Проектирование сервиса для разметки данных

Product Requirements Document (PRD)

Описание продукта

Что представляет собой продукт

Сервис разметки данных предназначен для помощи компаниям, занимающимся разработкой моделей машинного обучения, в организации процесса разметки больших объемов данных. Сервис позволяет управлять проектами разметки, обеспечивать высокое качество разметки, проводить обучение и проверку разметчиков.

Какую проблему и для кого он решает

Продукт решает проблему сложности и трудоемкости ручной разметки данных для машинного обучения, предоставляя удобный инструмент для Заказчиков и разметчиков. Основными пользователями сервиса являются:

- Компании, занимающиеся машинным обучением, которым необходимо быстро и качественно размечать большие объемы данных.
- Заказчики, управляющие процессом разметки.
- Разметчики, выполняющие задания по разметке данных.

Цели и задачи проекта

- Цели:
    - Обеспечить эффективный процесс разметки данных.
    - Повысить качество разметки данных.
    - Обеспечить гибкость и масштабируемость процесса разметки.
- Задачи:
    - Разработка удобного интерфейса для управления проектами разметки.
    - Создание инструментов для разметки данных (текст, изображения). ● Обеспечение системы мониторинга и проверки качества разметки. ● Разработка системы онбординга и обучения разметчиков.

Требования клиента

Как вы поняли требования клиента

- Система должна поддерживать создание и управление проектами разметки. ● Система должна предоставлять инструменты для разметки данных. ● Система должна обеспечивать возможность мониторинга и проверки качества работы разметчиков.
- Система должна включать механизм онбординга и обучения разметчиков. ● Система должна автоматически перераспределять задачи при необходимости.

Анализ конкурентов

1. Labelbox

Основные функции:

- Инструменты для разметки данных: Поддержка различных типов данных, включая изображения, текст и видео. Инструменты позволяют разметчикам выделять области, создавать аннотации и метки.
- Управление проектами: Возможность создания и управления проектами разметки, назначения задач и отслеживания прогресса.
- Контроль качества: Механизмы для проверки и ревизии разметки, включая автоматическую проверку качества и консенсус между разметчиками. ● Интеграции: Поддержка интеграции с платформами машинного обучения, такими как AWS, Google Cloud и другие.
- Совместная работа: Функции для командной работы, включая комментарии и совместное редактирование.

Преимущества:

- Широкий набор инструментов для разметки данных.
- Поддержка различных типов данных.
- Мощные функции контроля качества.
- Хорошая интеграция с платформами машинного обучения.

Недостатки:

- Высокая стоимость для малых и средних компаний.
- Могут быть сложности с освоением системы для новичков.
- Ограниченная функциональность для работы с большими видеофайлами.

2. Scale AI

Основные функции:

- Высококачественная разметка данных: Поддержка различных типов данных, включая изображения, текст и 3D-данные. Автоматизация части процессов разметки.
- Контроль качества: Системы проверки качества, включающие машинное обучение для обнаружения ошибок в разметке и улучшение качества данных. ● Управление проектами: Создание и управление проектами разметки, назначение задач разметчикам и отслеживание выполнения.
- Гибкость и масштабируемость: Возможность быстро масштабировать объемы разметки в зависимости от потребностей.
- Интеграции: Поддержка интеграций с основными платформами для машинного обучения и данными.

Преимущества:

- Высокое качество разметки данных благодаря автоматизации.
- Поддержка сложных типов данных, таких как 3D и видео.
- Хорошие возможности масштабирования.
- Надежные механизмы контроля качества.

Недостатки:

- Высокая стоимость услуг.
- Ограниченная кастомизация процессов разметки.
- Зависимость от автоматизации может не всегда соответствовать специфическим требованиям.

Требования к продукту на основе анализа конкурентов и

требований клиента

Что обязательно должен уметь ваш продукт:

- Поддержка различных типов разметки данных: Разметка изображений, текста, аудио.
- Управление проектами: Возможность создания проектов, управления задачами, назначения разметчиков и отслеживания выполнения.
- Инструменты для разметки: Удобные и функциональные инструменты для разметки данных.
- Контроль качества: Механизмы проверки и ревизии разметки, включая онбординг разметчиков и контроль качества.
- Совместная работа: Поддержка комментариев и совместного редактирования.
- Автоматическое перераспределение задач: Возможность автоматического распределения задач между разметчиками при необходимости.

Out of scope – Что пока не должен уметь ваш продукт

- Поддержка разметки видео.
- Полная автоматизация разметки данных.
- Интеграция с внешними платформами машинного обучения.

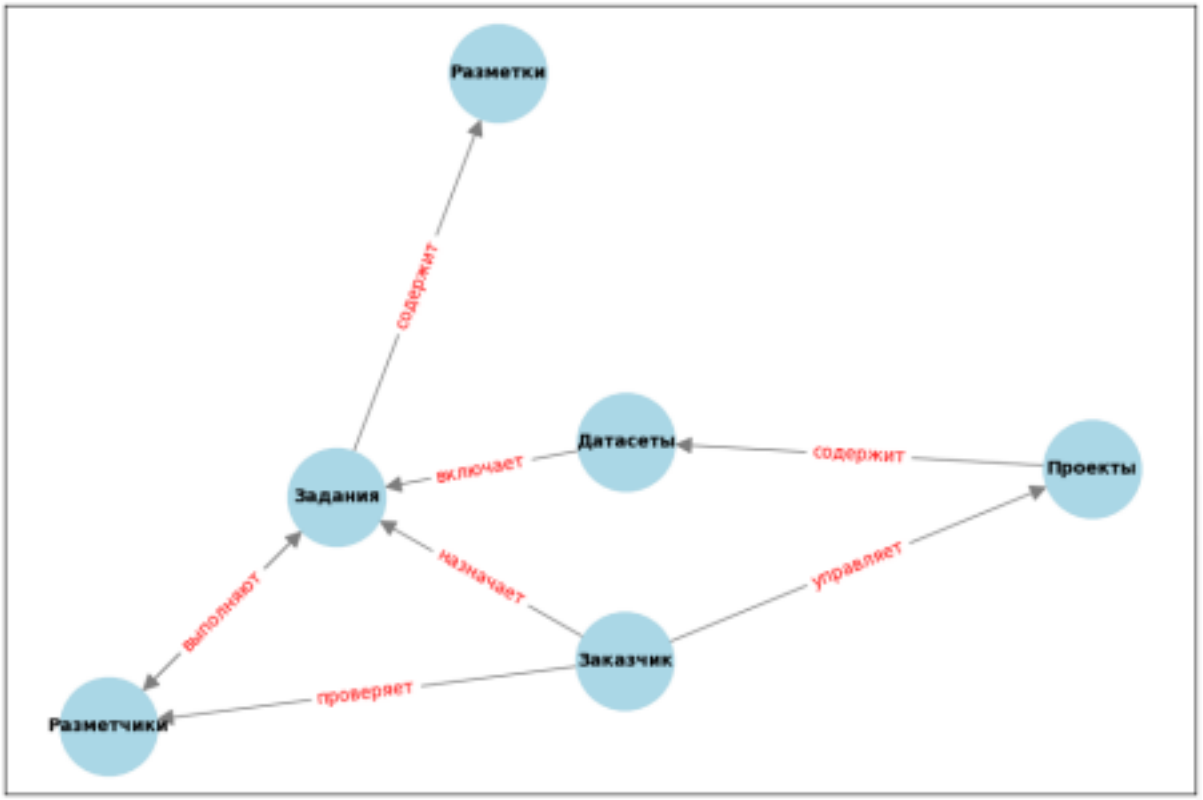
Описание разрабатываемой системы

Роли пользователей и их потребности

1. Заказчик:
    - Управление пользователями и проектами.
    - Настройка заданий и мониторинг выполнения.
    - Проведение онбординга и проверки разметчиков.
    - Скачивание размеченных датасетов.
2. Разметчик:
    - Просмотр и выполнение заданий по разметке.
    - Сохранение и отправка результатов.
    - Прохождение онбординга и обучения.

