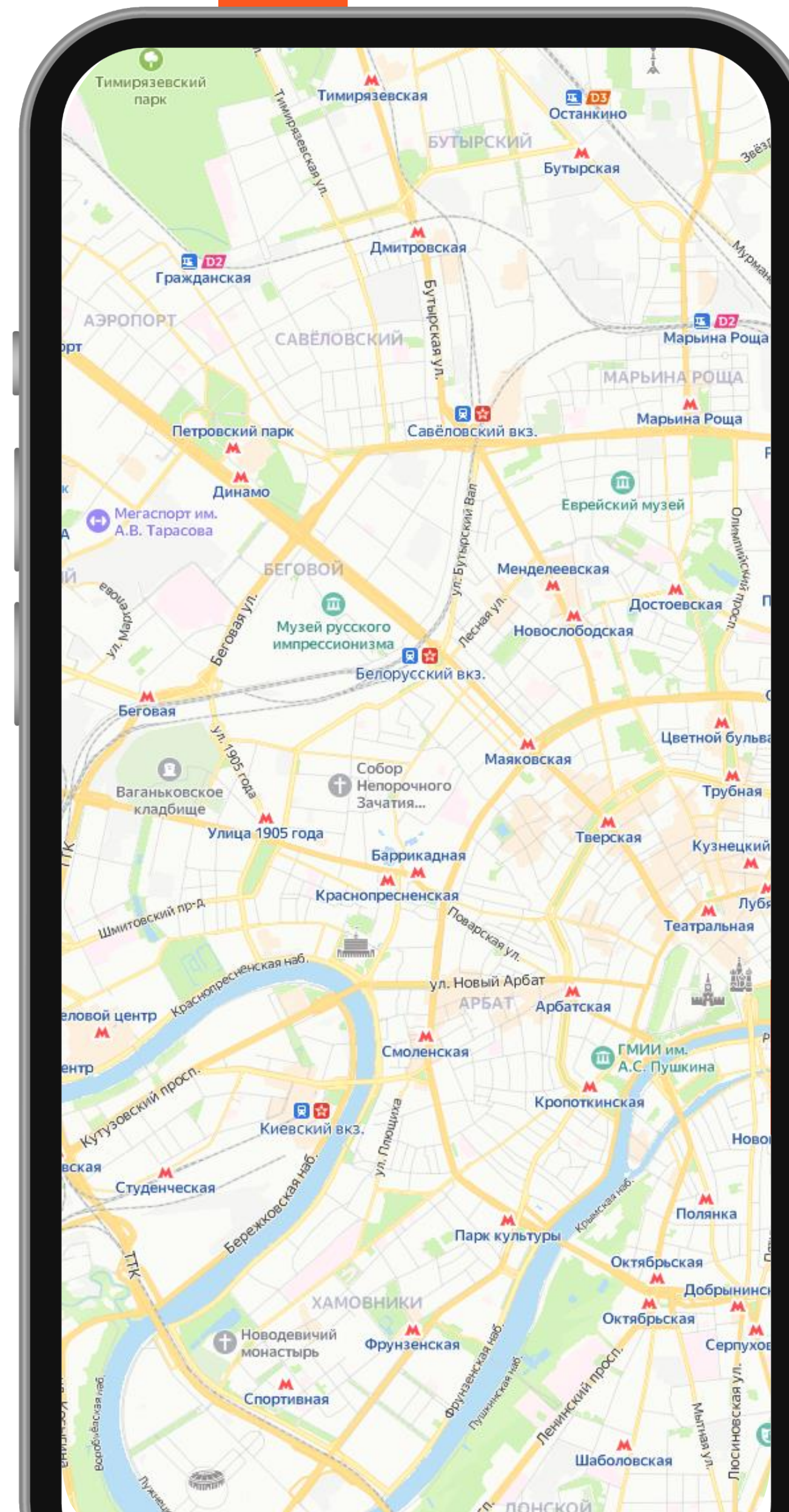


# Яндекс Карты

## Офлайн метрики базы организаций

Леонид Медников  
Ведущий аналитик



# О чём расскажу

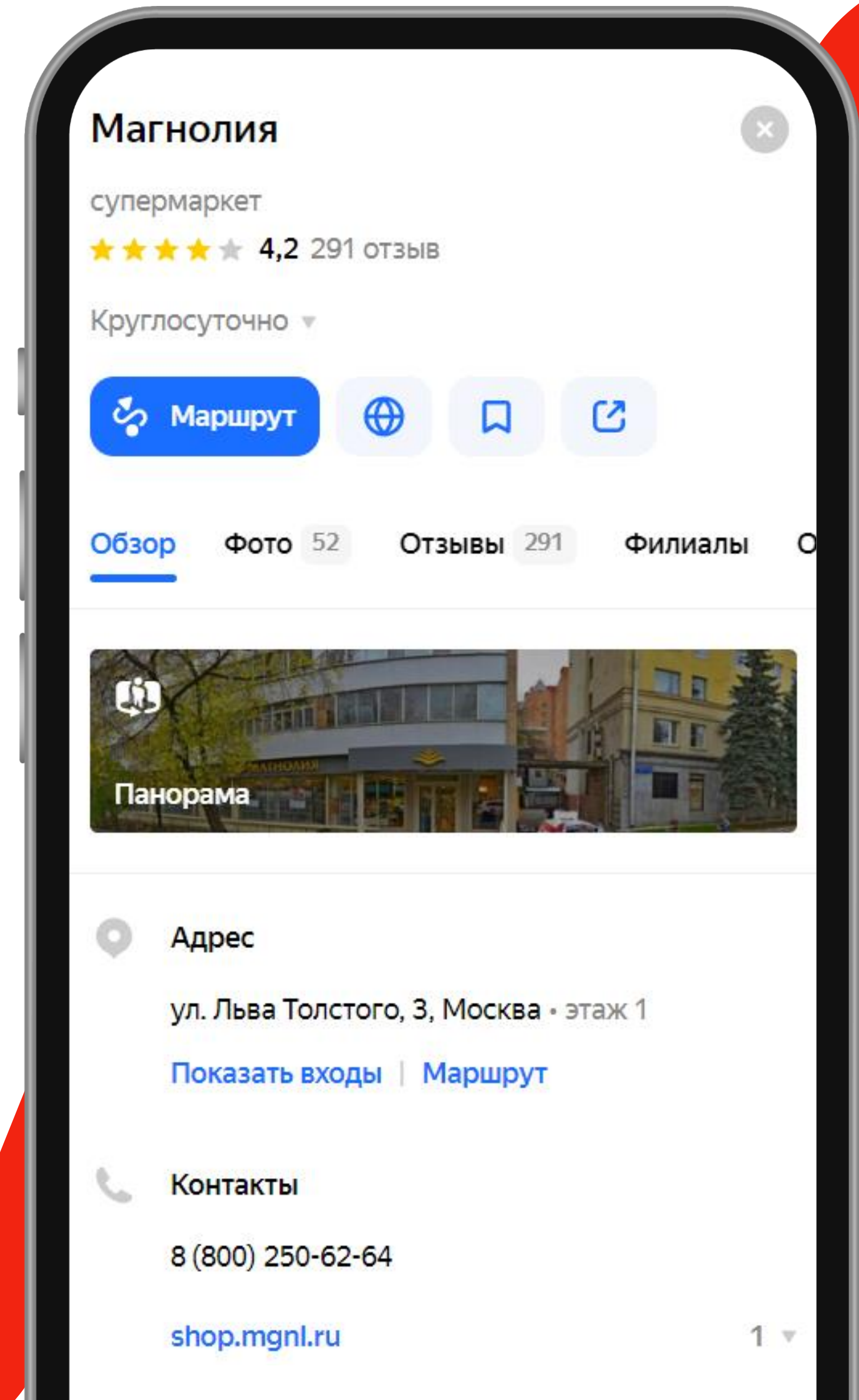
- 01 Про базу организаций
- 02 Почему не подходят пользовательские метрики?
- 03 Разметка данных асессорами
- 04 Разметка данных в реальном мире
- 05 Взвешивание оценок



# Справочник

Хранит данные об организациях (кафе, магазины, поликлиники) и других объектах (памятники, детские площадки).

- ★ Название организации
- 📍 Адрес, точное положение организации и входов
- ⚙ Вид деятельности (рубрика)
- 🕒 Часы работы
- ☎ Телефон
- 🌐 Сайт



# Источники данных

 Базы данных организаций

 Интернет

## Балатово

38 мин • 4.76 ★

ул. Мира, 100

Доставка и самовывоз

Пн-Вс: 09:00 — 00:00

Ресторан

Пн-Вс: 09:00 — 00:00

## Парковый

39 мин • 4.8 ★

пр-т Парковый, 33

Доставка и самовывоз

Пн-Вс: 09:00 — 00:00

Ресторан

Пн-Вс: 09:00 — 00:00

## Закамск

39 мин • 4.8 ★

ул. Ласьвинская, 35/1

Доставка и самовывоз

Пн-Вс: 10:00 — 00:00

Ресторан

Пн-Вс: 10:00 — 00:00

## Садовый

42 мин • 4.65 ★

ул. Уинская, 10

Доставка и самовывоз

Пн-Вс: 09:00 — 00:00

Ресторан

Пн-Вс: 09:00 — 00:00

## Колизей

46 мин • 4.72 ★

ул. Ленина, 60





Доставка и самовывоз

Пн-Вс: 09:00 — 00:00

Ресторан

Пн-Вс: 08:00 — 00:00

# Источники данных

-  Базы данных организаций
-  Интернет
-  Фиды от партнёров и сетей
-  Данные от владельцев бизнесов

Я Бизнес

Реклама

Организации

Заявки

Поиск

DREAM TEAM

Дрим-Тим

Центральный федеральный округ, Москва, Троицкий административный округ, поселение Новофёдоровское, деревня Кузнецово, 3-й Заречный переулок, 2

Главная

О компании

Данные

Фото и видео

Публикации

Отзывы

Рейтинг компании

Товары и услуги

Доставка

Промоматериалы

Карта на сайт

Изменения

Конкуренты

Реклама

Статистика

Форма заявки

Сайт

Доступы

Данные

Есть изменения на модерации

Основные

Особенности

Реквизиты

Статус в Профиле и Картах

Измените статус работы на актуальный, если вы больше не работаете или закрылись на ремонт и хотите сообщить об этом пользователям.

