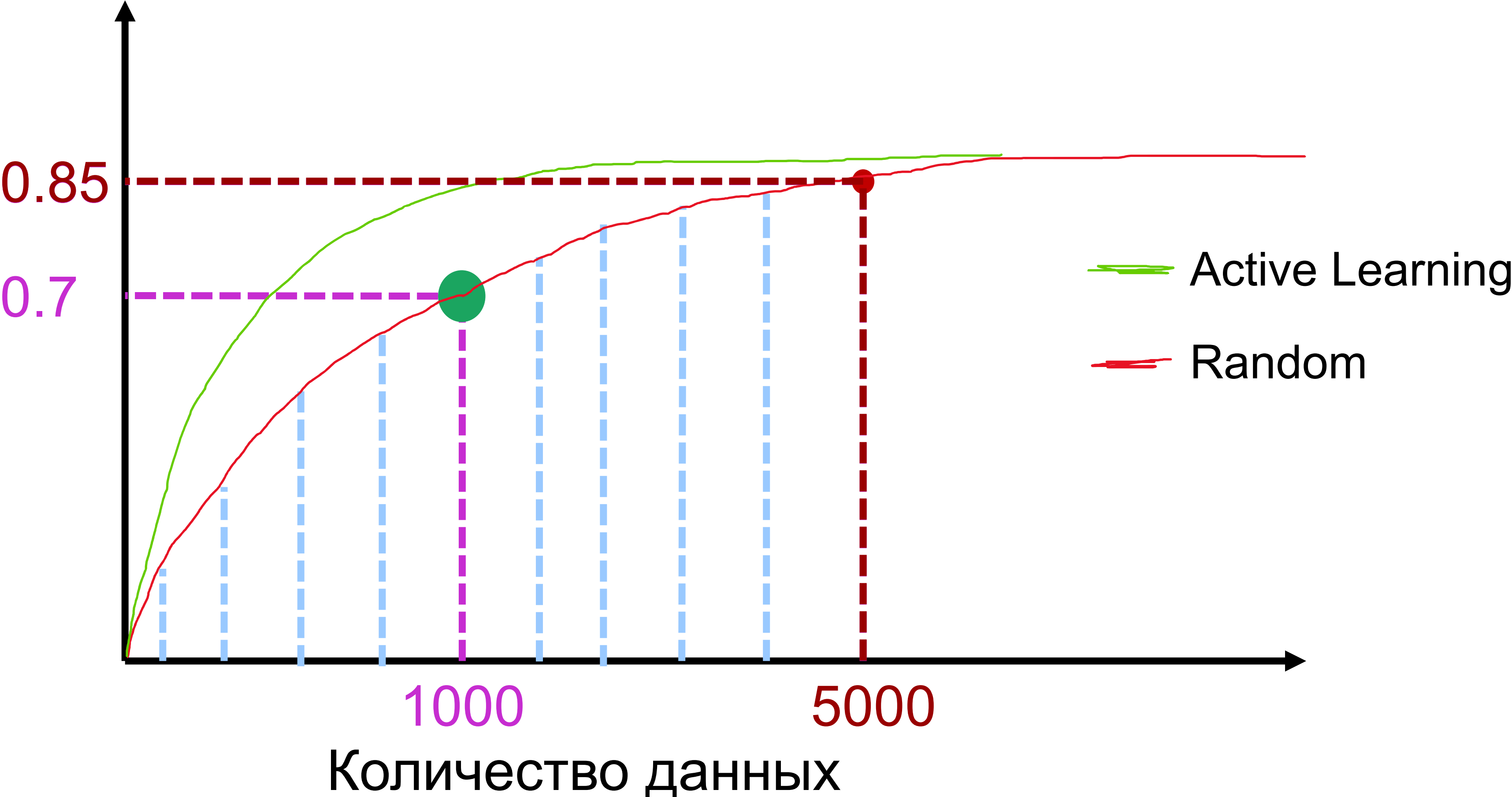
Сначала размечаются данные с низкой уверенностью в ответе.
Как правило они оказываются сложными и алгоритм с ними плохо справляется.