Концептуальная схема системы

- Пользователи (Заказчик, Разметчик)
- Сущности (Проекты, Задания, Датасеты, Разметки)
- Связи (Заказчик управляет Проектами, Разметчики выполняют Задания, Проекты содержат Датасеты, Датасеты содержат Разметки)



Жизненный цикл сущностей

1. Создание задания Заказчиком:
    - Заказчик создает новое задание в системе, добавляет описание, выбирает датасет для разметки, определяет требования и указывает коэффициент сложности (x2, x3, x4).
    - Заказчик также задает сроки выполнения задания и определяет критерии качества разметки.
2. Назначение задания разметчику:
    - Заказчик распределяет задания между разметчиками, учитывая их загруженность, квалификацию и текущую доступность.
    - Разметчики получают уведомление о назначенных им заданиях через систему.
3. Выполнение разметки разметчиком:
    - Разметчик приступает к выполнению задания, используя предоставленные инструменты для разметки данных (выделение объектов, аннотирование текста и т.д.).
    - Разметчик может сохранять промежуточные результаты, а также отправлять задания на ревизию по мере их завершения.
4. Проверка и утверждение разметки Заказчиком:
    - Заказчик проверяет выполненные разметки, используя механизмы контроля качества, такие как автоматическая проверка, двойная разметка и консенсус разметчиков.
    - В случае обнаружения ошибок или несоответствий, Заказчик может отправить задание на доработку разметчику.
    - После успешной проверки задания утверждаются Заказчиком.
5. Скачивание конечного размеченного датасета:
    - По завершении всех заданий и утверждении разметок, Заказчик может скачать конечный размеченный датасет.
    - Датасет экспортируется в формате, удобном для использования в моделях машинного обучения (например, CSV, JSON, XML).

- Размеченные данные передаются в отделы, занимающиеся построением и обучением моделей машинного обучения.

6. Обратная связь и улучшение процесса:
   - Заказчик собирает обратную связь от разметчиков о сложности задания и возможных проблемах.
   - Анализируется качество разметки и корректируются процессы для улучшения будущих заданий.
   - В зависимости от результатов, могут быть пересмотрены требования к разметке, обновлены инструкции и обучающие материалы для разметчиков.

Пользовательские истории/сценарии

Разбиение по юзерам и эпикам

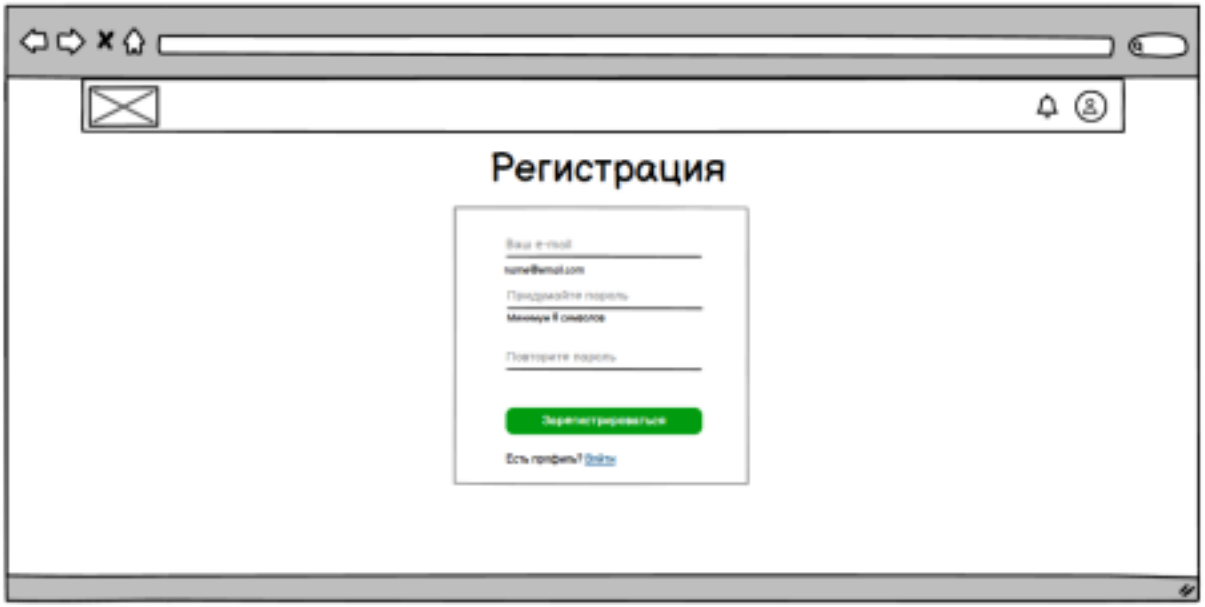
1. Заказчик:
   - Как Заказчик, я хочу создать новый проект разметки, чтобы организовать данные для разметки.
   - Как Заказчик, я хочу добавить описание, задание для разметчиков, датасет и коэффициент сложности (x2, x3, x4), чтобы разметчики знали, что делать.
   - Как Заказчик, я хочу мониторить выполнение заданий и проверять качество работы разметчиков, чтобы убедиться в правильности разметки.
   - Как Заказчик, я хочу проводить онбординг разметчиков, чтобы они могли правильно выполнять задания.
   - Как Заказчик, я хочу перераспределять задачи между разметчиками, чтобы обеспечить непрерывность работы.
   - Как Заказчик, я хочу скачать размеченный датасет по завершению разметки, чтобы использовать его для создания моделей машинного обучения.
2. Разметчик:
   - Как разметчик, я хочу просмотреть задание и получить инструкции, чтобы понять, что от меня требуется.
   - Как разметчик, я хочу выполнить разметку данных, используя доступные инструменты, чтобы выполнить задание.
   - Как разметчик, я хочу сохранять и отправлять результаты разметки, чтобы Заказчик мог их проверить.
   - Как разметчик, я хочу пройти онбординг и тестовые задания, чтобы быть уверенным в своих действиях.