Работает

Логотип

На модерации

DREAM TEAM

Логотип поможет оформить профиль организации в Поиске, Картах и других сервисах Яндекса в корпоративном стиле.

Вид деятельности

Консьерж-сервис (Основной)

Добавить категорию

не больше трёх

График работы

будни 9:00-19:00

Будни↑ Выходные↑ 24/7↑ Ежедневно↑ Перерыв↑

Добавить режим работы на дату

Контакты

Для связи и перехода на ваши социальные сети. Будут отображаться в профилях.

В профиле будет опубликована ссылка только на один сайт. Укажите адрес сайта, который наиболее важен для привлечения клиентов, и контакты в соцсетях. Прочие адреса сайтов лучше удалить.

Адрес сайта

Создать сайт на Яндекс Бизнес

ВКонтакте

Логин или публик в ВК

YouTube

Ссылка на канал в Youtube

Одноклассники

Логин или публик в Одноклассниках

Ваша организация в Яндексе

Я

Поиск

Я

Карты

5



# Источники данных

- Базы данных организаций
- Интернет
- Фиды от партнёров и сетей
- Данные от владельцев бизнесов
- Фидбек от пользователей и из Народной карты

Добавить организацию

Если вы представитель компании, лучше добавьте организацию в [Яндекс Бизнесе](#): там можно подробнее о ней рассказать.

Название \*

Введите название организации

Адрес \*

Укажите местоположение организации с помощью метки на карте

Воронеж, Кольцовская улица, 44

Вид деятельности \*

не более трёх

Начните печатать и выберите из списка

Время работы

Дни недели ▾

Время работы ▾

Без перерыва ▾






Контакты ▾

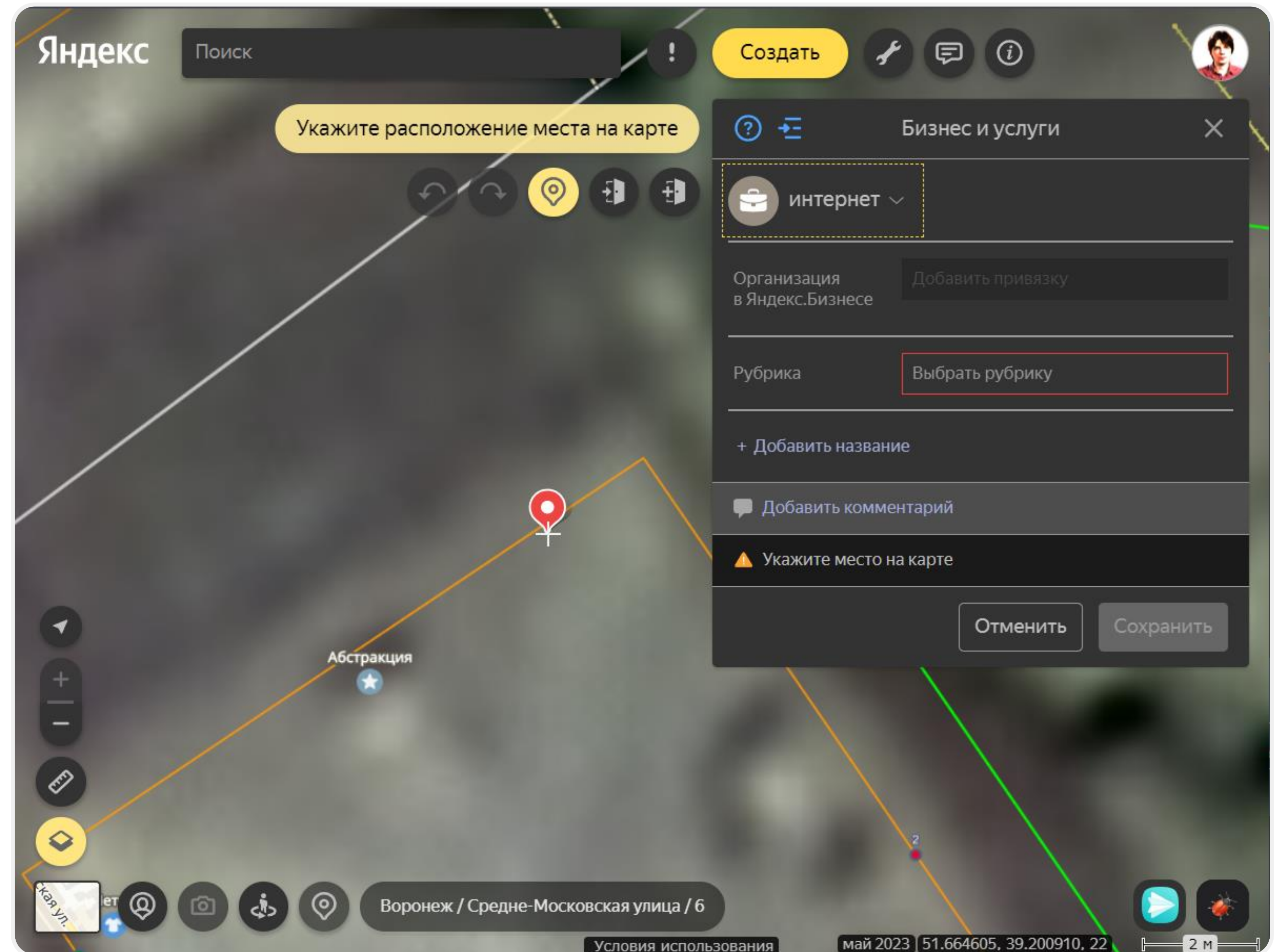
Отправить

Добавить вход

The map shows a street intersection area. At the top, there's a coffee shop 'Chicco Di Caffè' with an orange cup icon. Below it are two shops: 'Продуктофф' (Productoff) with a shopping basket icon and 'Ozon' with a shopping bag icon. A blue pin icon is placed on the map, indicating a location. The street name 'Средне-Московская ул.' (Sredne-Moskovskaya St.) is written diagonally across the bottom right. A blue square button with a white crosshair icon is also visible on the map.








# Источники данных

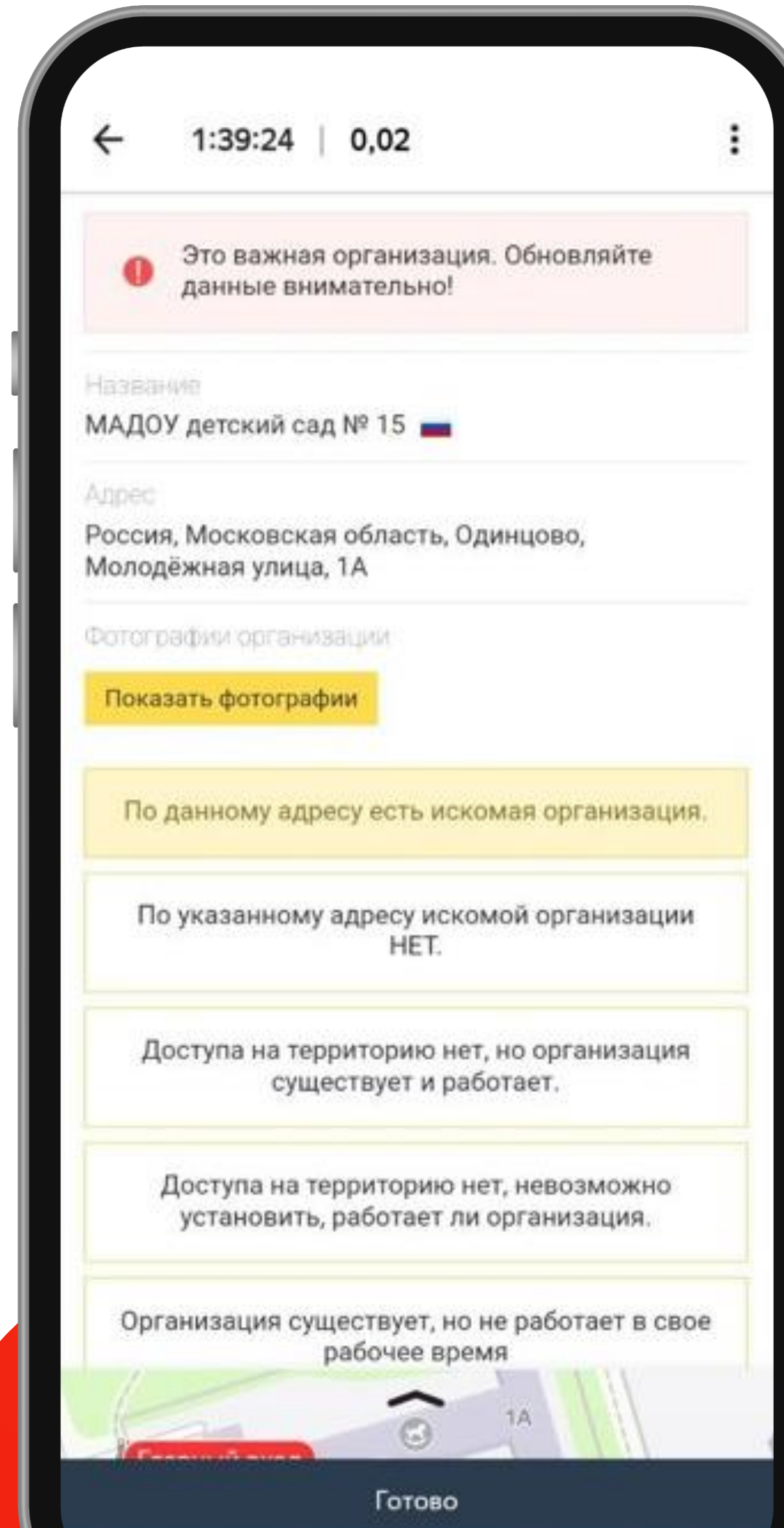
-  Базы данных организаций
-  Интернет
-  Фиды от партнёров и сетей
-  Данные от владельцев бизнесов
-  Фидбек от пользователей и из Народной карты





# Источники данных

-  Базы данных организаций
-  Интернет
-  Фиды от партнёров и сетей
-  Данные от владельцев бизнесов
-  Фидбек от пользователей и из Народной карты
-  Проверка через колл-центр
-  Сбор данных пешеходами



← 1:39:24 | 0,02

❗ Это важная организация. Обновляйте данные внимательно!