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

Active Learning



Active Learning



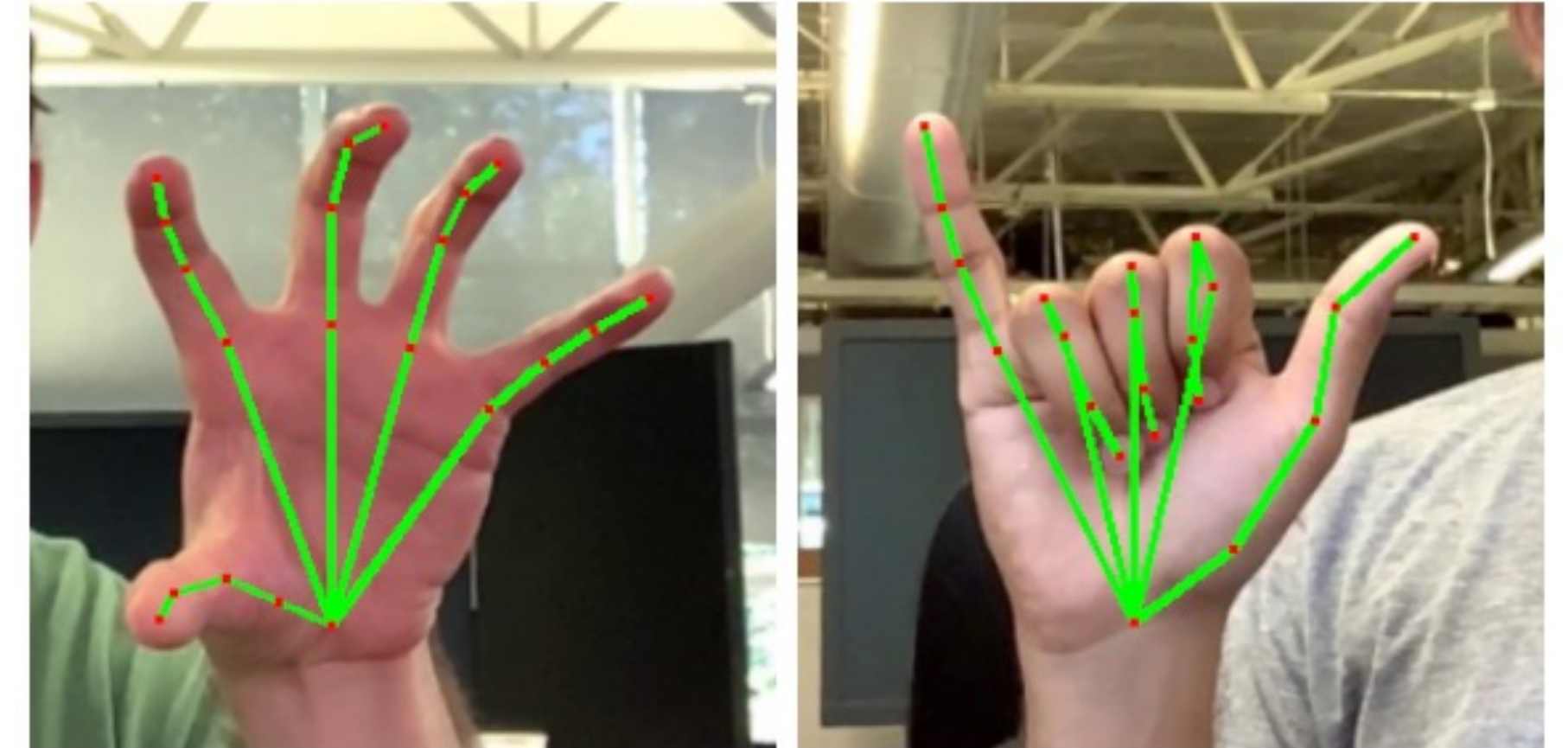
Synthetic datasets



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
- Проблемы с переносом на основной домен
- Обычно хорошо заходит СМЕСЬ синтетики и реальных данных
- Синтетику можно за очень дешево очень много нарендерить