Пользовательское взаимодействие и дизайн

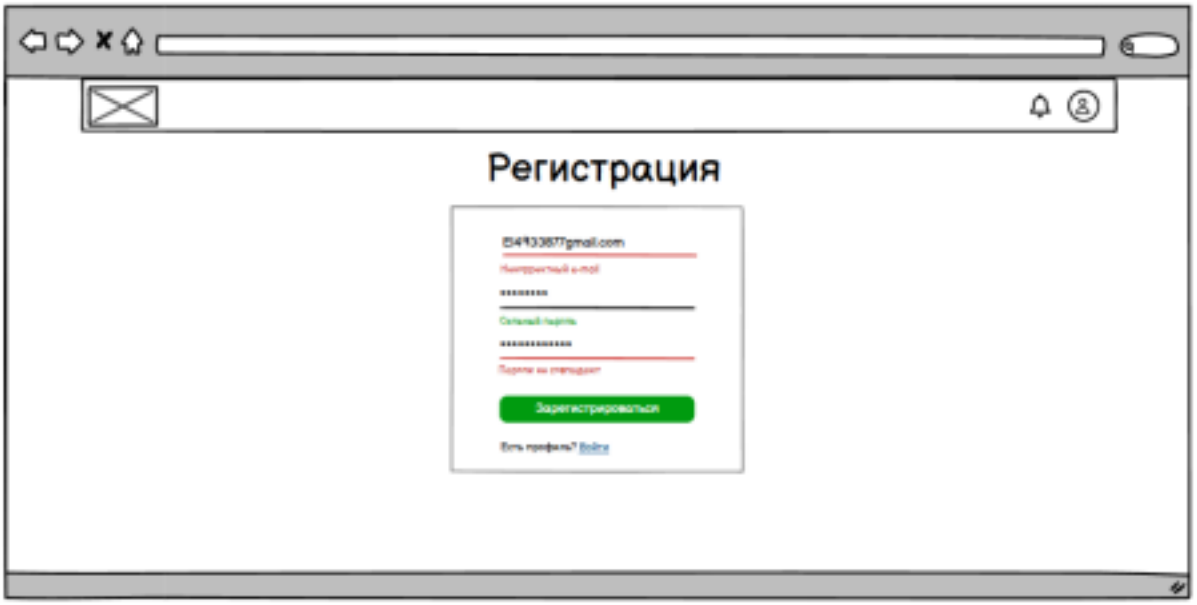
Мокапы с описанием

Ниже будут приведены мокапы основных экранов в Balsamiq:

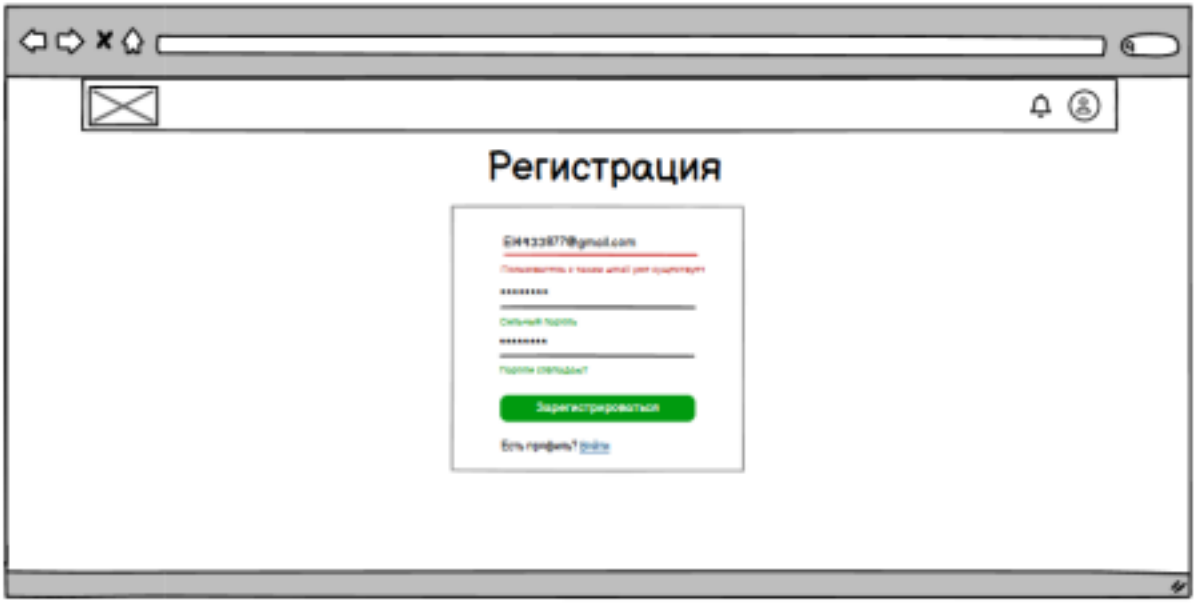
1. Экран входа в систему:



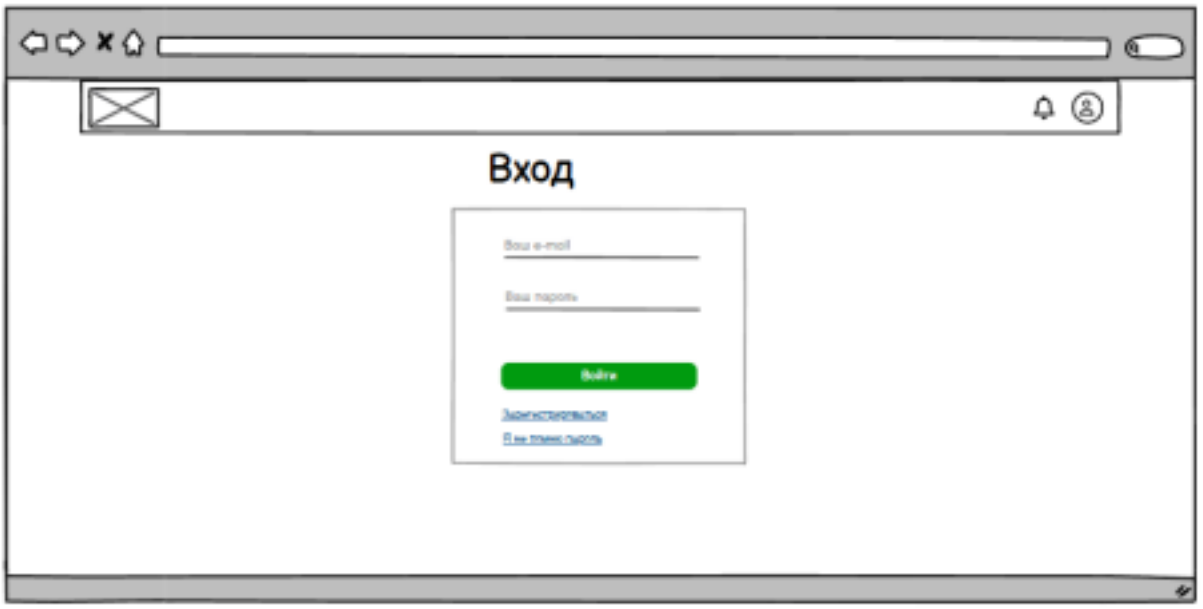
Регистрация



Регистрация. Исключение №1



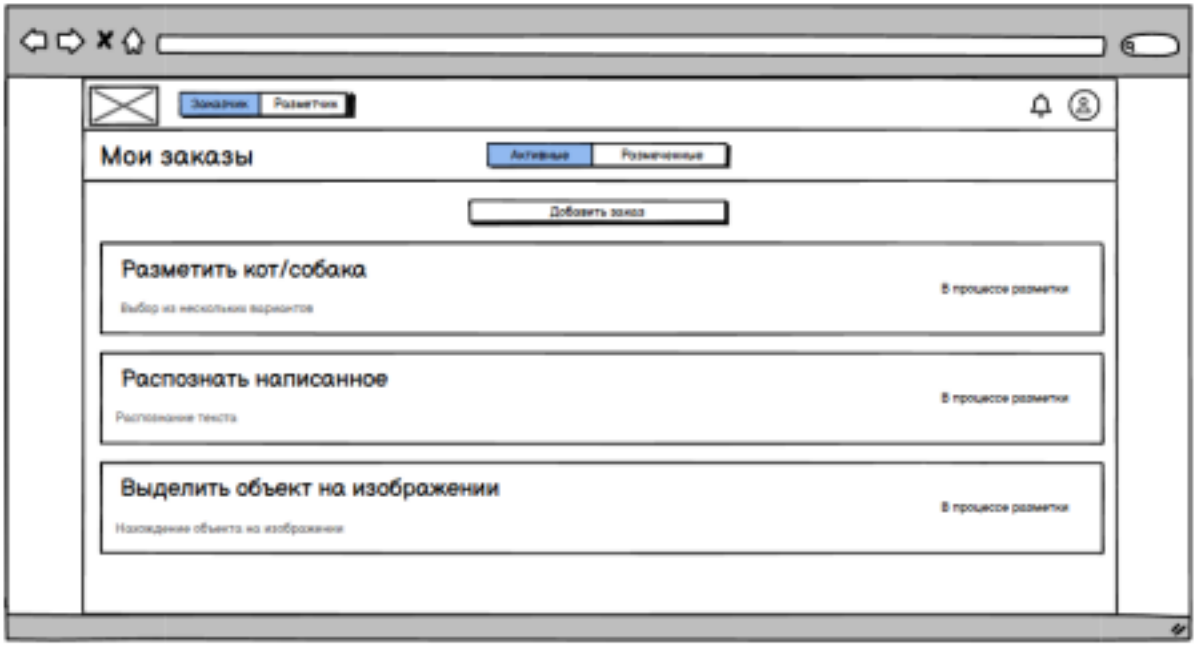
Регистрация. Исключение №2



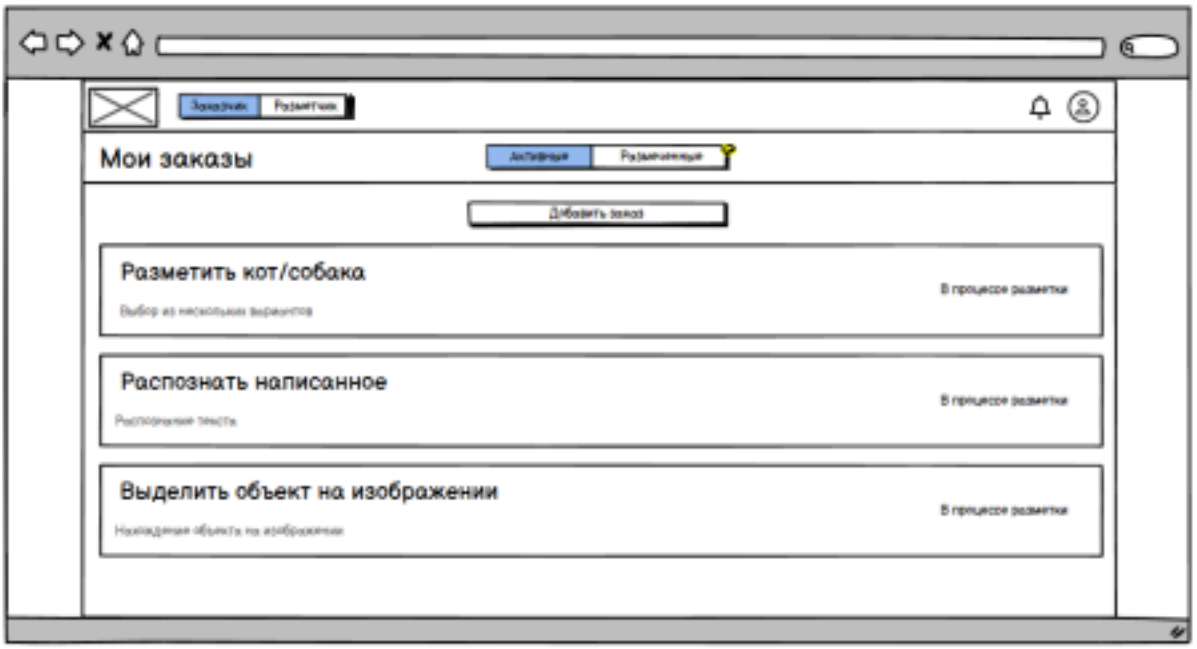
Вход

Вход. Исключение № 1

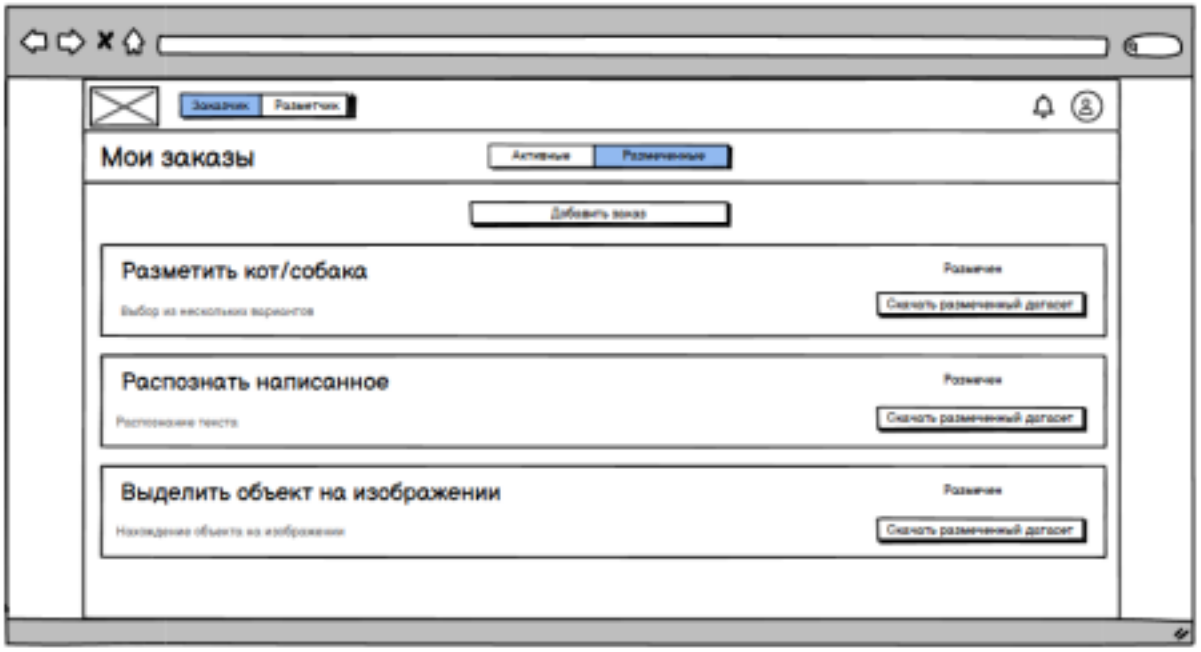
2. Панель управления проектами:



Заказчик. Активные заказы



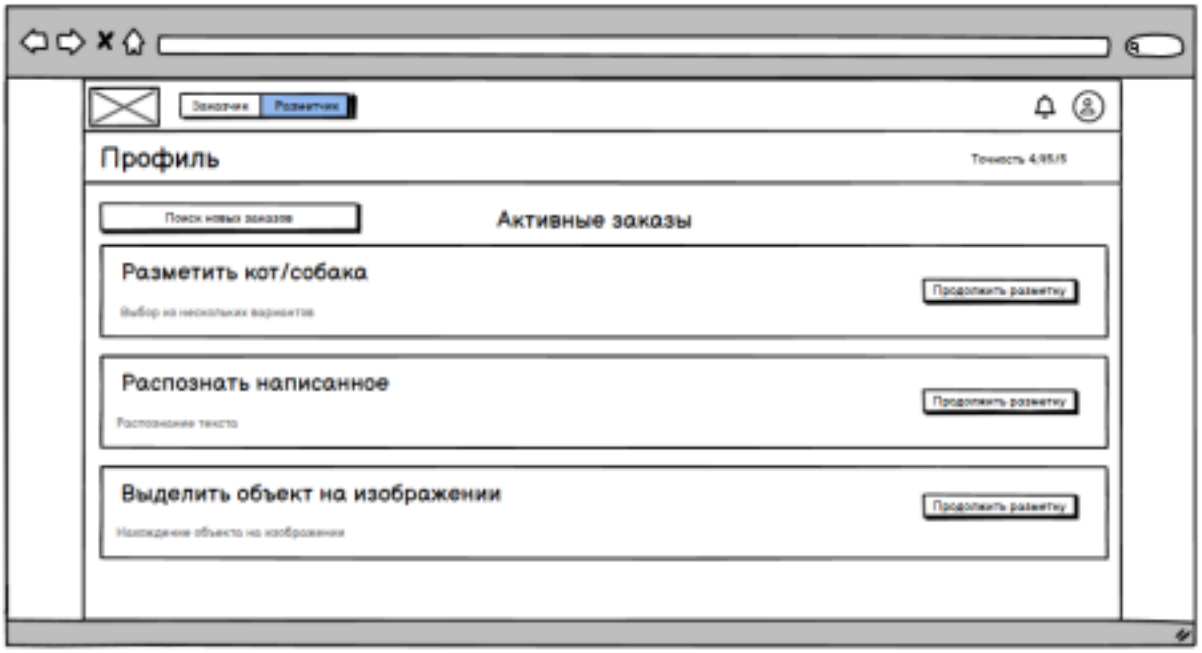
Оповещение о завершении разметки



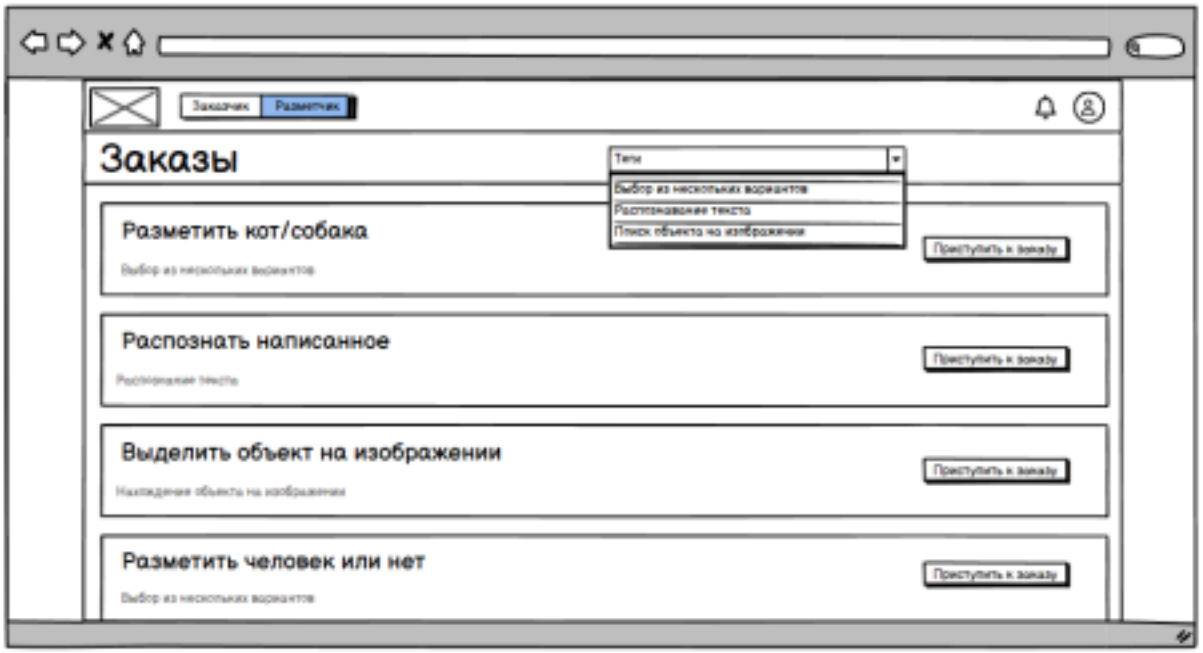
Заказчик. Размеченные заказы

Заказчик. Добавить заказ

3. Экран разметки данных:

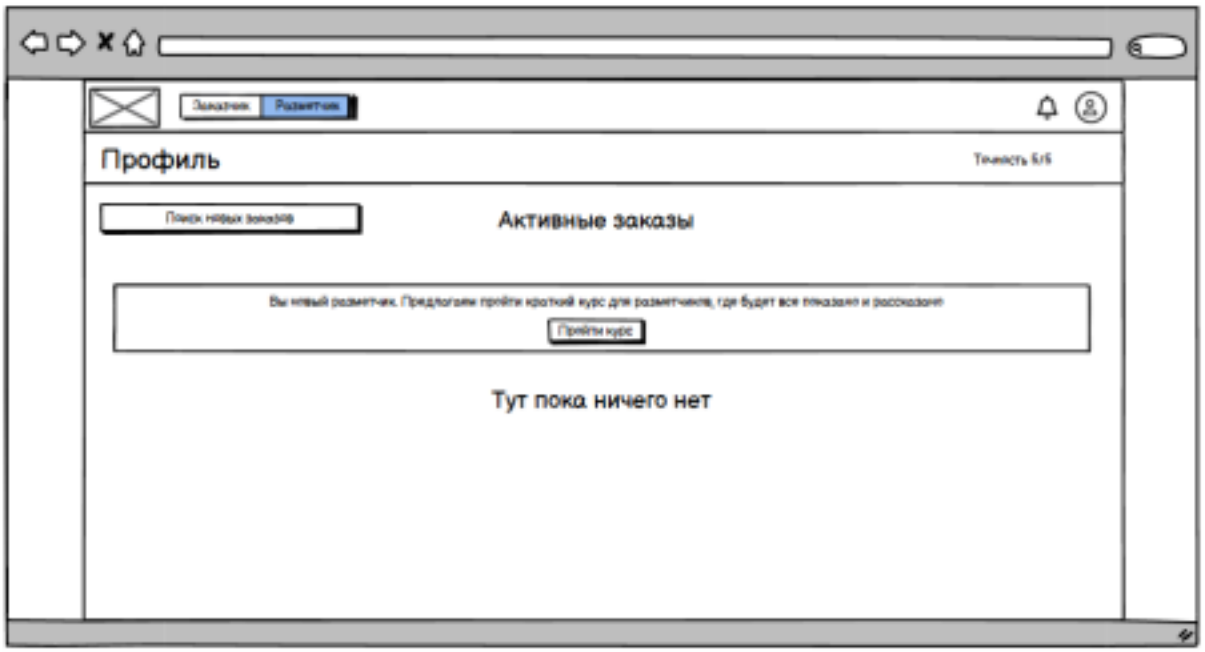


Профиль разметчика

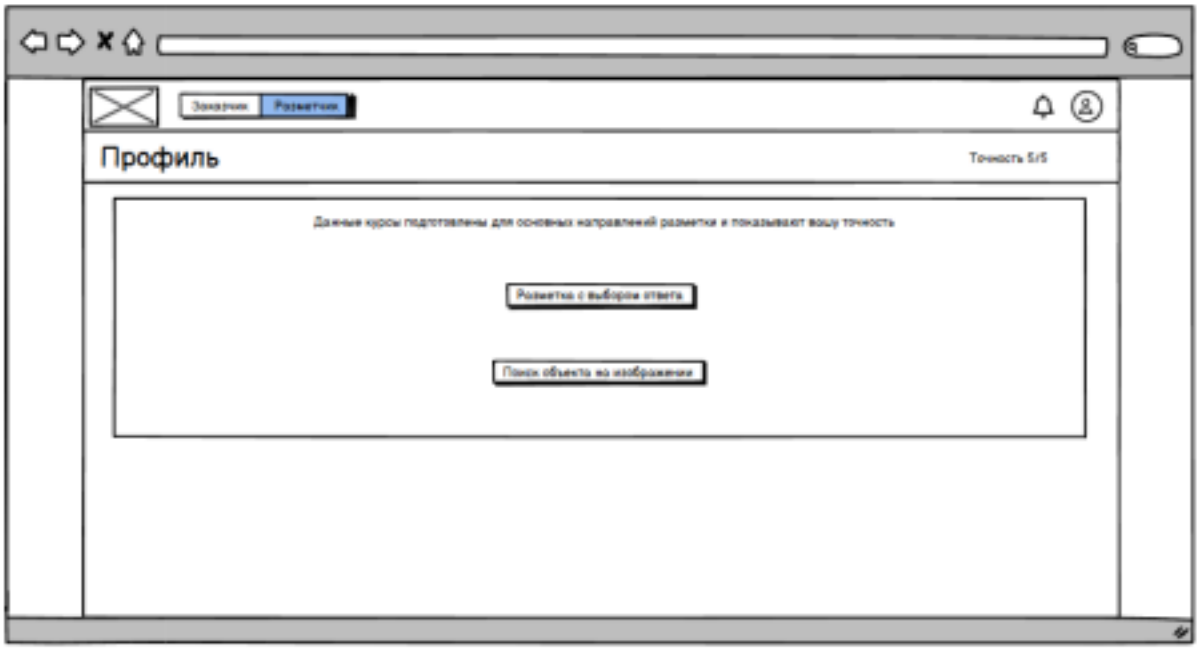


Список заказов для разметки

4. Экран онбординга:

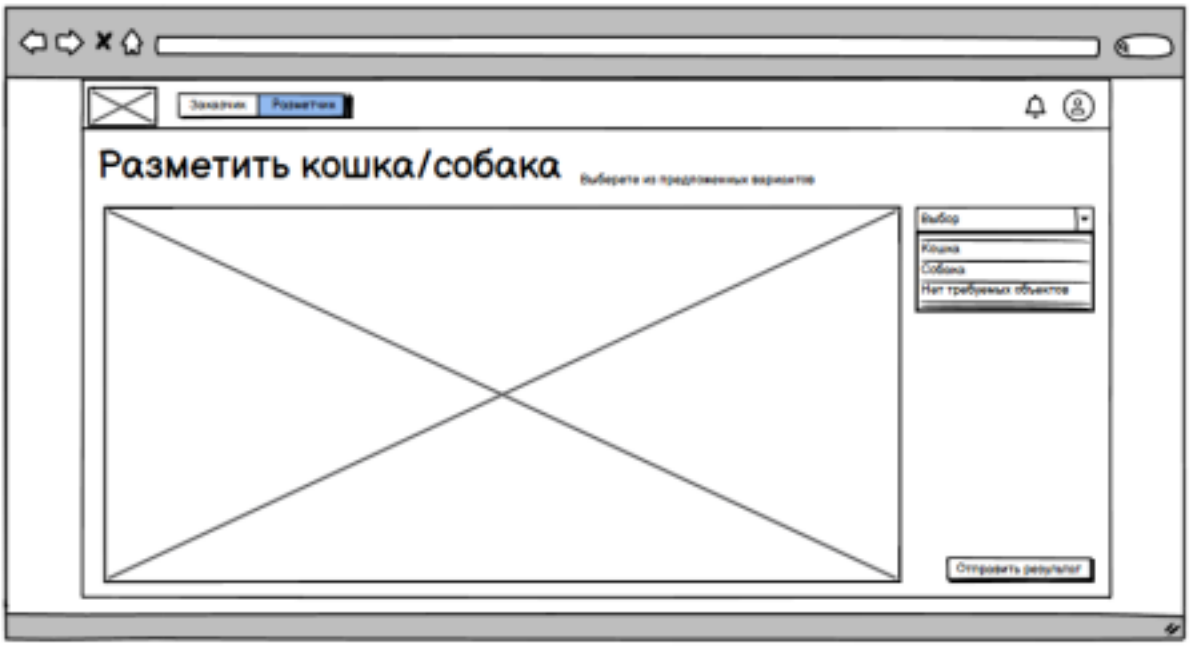


Краткое обучение



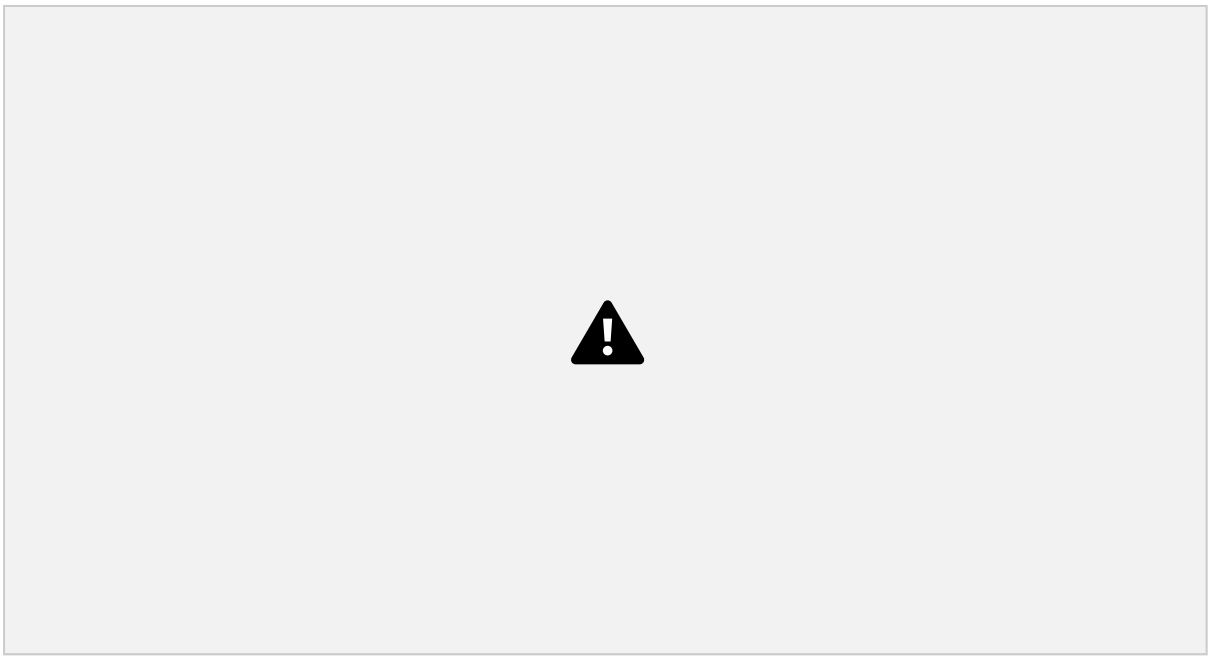
Обучение

5. Экран ревизии разметки:

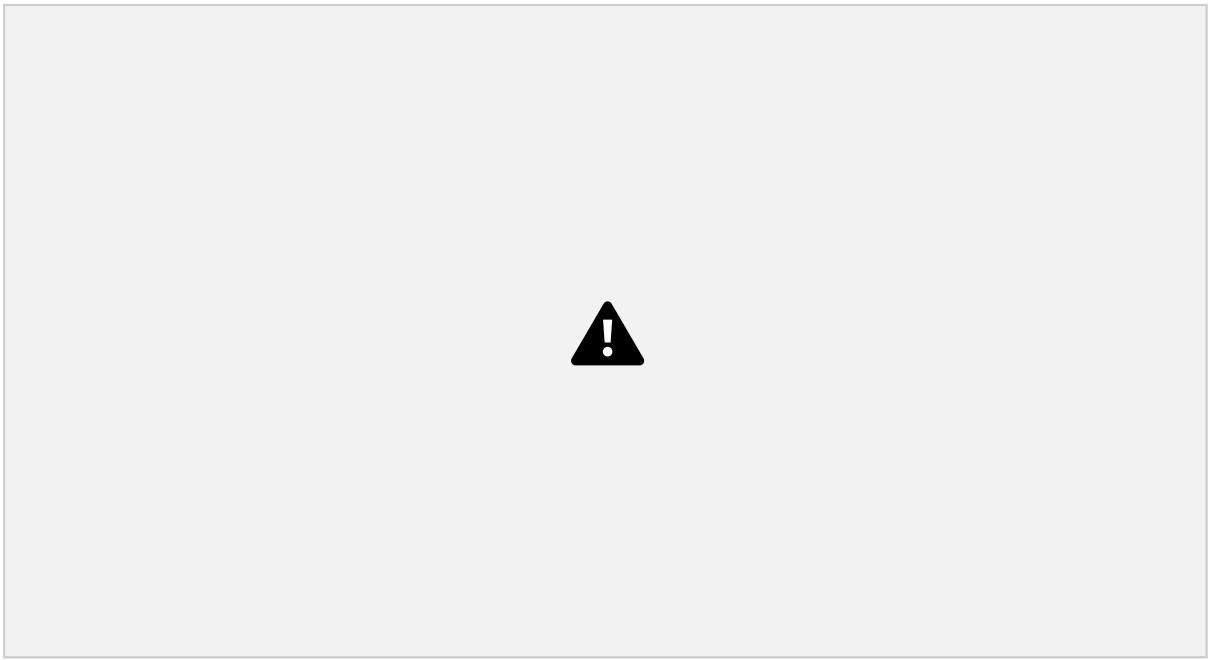


Разметка с вариантами

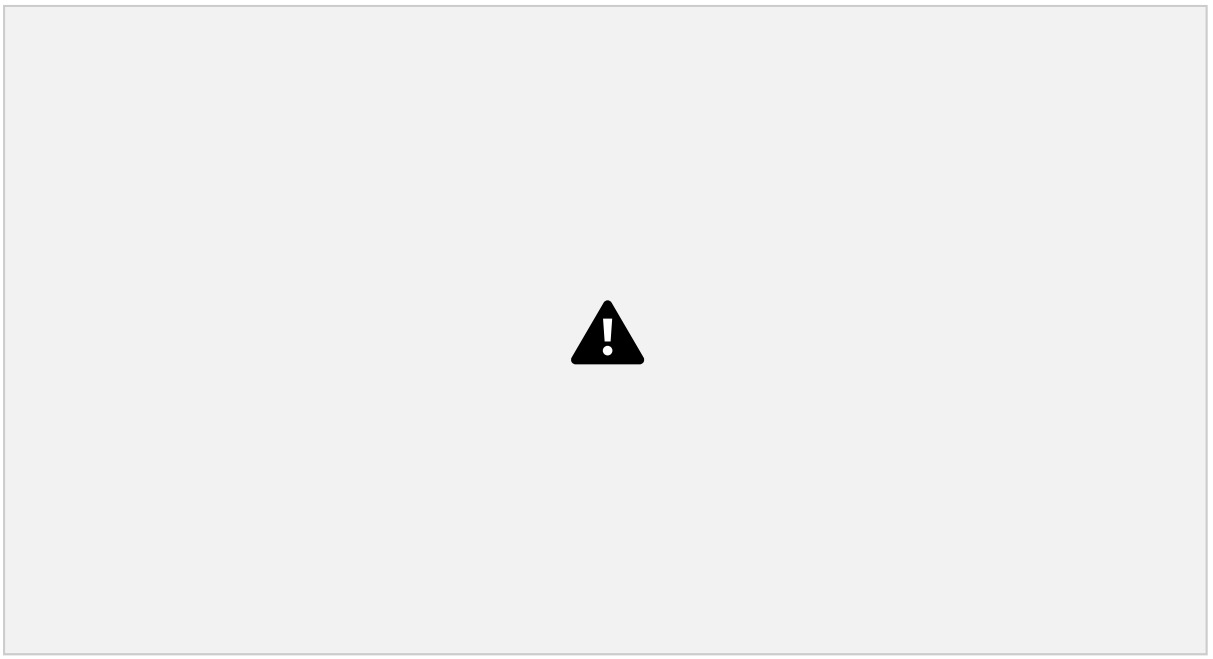
Разметка с вариантами обучения



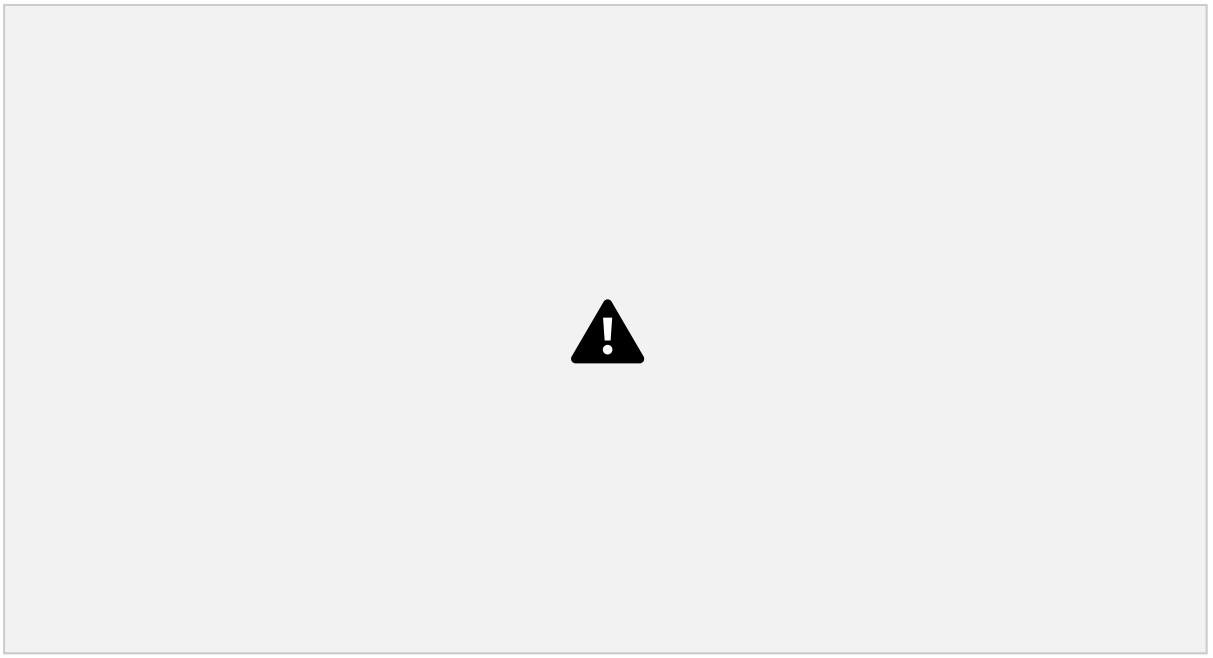
Разметка с вариантами обучения. Пройдено