Название  
МАДОУ детский сад № 15 🇷🇺

Адрес  
Россия, Московская область, Одинцово, Молодёжная улица, 1А

Фотографии организации  
Показать фотографии

По данному адресу есть искомая организация.

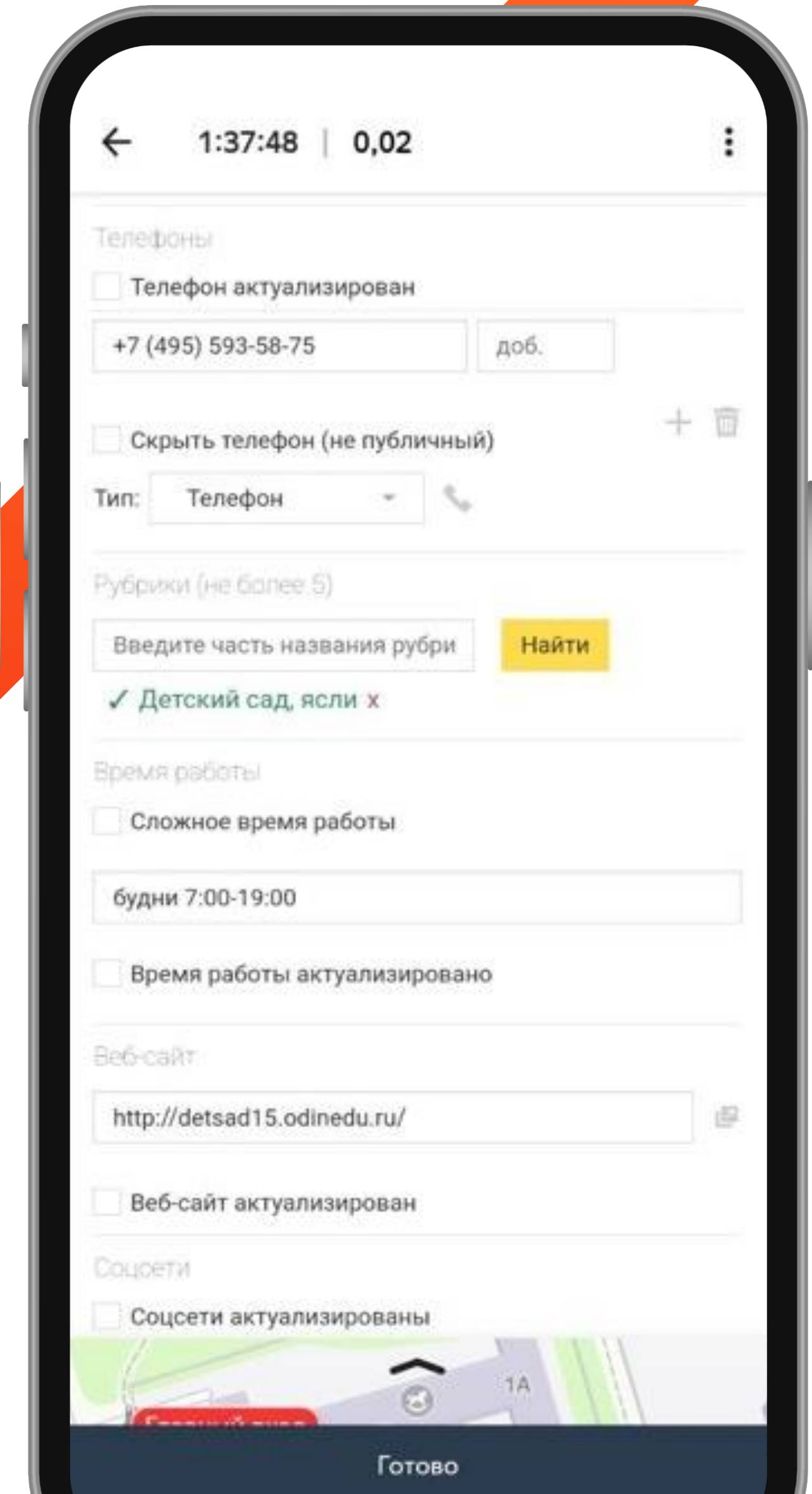
По указанному адресу искомой организации НЕТ.

Доступа на территорию нет, но организация существует и работает.

Доступа на территорию нет, невозможно установить, работает ли организация.

Организация существует, но не работает в свое рабочее время

Готово



← 1:37:48 | 0,02

Телефоны

☐ Телефон актуализирован

+7 (495) 593-58-75 доб.

☐ Скрыть телефон (не публичный) + 🗑

Тип: Телефон 📞

Рубрики (не более 5)

Введите часть названия рубри Найти

✓ Детский сад, ясли x

Время работы

☐ Сложное время работы

будни 7:00-19:00

☐ Время работы актуализировано

Веб-сайт

<http://detsad15.odinedu.ru/> 📄

☐ Веб-сайт актуализирован

Соцсети

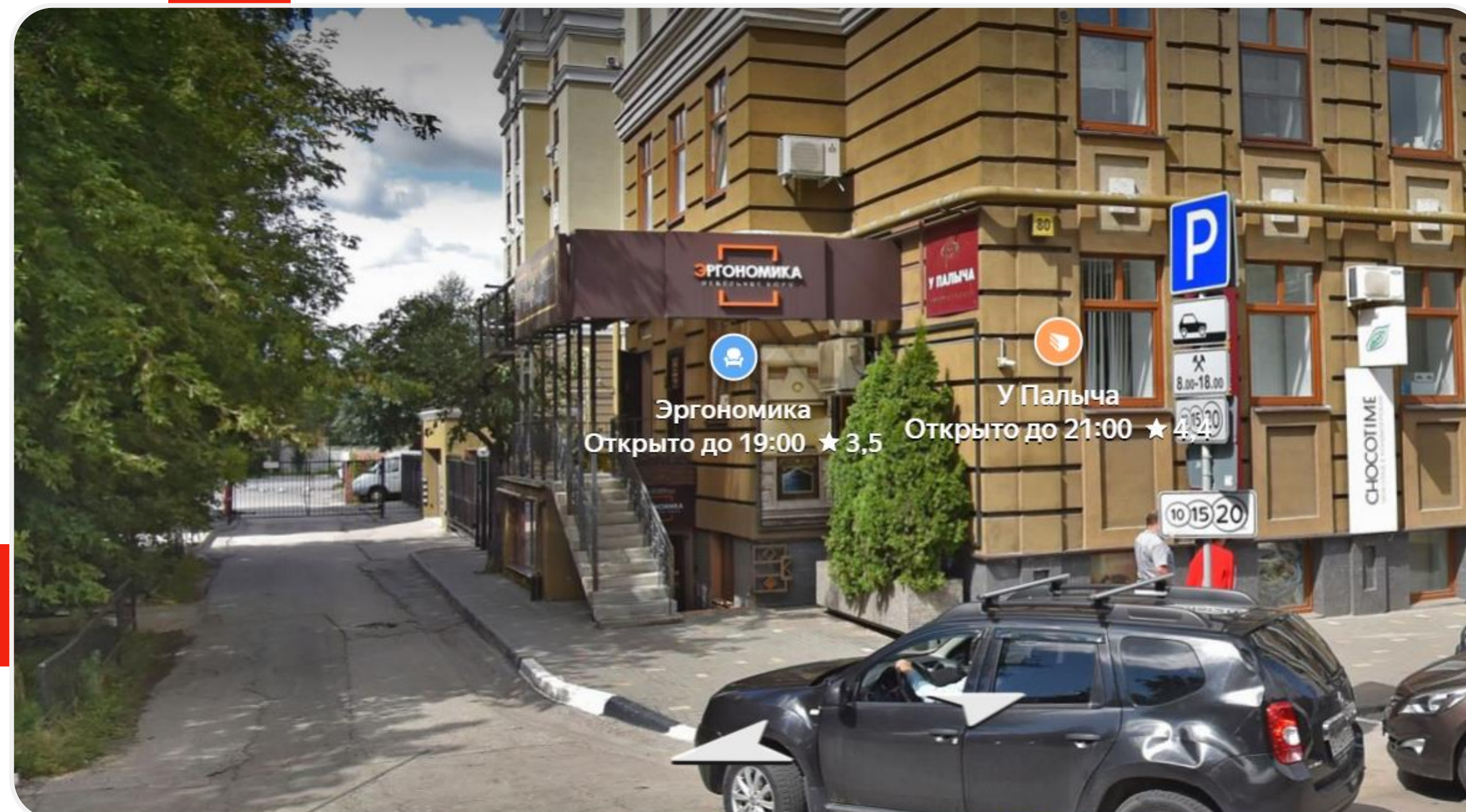
☐ Соцсети актуализированы

Готово



# Источники данных

- ☁ Базы данных организаций
- 🌐 Интернет
- 📋 Фиды от партнёров и сетей
- 👤 Данные от владельцев бизнесов
- 📖 Фидбек от пользователей и из Народной карты
- 👁 Проверка через колл-центр
- 💬 Сбор данных пешеходами
- 📷 Панорамы и фотографии