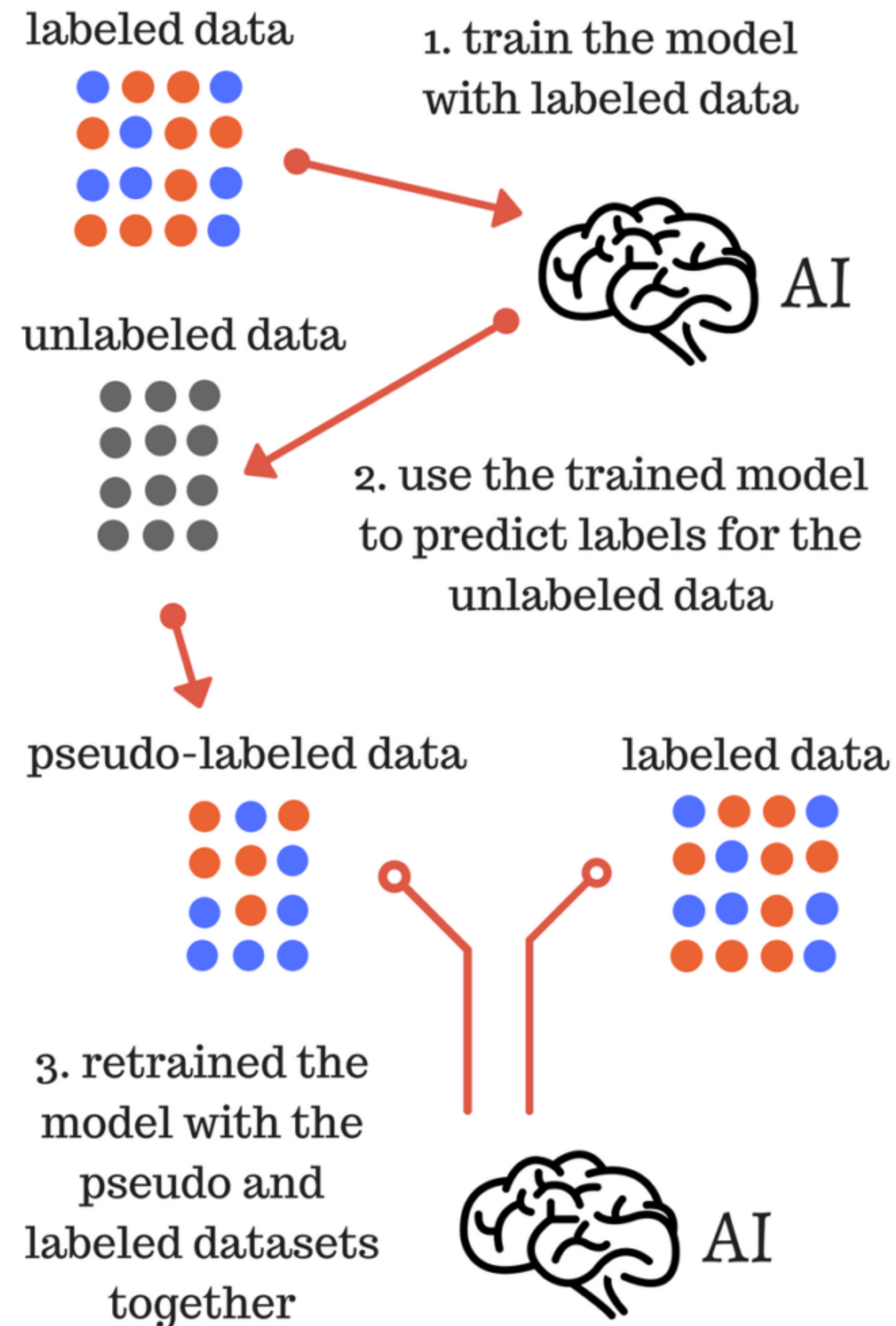


Synthetic datasets



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



Cat

Dog

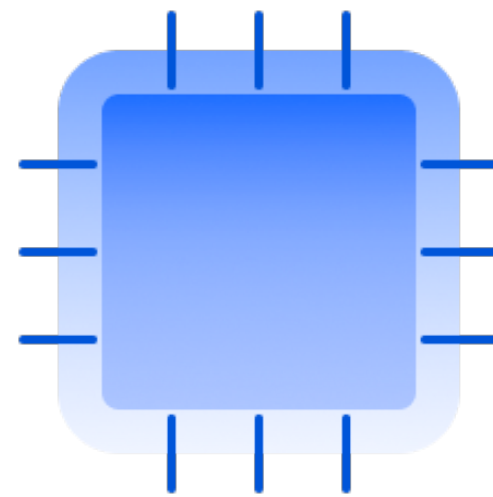
Cat



Dog

Cat

Cat



Dog

Cat

Cat

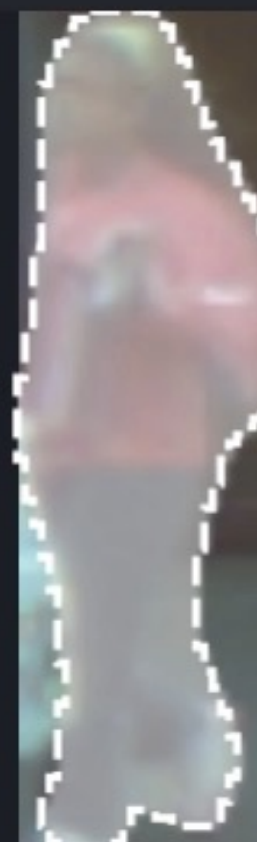
Поиск ошибок в разметке с помощью ML модели



Error type: Low iou
iou: 88.53%



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



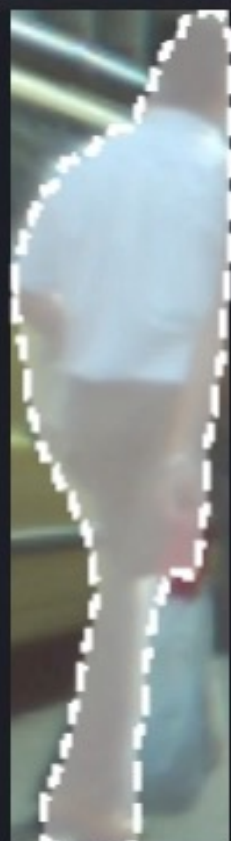
Error type: Missing label
Predicted: Pedestrain (97.23%)