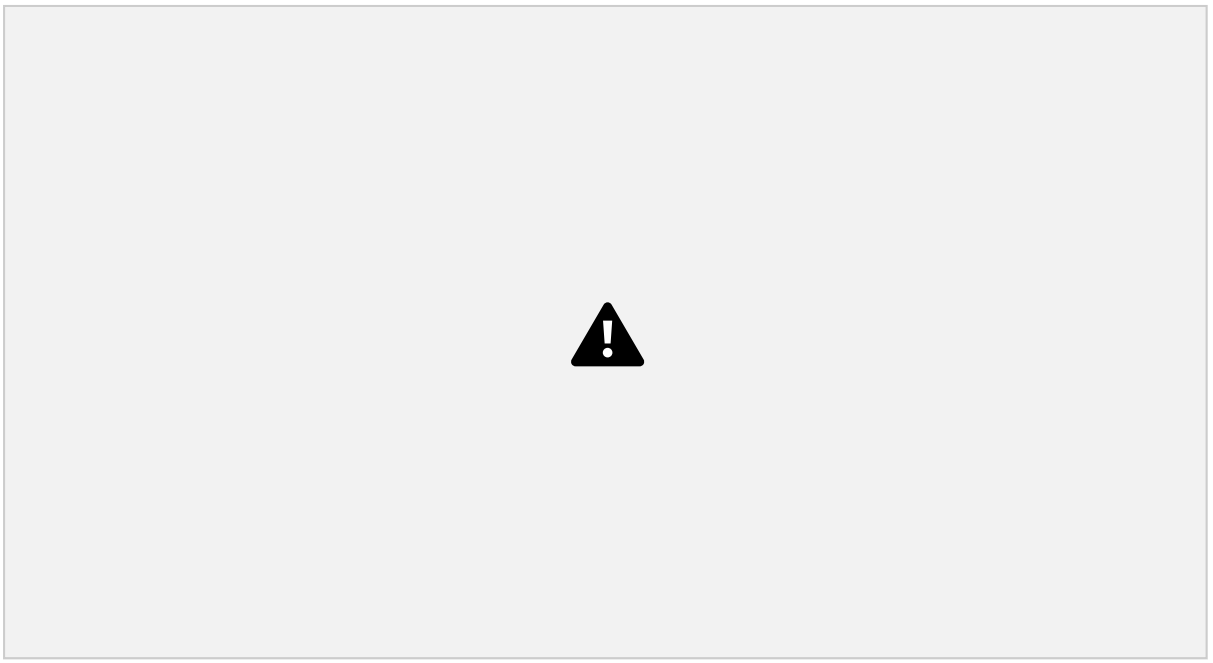
Разметка с вариантами обучения. Пройти еще раз



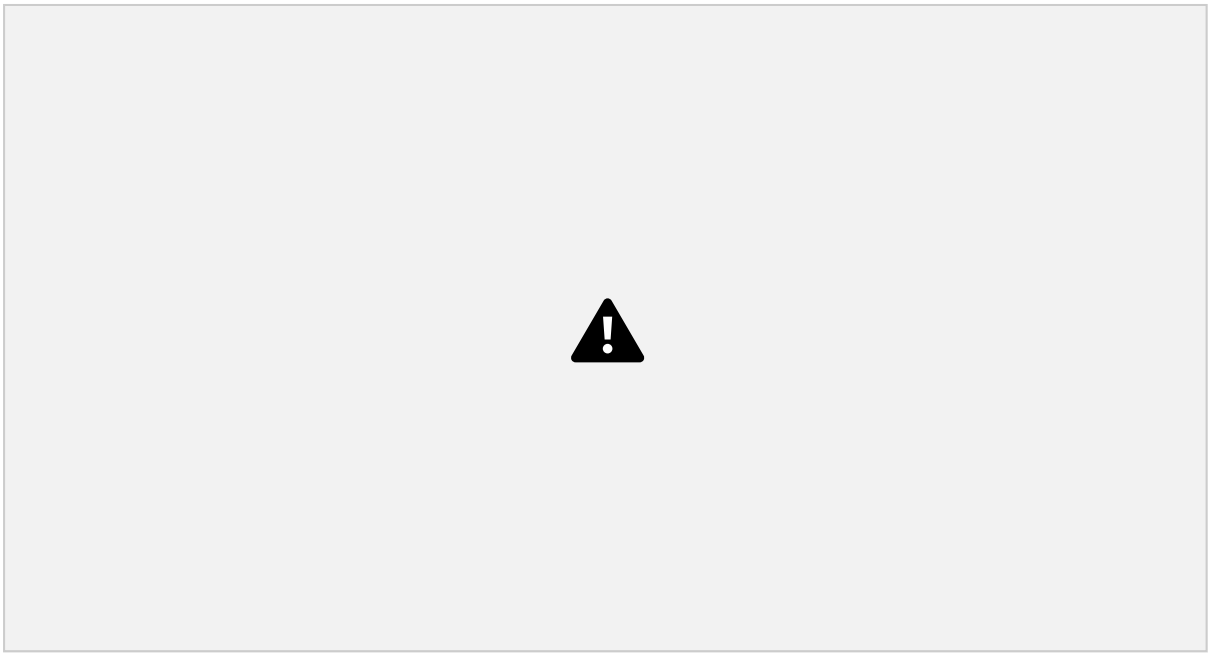
Поиск объекта на изображении



Разметка объекта на изображении. Обучение

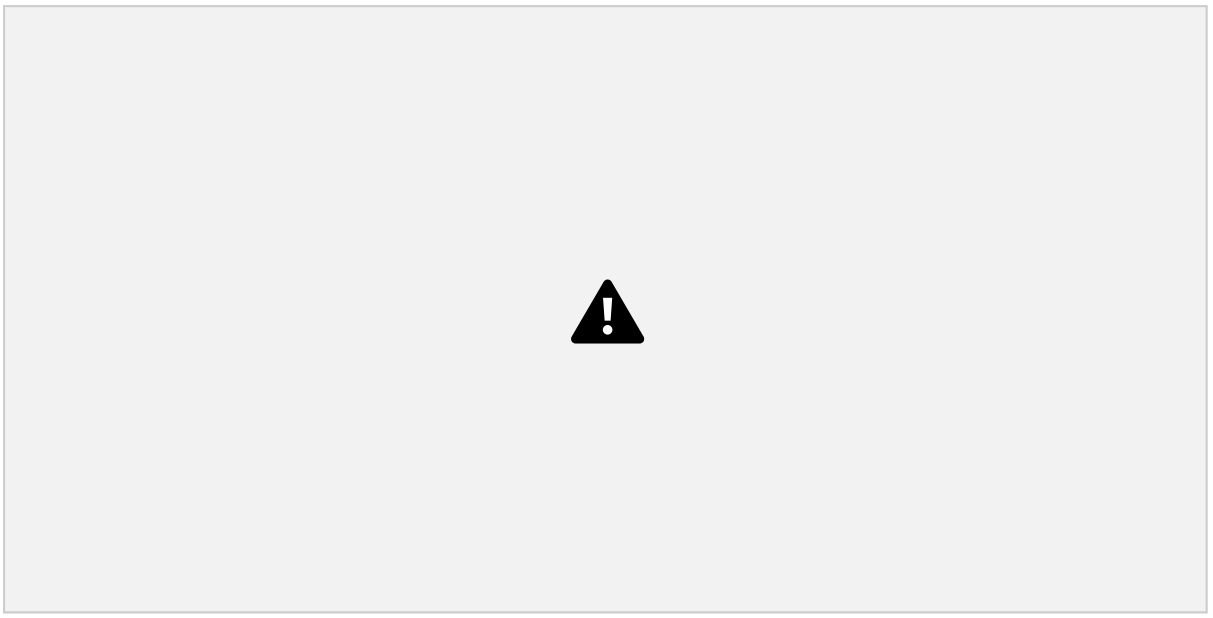


Разметка объекта на изображении. Обучение пройдено



Разметка объекта на изображении. Обучение не пройдено

Экран ограничения за плохую точность:



Временная блокировка

Доска KanbanFlow