# Конфликт источников

Название	Адрес	Координаты	Время работы	Телефон	Рубрики
Domashniy Vkus (main)	Россия, Москва, Малахитовая улица, 7	37.657540 55.832940			Ресторан
Domashniy Vkus (main)		coordinates			
Домашний вкус (main)					
Домашний вкус (main)	Россия, Москва, Малахитовая улица, 7	37.657411 55.832938	ежедневно 9:00-23:00	+7 (925) 203-39-96 +7 (903) 787-98-60	Пекарня Кафе
Домашний вкус (short)		original			
Домашний Вкус (synonym)					
Столовая Домашний вкус (synonym)					



# Метрики

Онлайн vs. офлайн

## Онлайн метрики

Пользовательские метрики  
не позволяют измерить  
качество

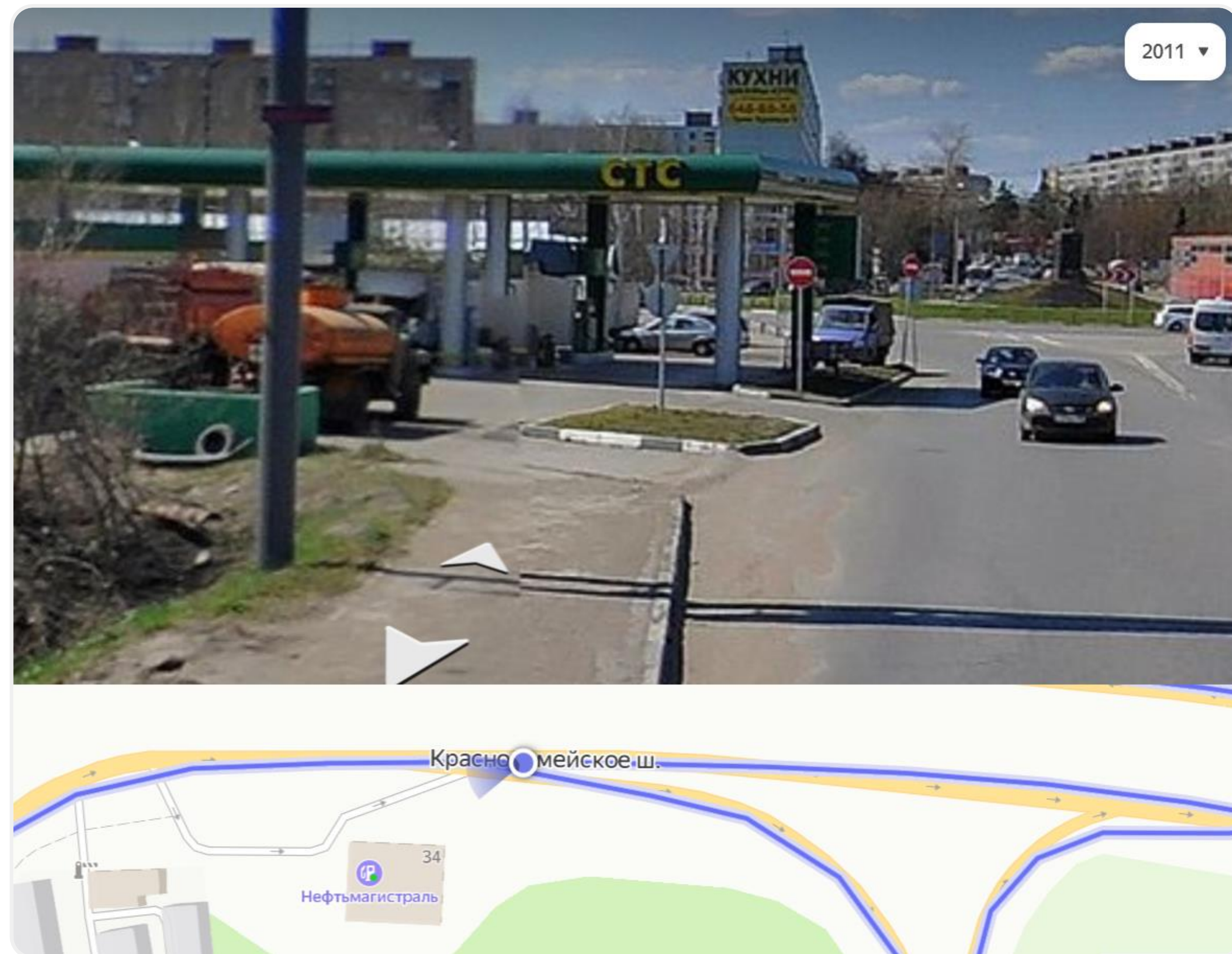
## Офлайн метрики

Напрямую измеряют  
качества данных

- Какие метрики
- Где взять эталон
- Правильная выборка

# Какие проблемы такие метрики

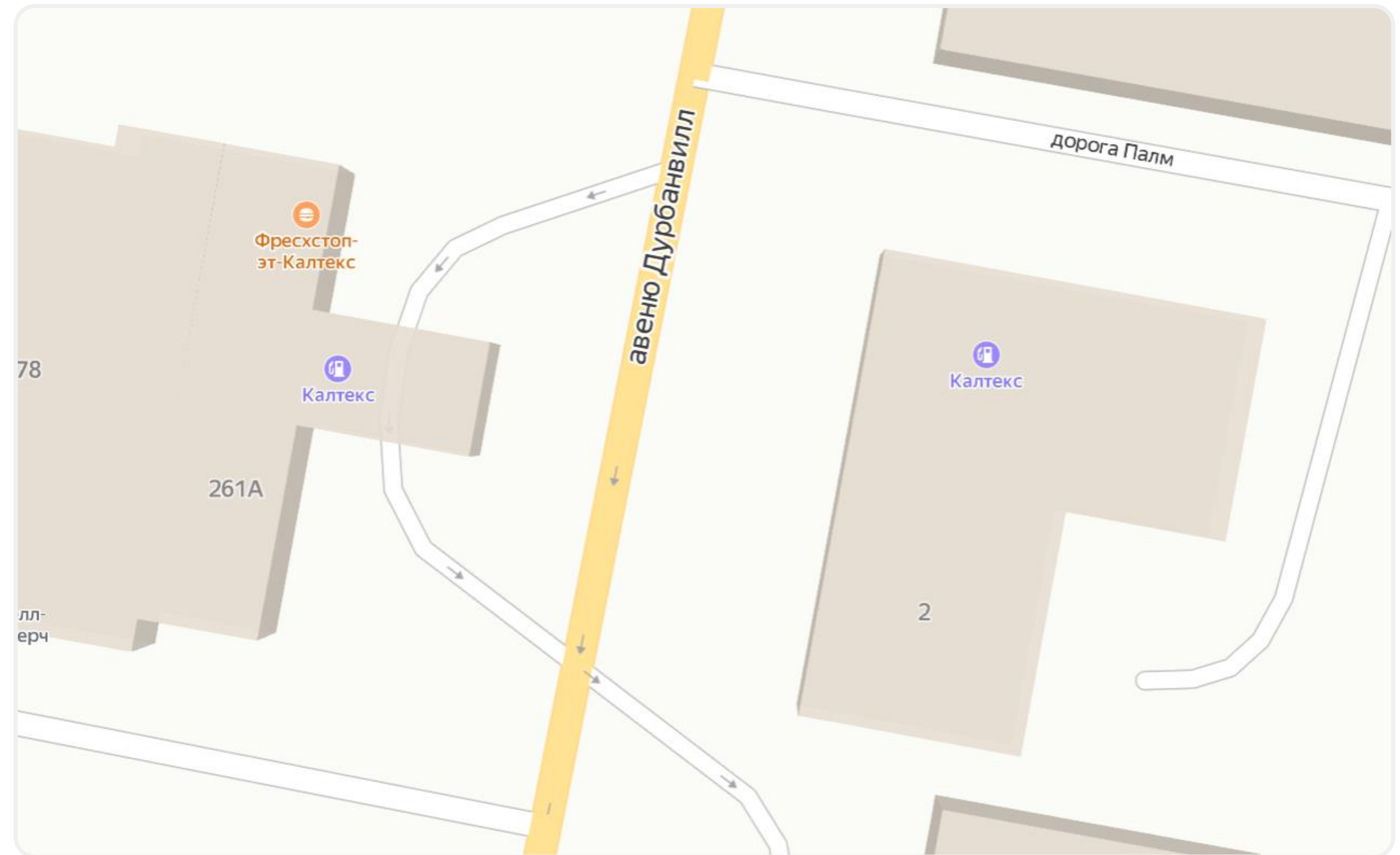
- Показываем закрытые организации
- Не показываем открытые
- Показываем не там
- Отсутствует или неправильное расписание работы, телефон, сайт
- Неточное название, рубрика
- Нет фотографий, отзывов
- Показываем организацию дважды





# Какие проблемы такие метрики

- Показываем закрытые организации
- Не показываем открытые
- Показываем не там
- Отсутствует или неправильное расписание работы, телефон, сайт
- Неточное название, рубрика
- Нет фотографий, отзывов
- Показываем организацию дважды



# Дубли

Из разных источников приходят данные про одну компанию.

Если данные отличаются, мы можем решить, что это 2 разные компании.

Как измерить дубли?