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



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



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



Error type: Low iou
iou: 87.28%



Error type: Low iou
iou: 86.96%



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

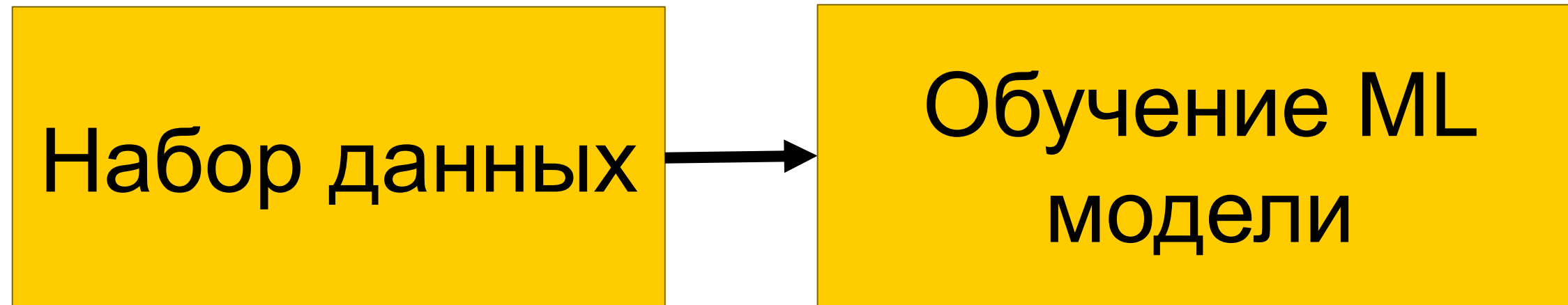


Система фиксации
нарушений

Human in the Loop

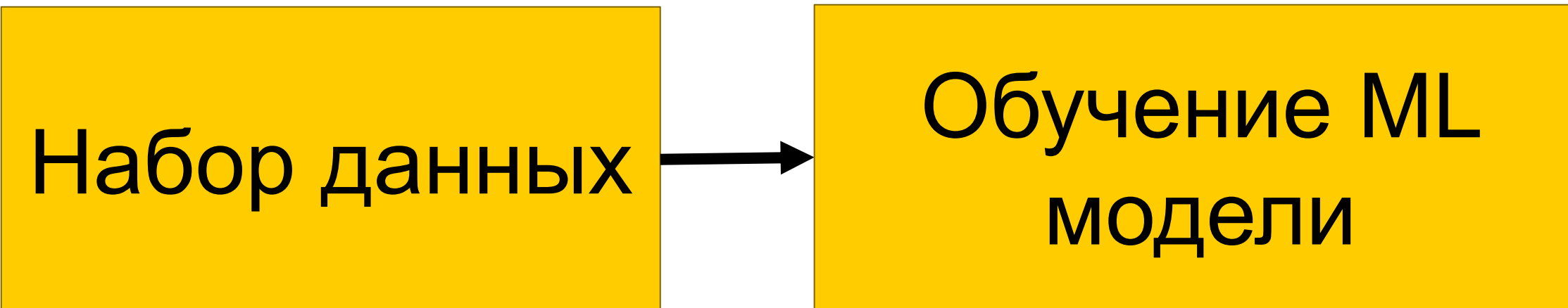
Набор данных

Human in the Loop

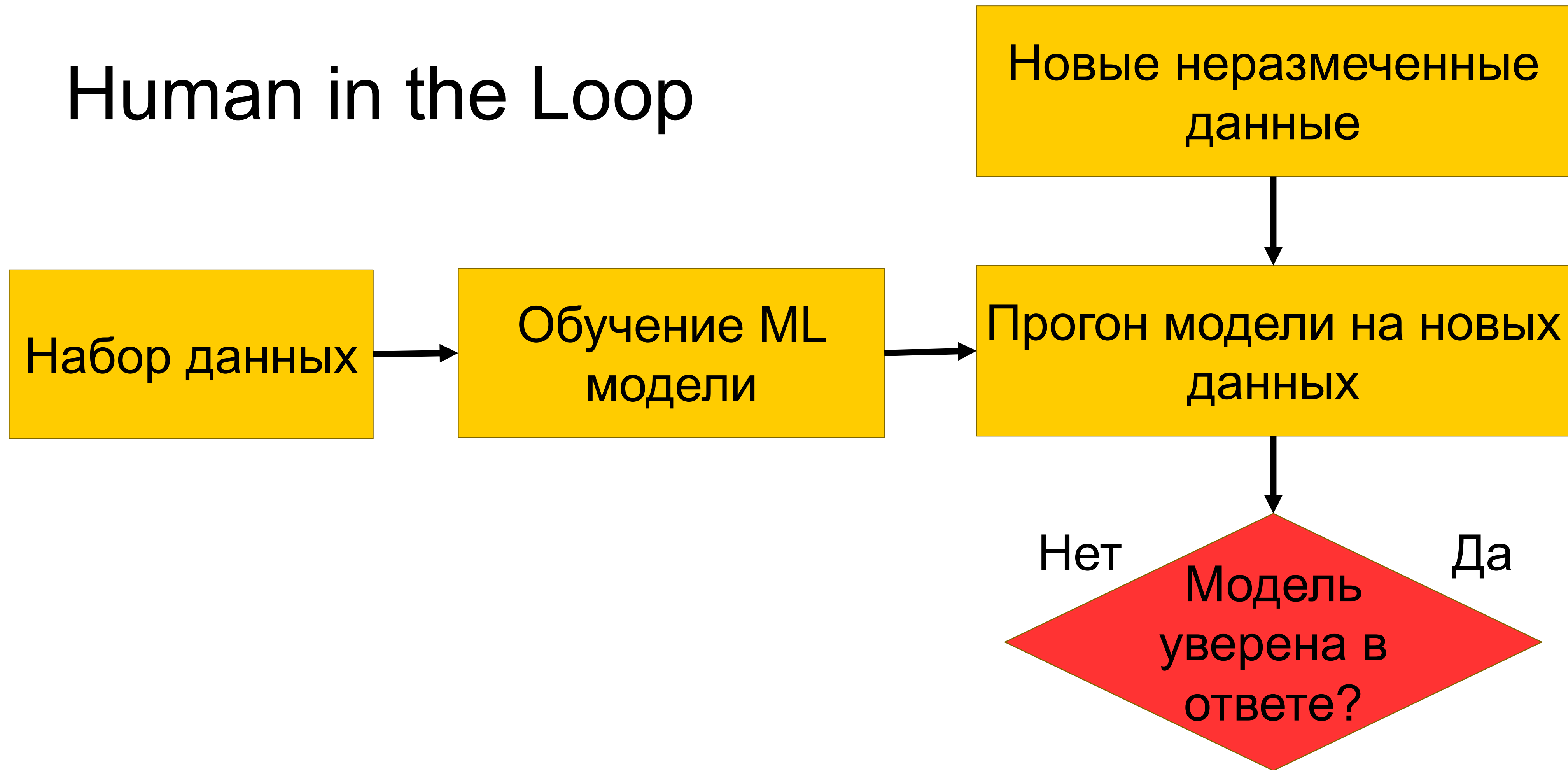


Human in the Loop

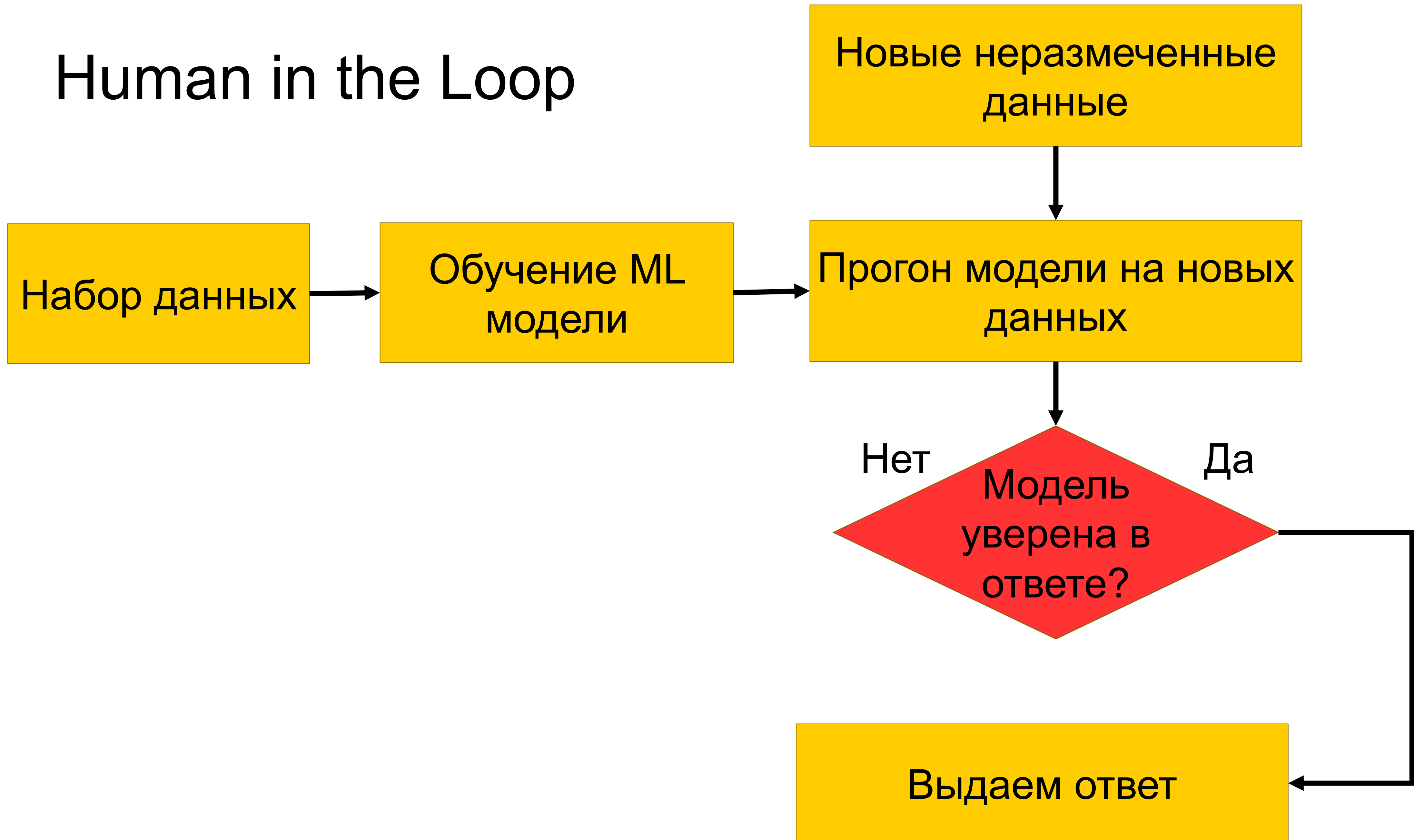
Новые неразмеченные
данные



