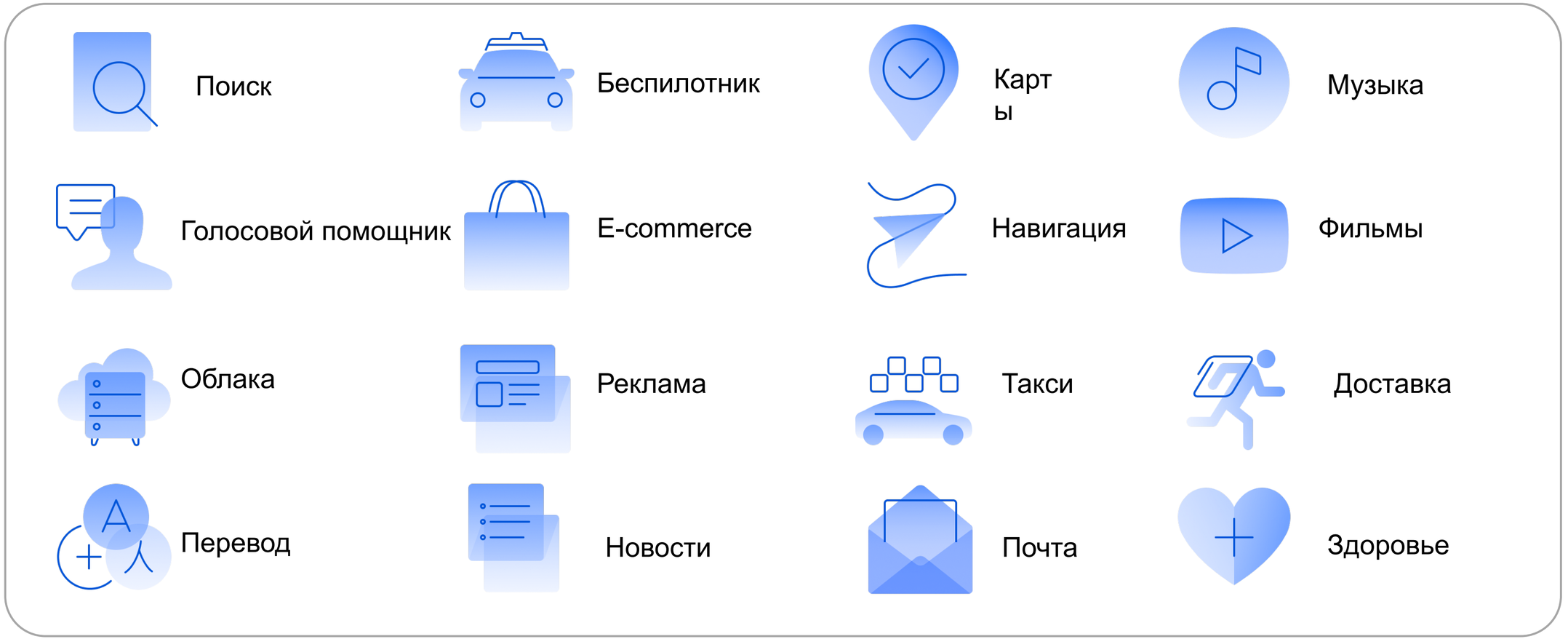
Toloka

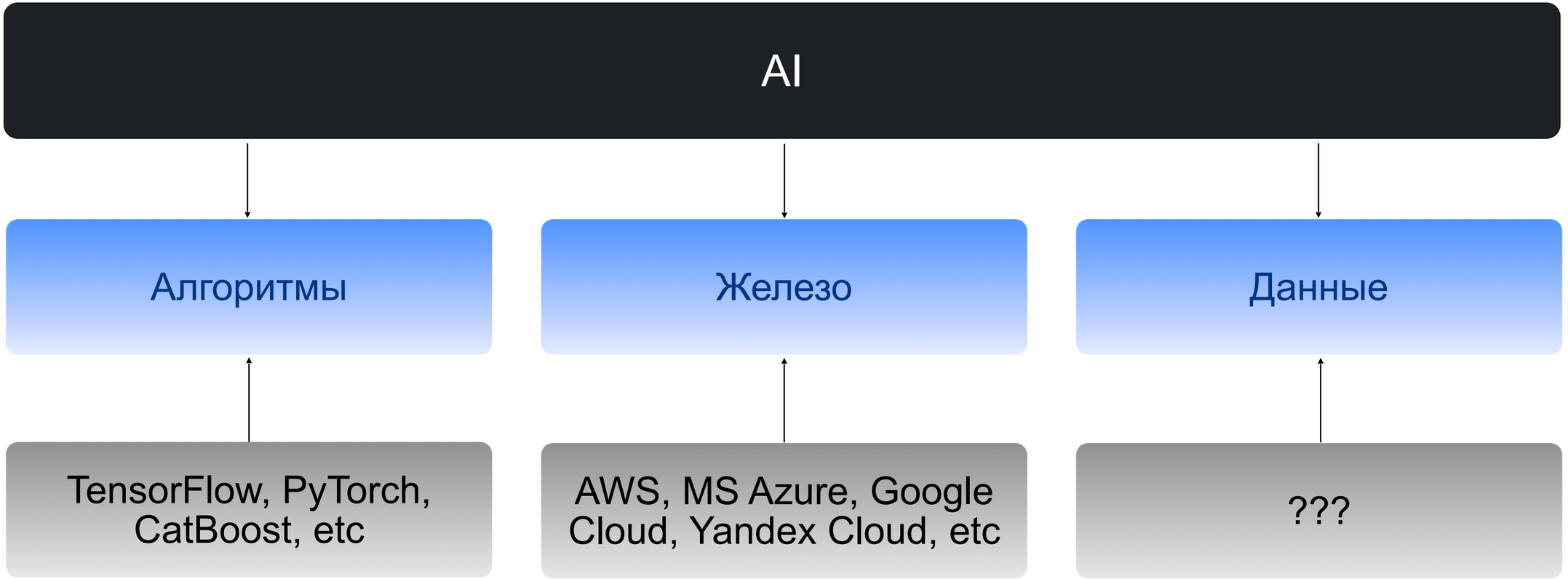
Сбор и разметка данных для машинного обучения