- Ищем похожие компании алгоритмами
- Размечаем дубли ассессорами

name	Калина ойл	Калина Ойл
name_en	Kalina oyl	Kalina Oil
phones	+7 (930) 760-78-84	+7 (920) 229-06-68;+7 (951) 556-28-60
address	Курск, Станционная улица, 39А	Курск, Станционная улица, 39А
main_url	-	<a href="https://kalina.ru/">https://kalina.ru/</a>
main_rubric_name_ru	A3C	A3C
rubric_names_ru	A3C	A3C;АГНС, АГЗС, АГНКС



# Проблемы метрики дублей

01

Алгоритм  
не находит  
дубли

02

Алгоритм может  
начать находить  
больше/меньше  
дублей

03

Асессоры по-разному  
оценивают наличие  
дубля

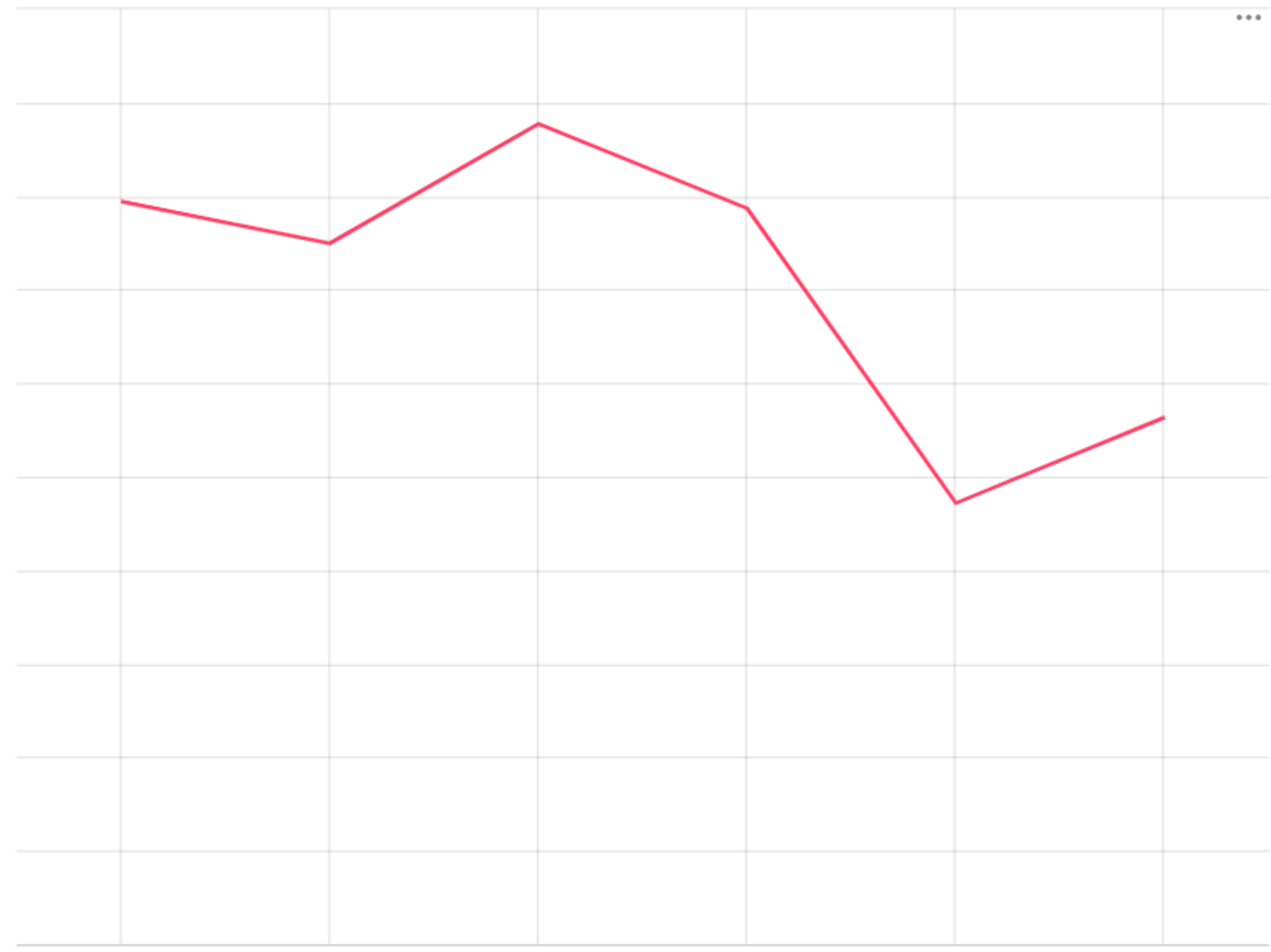
# Что можно сделать лучше?

- 01 Обучение
- 02 Тесты перед началом разметки
- 03 Нонеурот'ы
- 04 Перекрытие между ассессорами
- 05 Разбор расхождений
- 06 Улучшение инструкции
- 07 Отдельный статус для неопределённых ситуаций
- 08 Повторить



# Итоговая метрика дублей

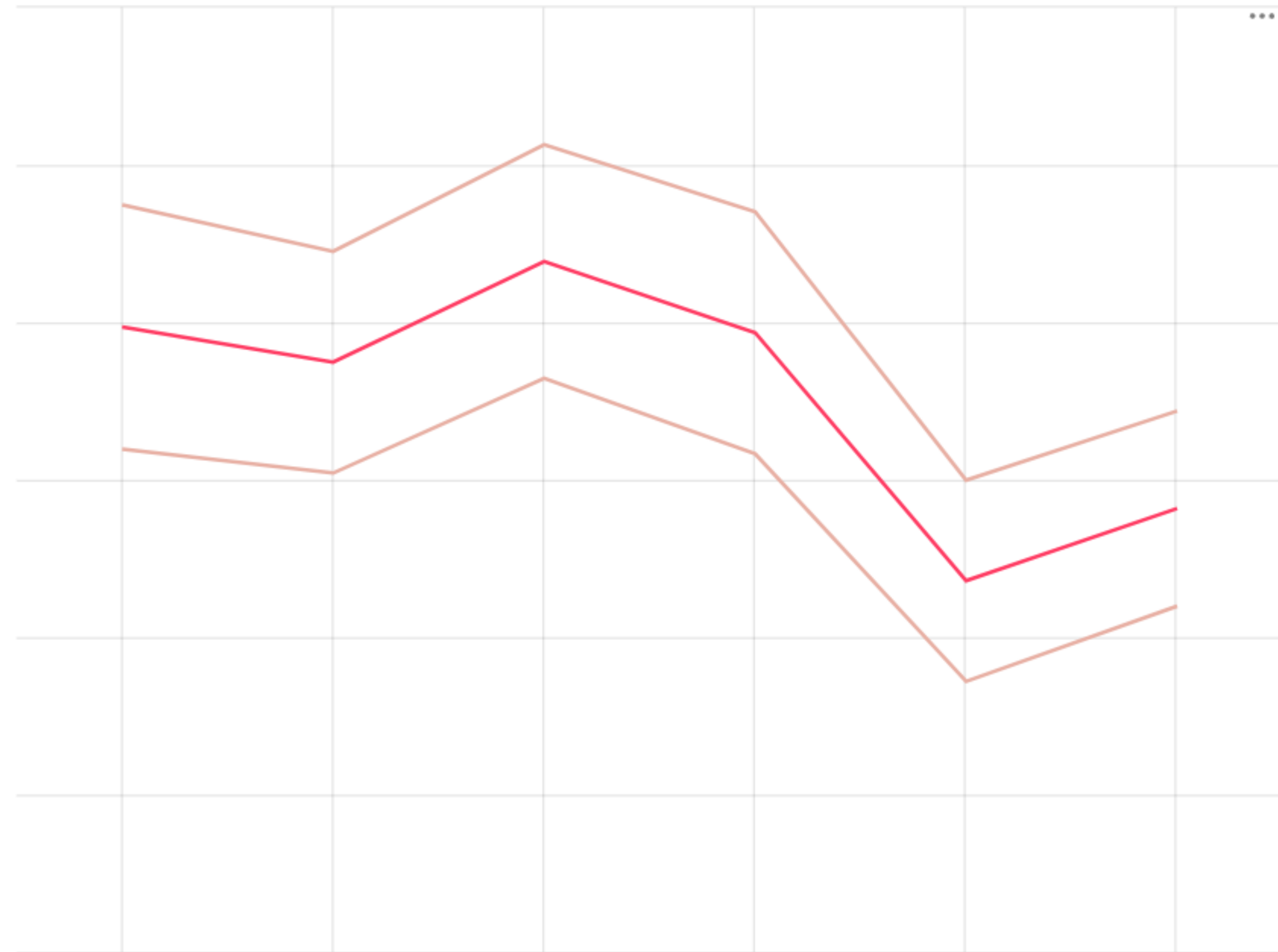
Доля организаций,  
у которой нашёлся  
хотя бы один дубль



# Итоговая метрика дублей

Чтобы понять значимость изменений отображаем доверительный интервал  $\pm 2$ -сигма, который получаем из расчёта дисперсии биномиального распределения.