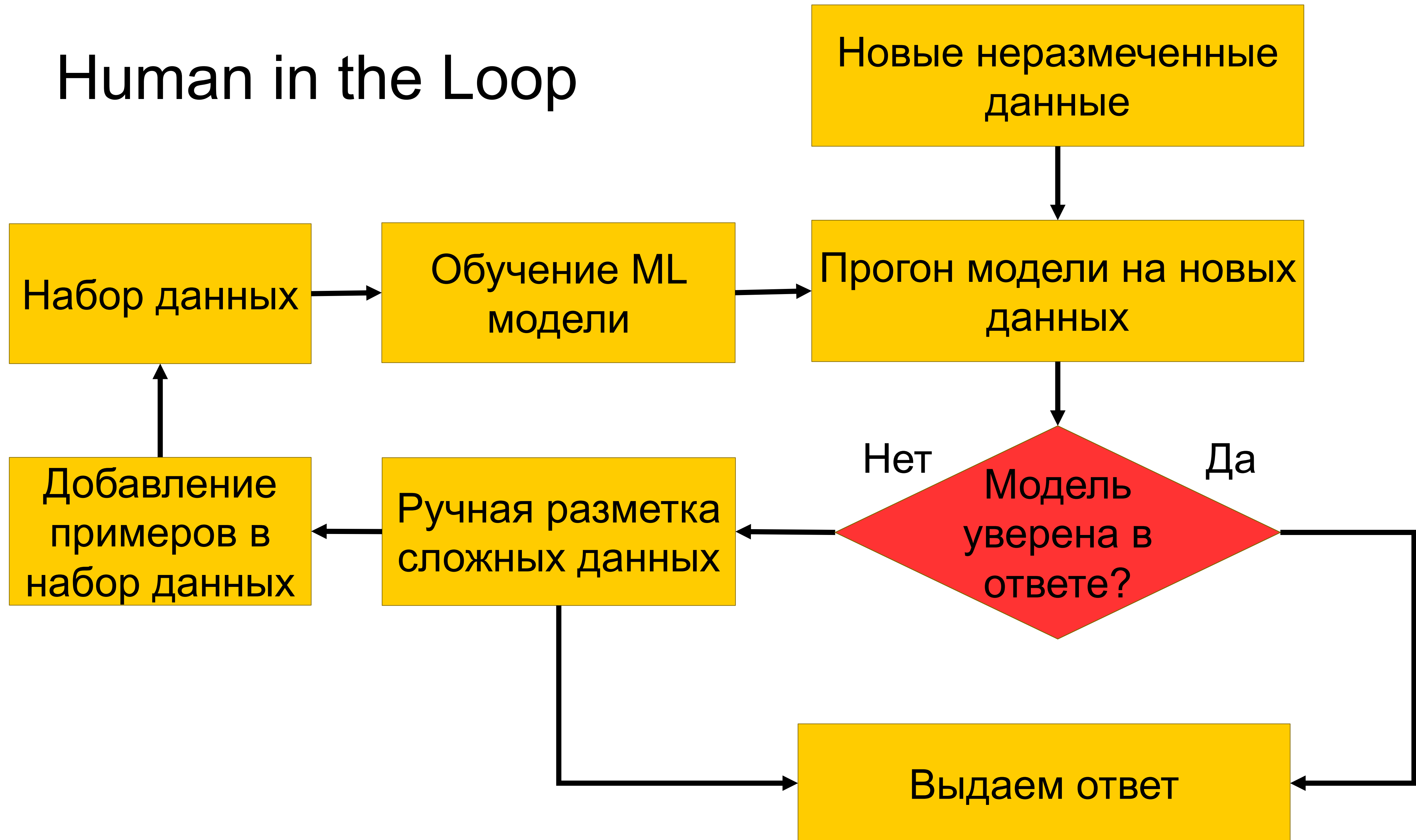
Human in the Loop



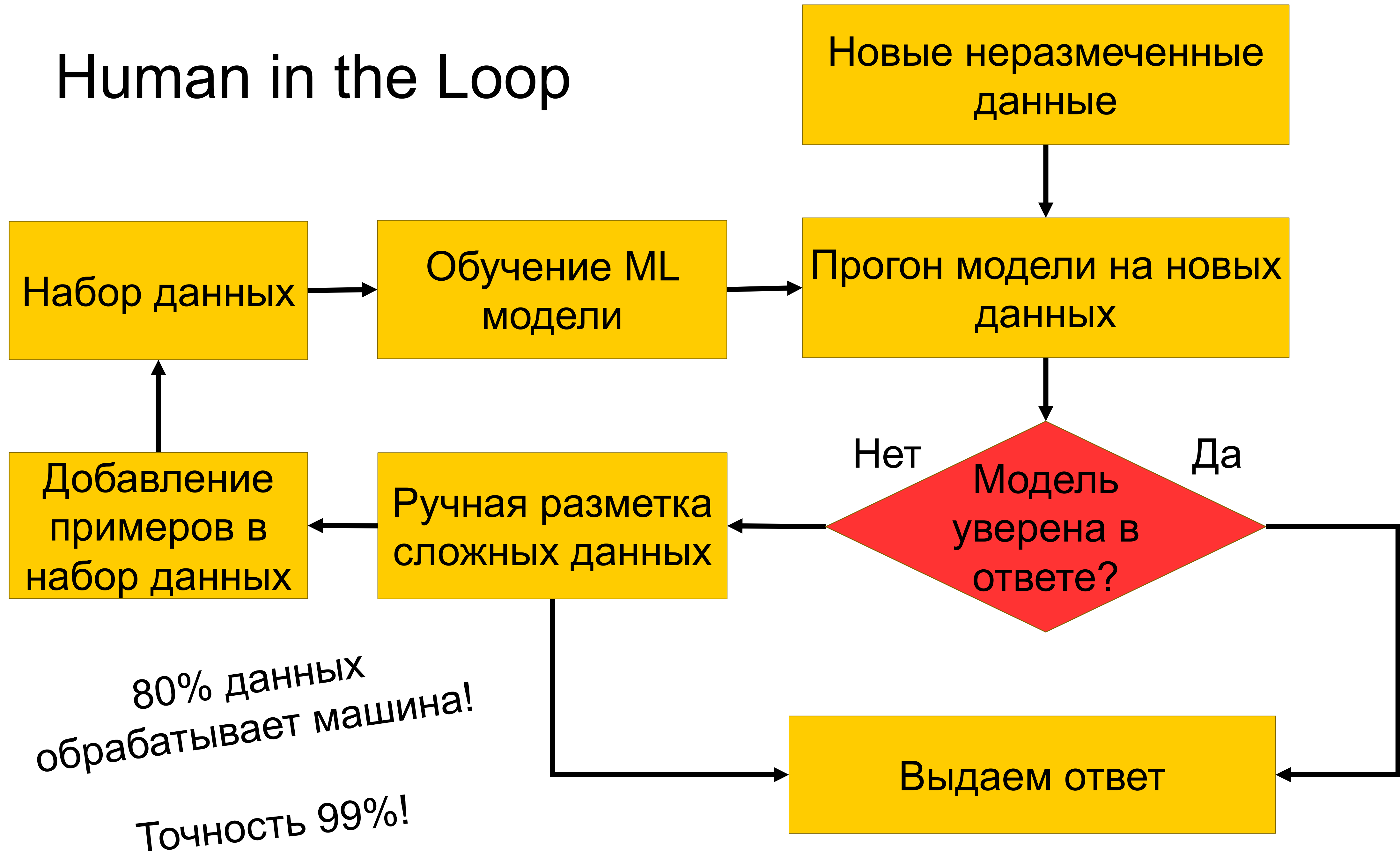
Human in the Loop



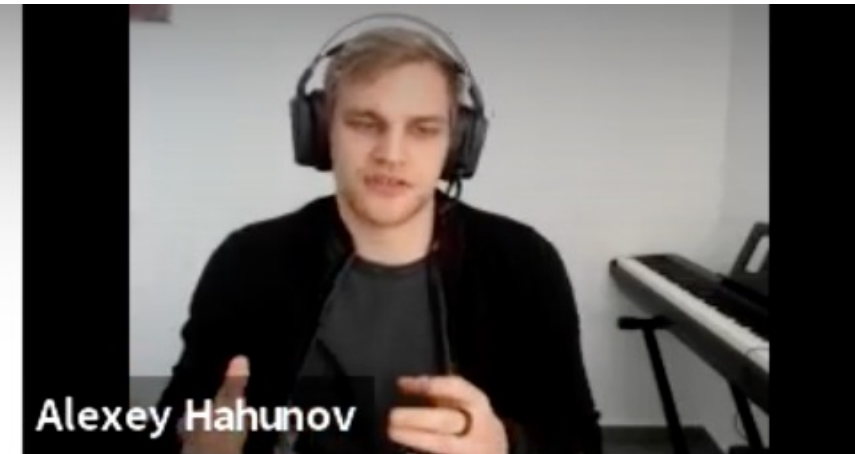
Human in the Loop



Human in the Loop



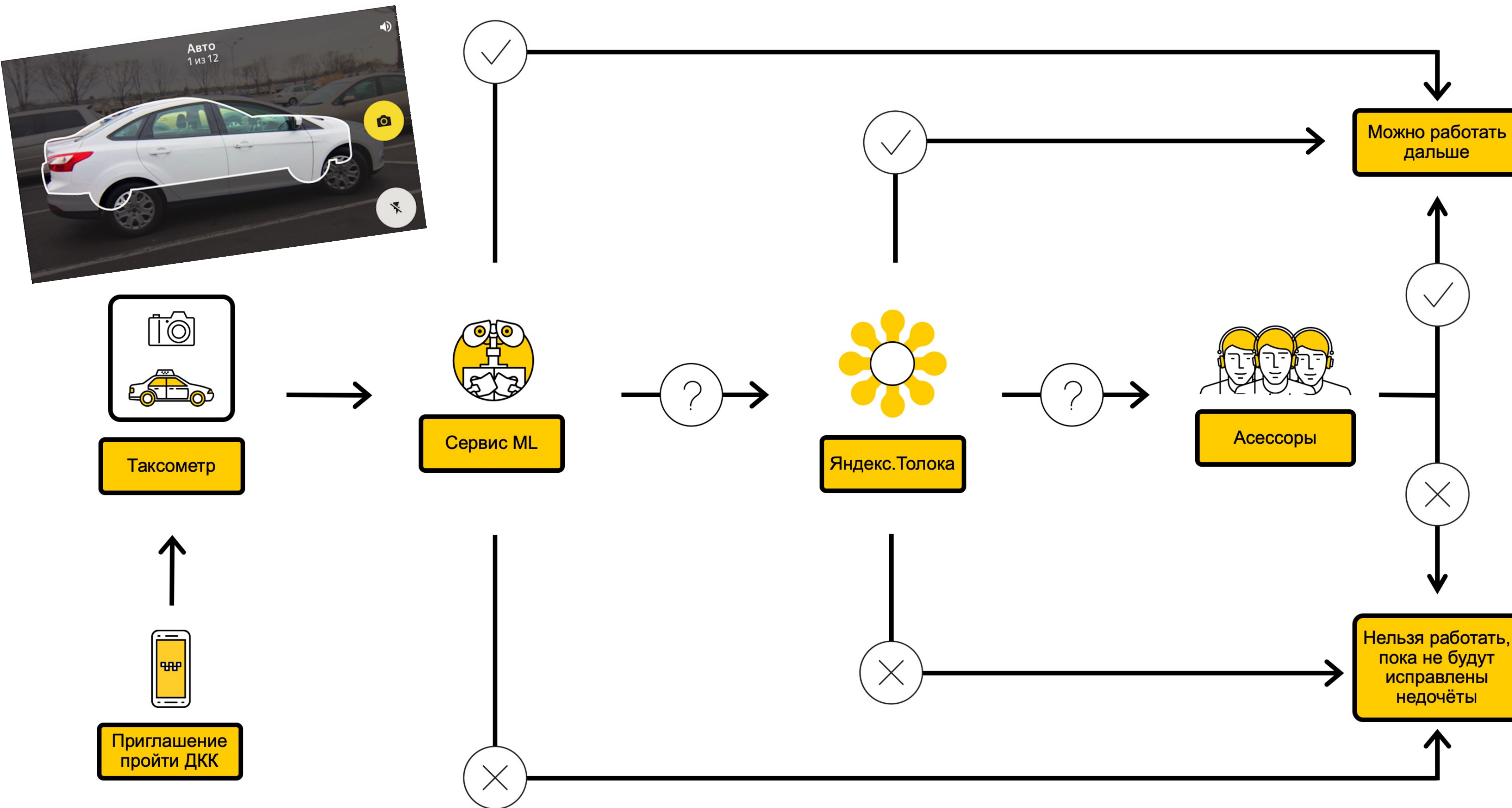
Human in the Loop в Dbrain