Интервал уменьшается пропорционально корню из числа измерений



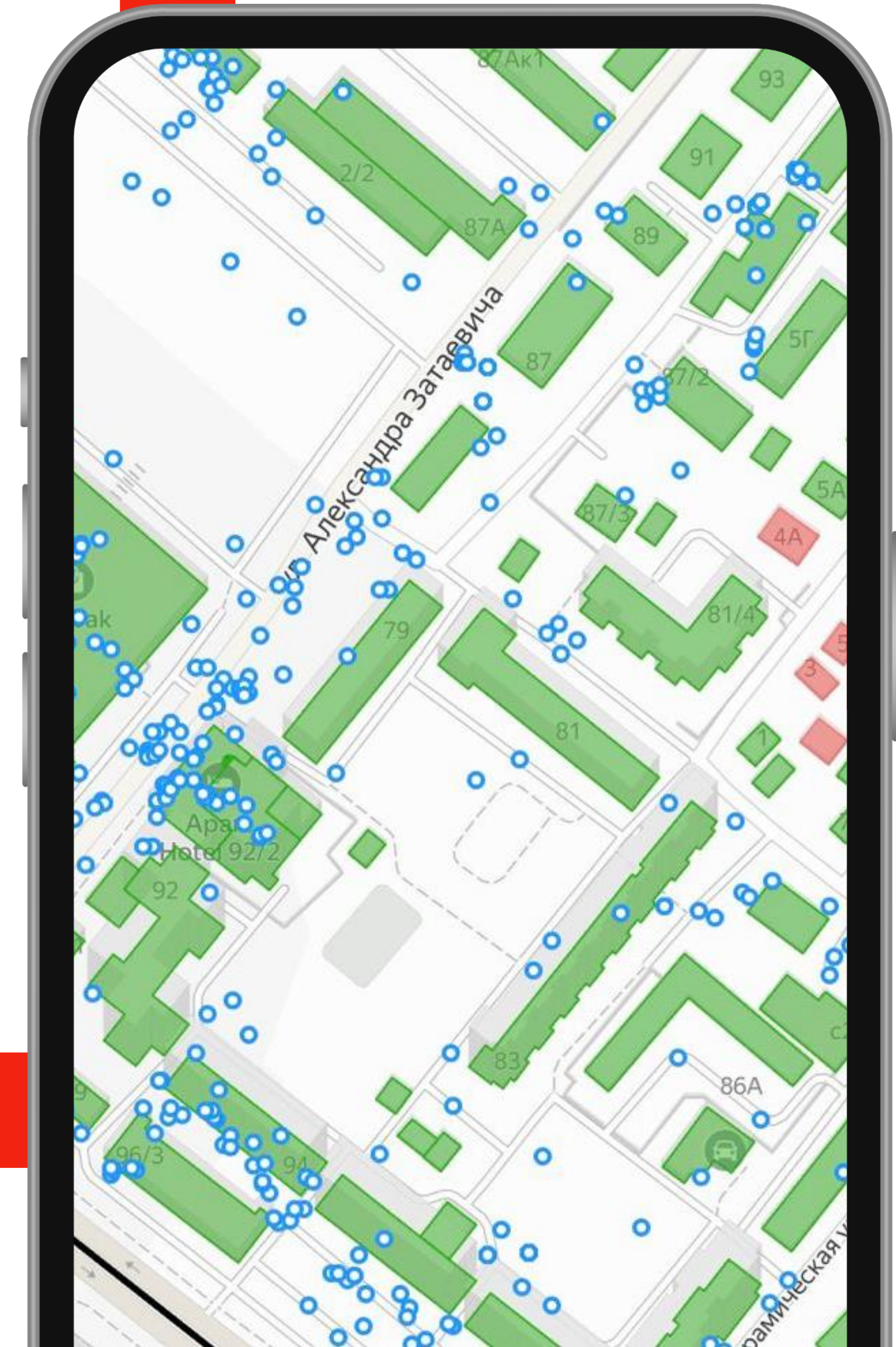


# Сверка с реальностью

# Пешеходы обходят заданную область и описывают реальность

## Но это не всегда возможно:

- Бывают организации, в которые не попадёшь без предварительной записи
- Бизнес центры
- Могут прийти в нерабочие часы





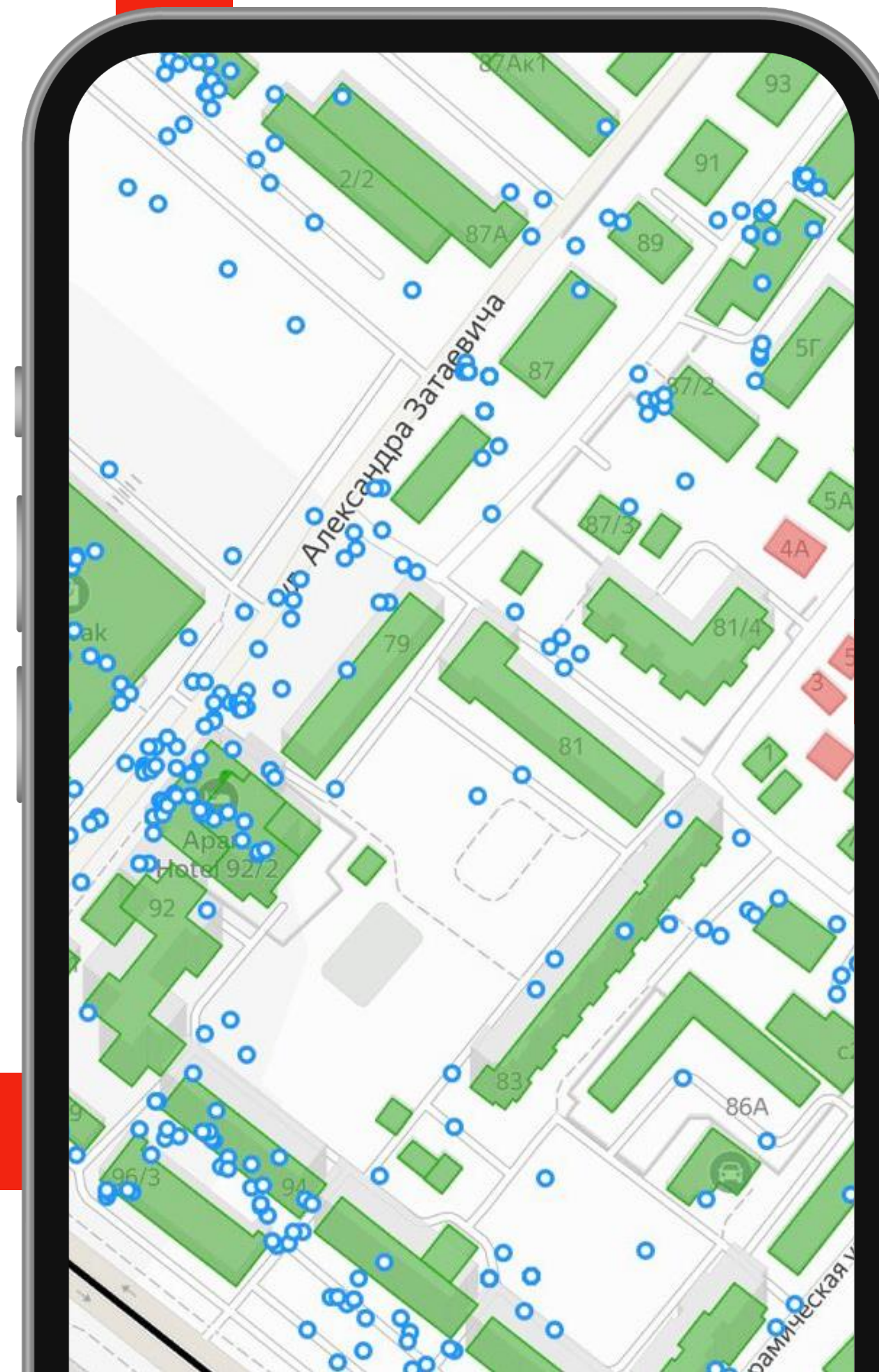
# Сверка с реальностью

## Пешеходы не идеальны

- Могут не найти организацию
- Могут подтвердить данные, не проверяя

## Что можно сделать:

- Добавлять honeypot по организациям и данным в них
- Наличие новой
- Отсутствие старой
- Позиция
- Название, расписание, телефон, адрес, сайт
- Координаторы отслеживают качество
- Качество влияет на вознаграждение пешеходов
- Отслеживать перемещения



# Как сверить данные?

## Названия

Новосибирская аптечная сеть = Муниципальная аптека?

Фудтрак1ff = 1 Food Factory?

## Телефоны

+7 (383) 347-00-01 = +7 (923) 777-42-84?

+7 (383) 335-42-08 = 8 (800) 234-56-48?

## Сайты

ekonika.ru = econika.ru?

ангарскаясосна.рф = angcosna.ru?