Отдельные сервисы для решения любой задачи



Human in the Loop в Яндекс.Такси



Выводы

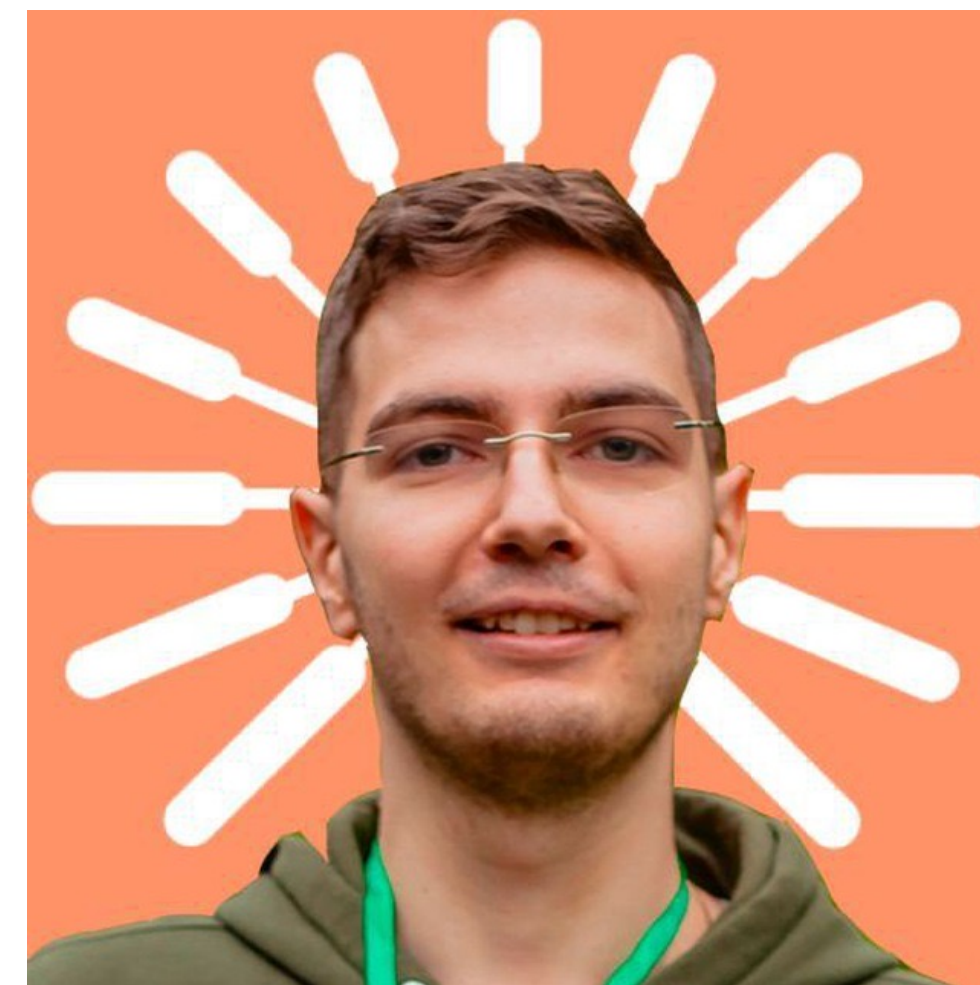
- Существует множество методов и техник, которые позволяют увеличить скорость и качество разметки
- Данные техники требуют времени на внедрение и настройку, но позволяют сильно снизить стоимость разметки

Тренды в разметке данных

1. Автоматизация разметки
2. Внедрение разметки в пайплайн разработки ML продукта
3. Увеличение объема размечаемых данных
4. Использование узких специалистов



Я: Куцев Роман
t.me/roman_kusev



ТГ канал: Рома ❤️ Толоку
t.me/roma_toloka