## Вид деятельности (рубрика)

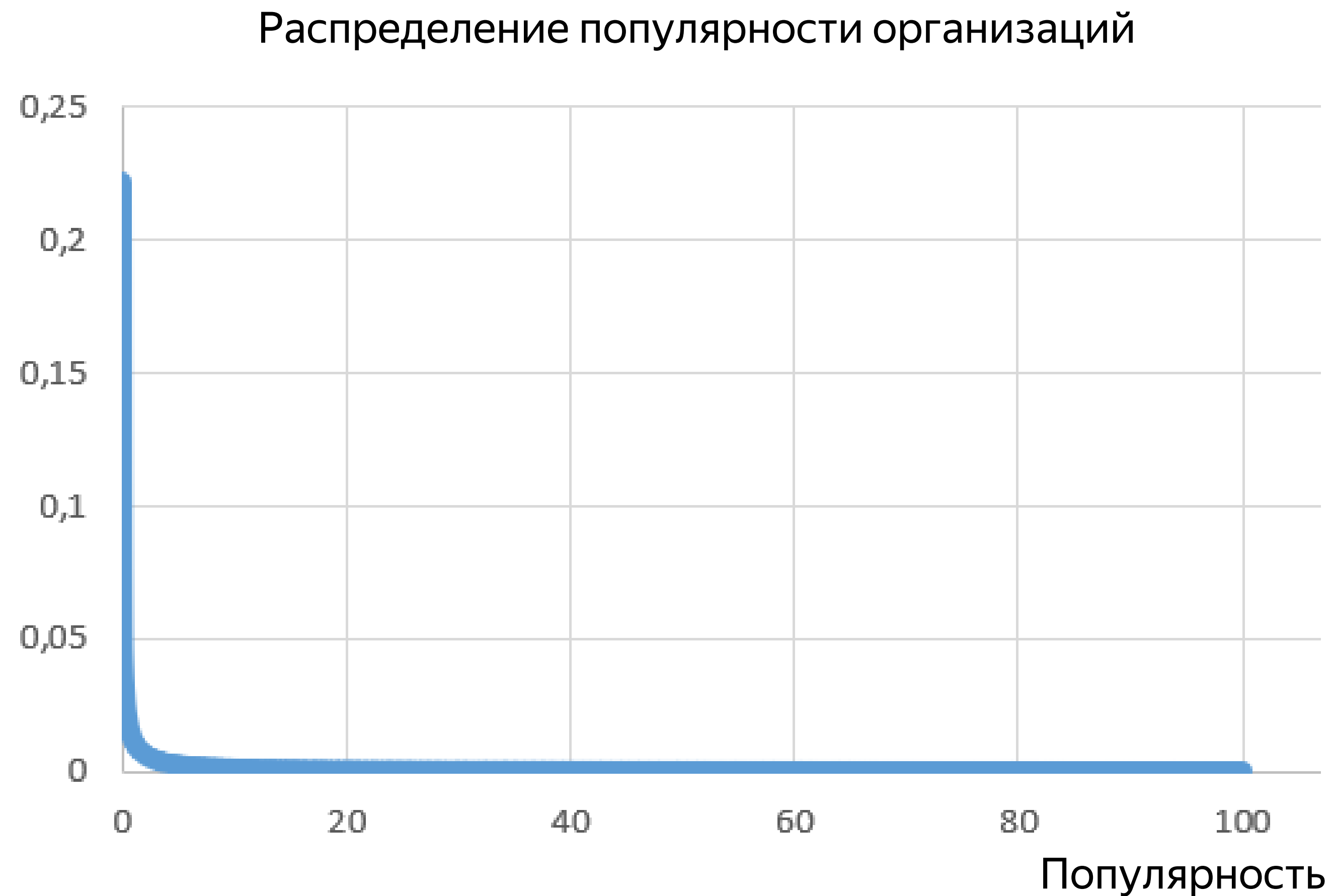
Ресторан, Кафе = Кофейня?

Поликлиника для взрослых = Пластическая хирургия?



# Правильное взвешивание

Ошибки в разных организациях затрагивают разное число пользователей





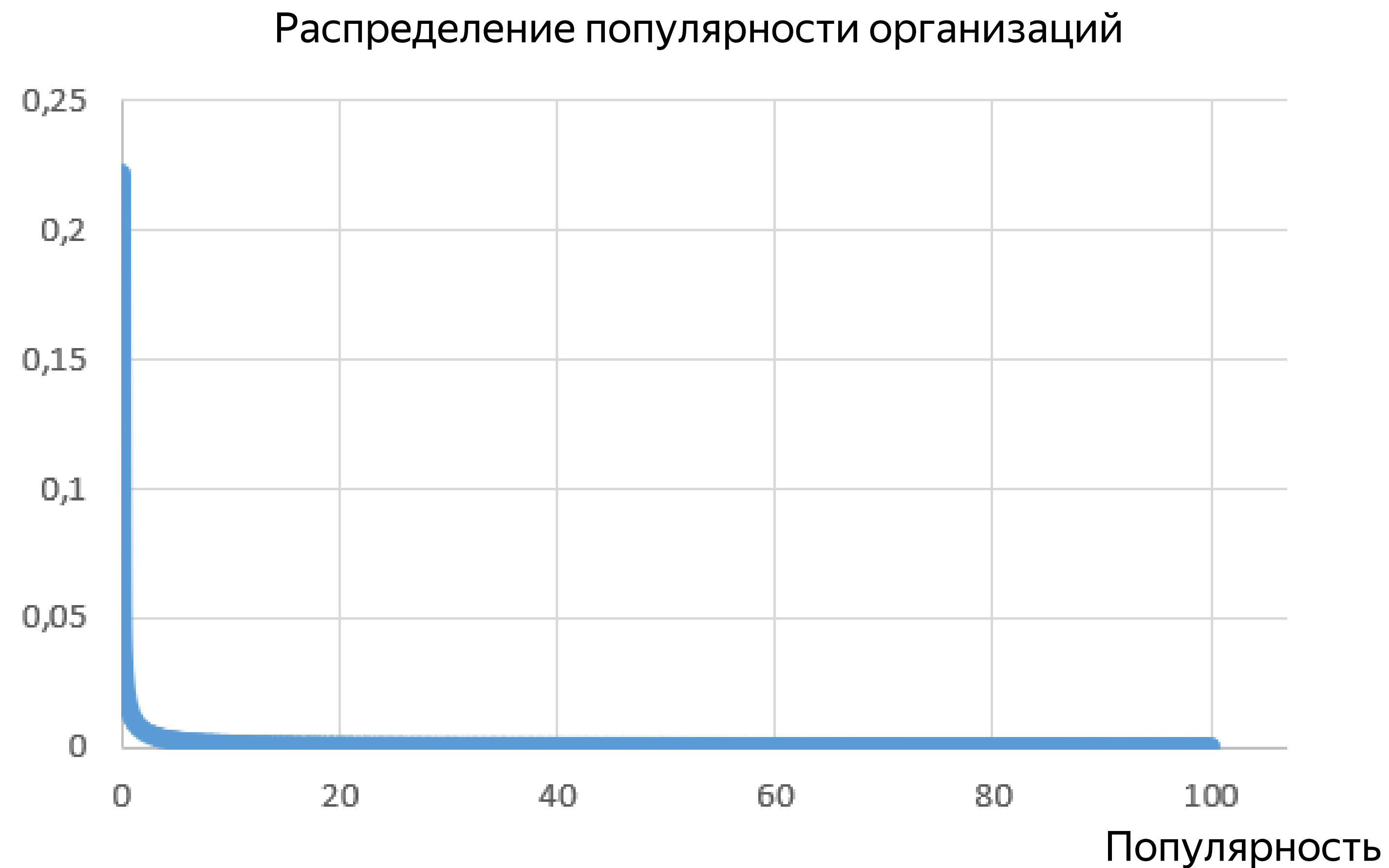
# Правильное взвешивание

Для оценки полноты нам нужно знать популярность новых компаний



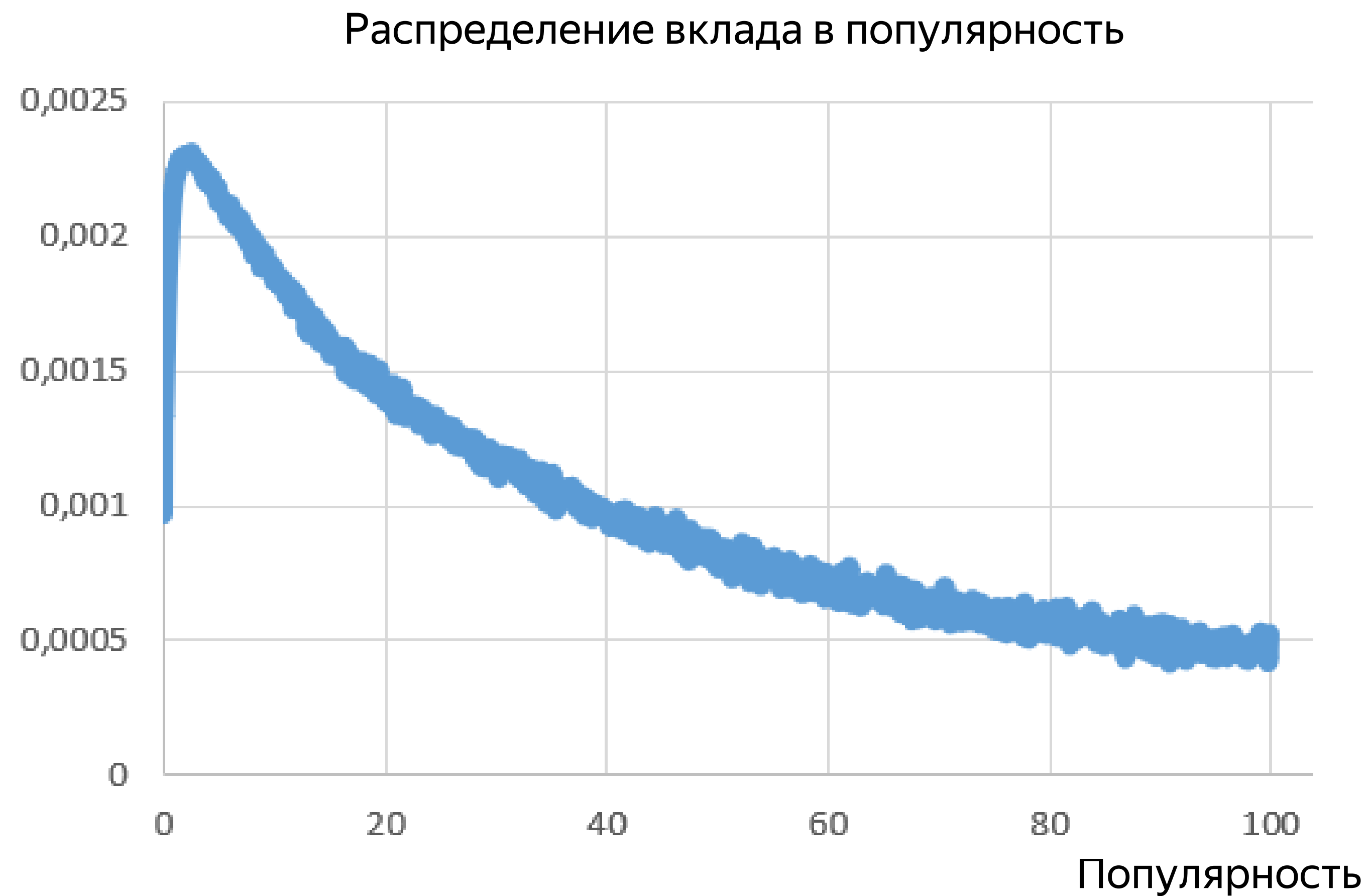
# Правильное взвешивание

Случайная выборка даст перекос в малопопулярные организации  
Но качество будет определяться отдельными популярными



# Правильное взвешивание

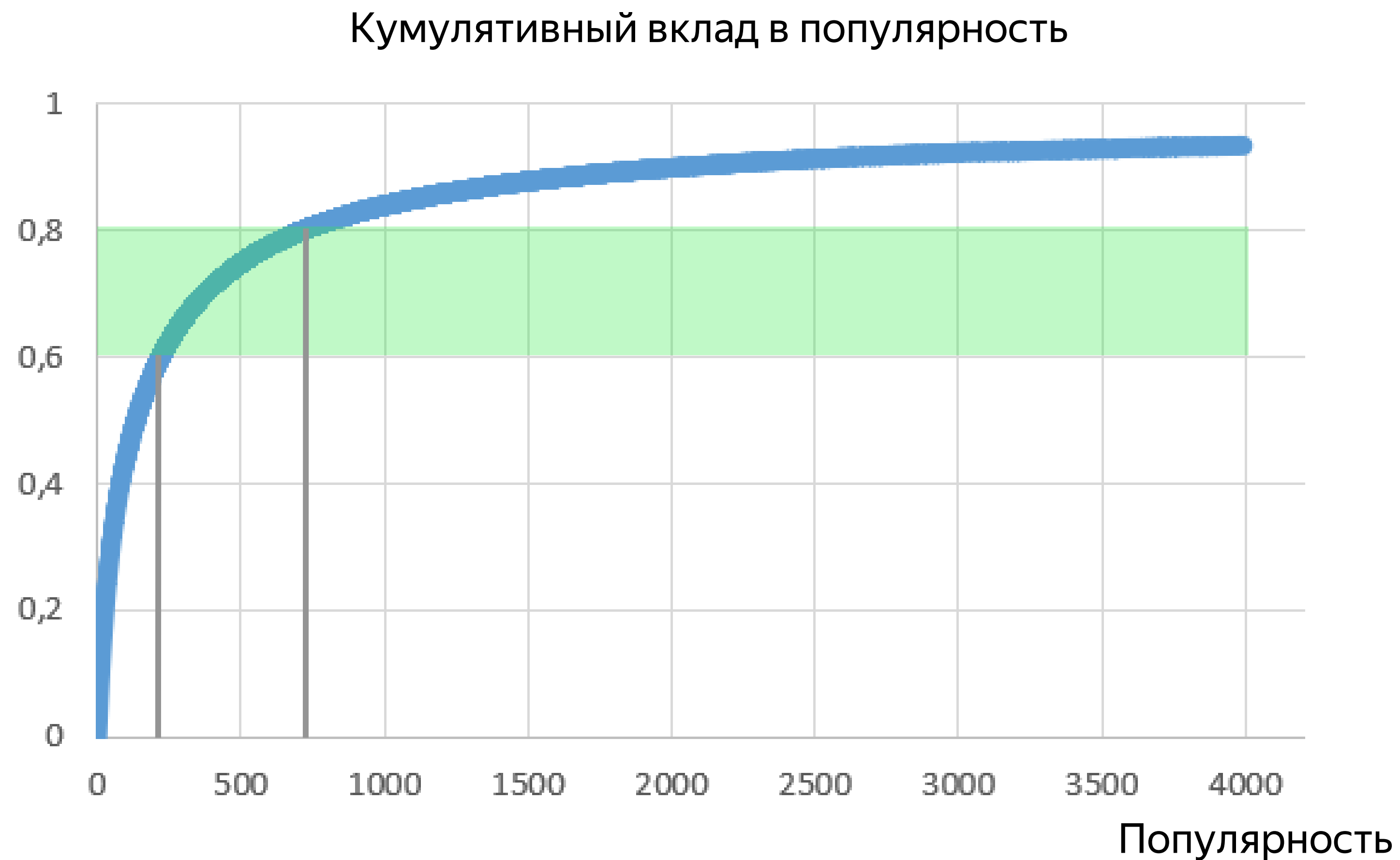
Число организаций помноженное на популярность.  
Уже нет острого пика у нуля.





# Правильное взвешивание

Взвешенное сэмплирование: делаем равномерную выборку вдоль оси Y



# Правильное взвешивание

Взвешенное сэмплирование: делаем равномерную выборку вдоль оси Y

$num\_of\_bins$

Число корзин

$q_i, w_i$

Качество и вес

$$weight\_of\_bin = \sum_i^{in\_bin} w_i$$

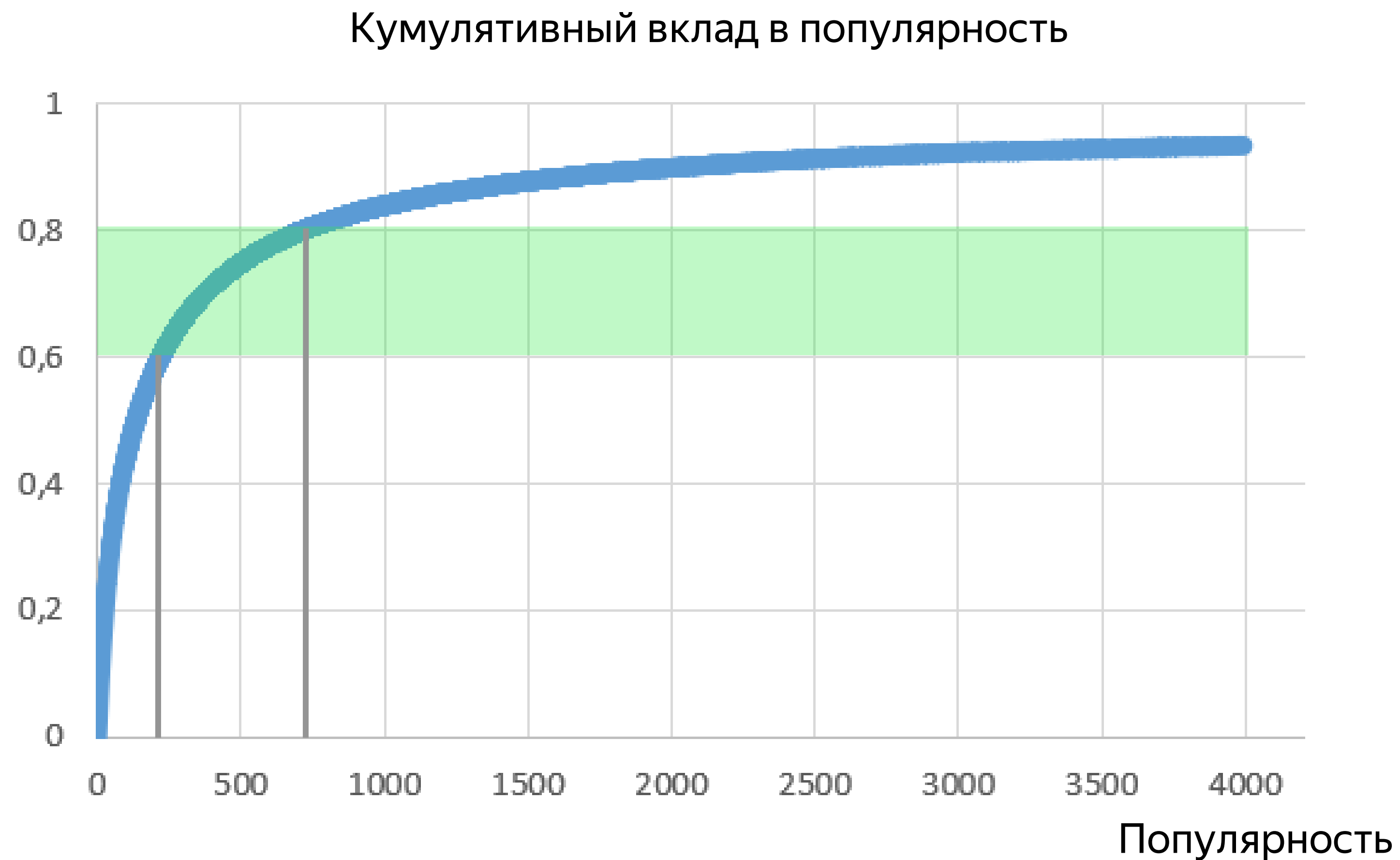
Суммарный вес в каждой корзине одинаков

$$quality = \frac{\sum_i^{all} q_i w_i}{\sum_i^{all} w_i} = \frac{\sum^{bins} \sum_i^{in\_bin} q_i w_i}{\sum^{bins} \sum_i^{in\_bin} w_i} \approx \frac{\sum^{bins} \sum_i^{in\_bin} q_i \frac{weight\_of\_bin}{N_{in\_bin}}}{\sum^{bins} weight\_of\_bin} = \frac{\sum^{bins} avr_{bin}(q)}{num\_of\_bins}$$

Взвешенное качество равно среднему качеству в каждой корзине с равным весом

# Правильное взвешивание

Взвешенное сэмплирование: делаем равномерную выборку вдоль оси Y





# Резюме

- ✓ Для измерения качества данных не подходят онлайн метрики
- ✓ Офлайн метрики построенные на людях требуют отладки и контроля: перекрытия, honeypot, кросс-валидация, метрики верности выполнения задания
- ✓ Нужно подбирать и вознаграждать исполнителей в зависимости от верности выполнения задания
- ✓ Когда нет понятного ответа, должен быть выбор «не знаю»
- ✓ Правильно выбирать и взвешивать задания