Евгения Суходольская, Data evangelist, Toloka

Какие технологии зависят от данных?



Три “столпа”


AI
Алгоритмы
Железо
Данные
TensorFlow, PyTorch, CatBoost, etc
AWS, MS Azure, Google Cloud, Yandex Cloud, etc
???

Три “столпа”: революция

AI

Алгоритмы

Железо

Данные

Привычный фокус

Потенциал роста

Можно ли использовать готовые датасеты?

Откуда берутся данные?

Kaggle:

→ Тебе выдали датасет - соревнование моделей

Реальность:

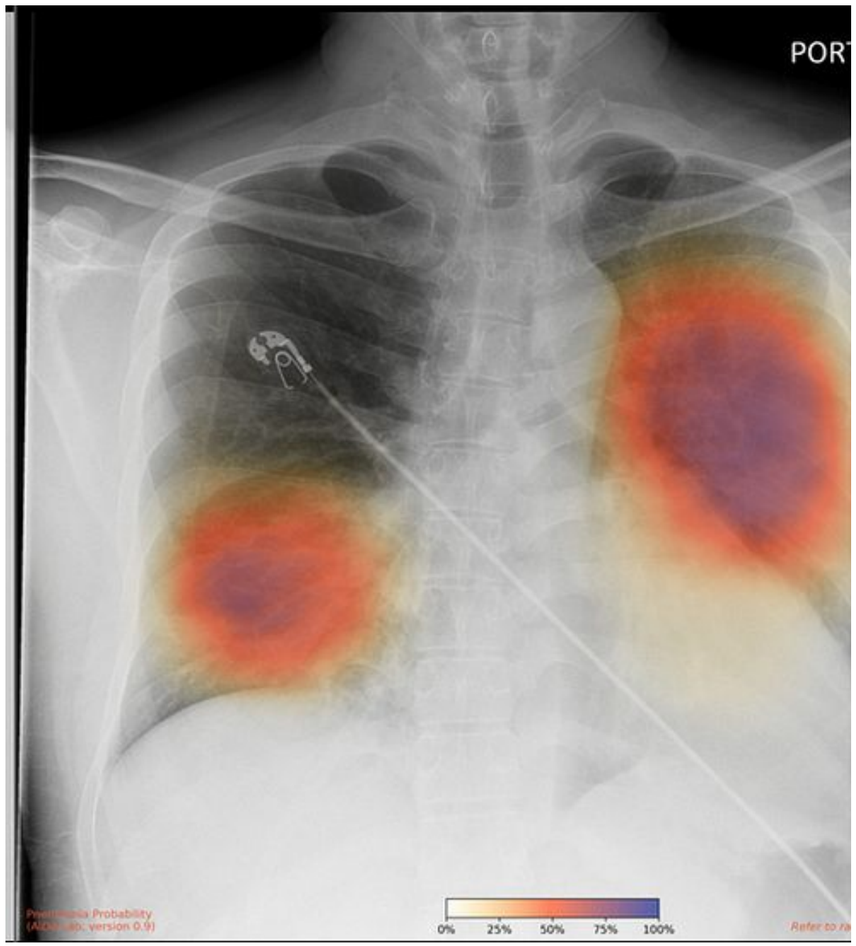
→ Соревнуются целые конвейеры ML производства

Пример 1: Перевод с токсичного

Original	Paraphrase
Reddit	
this is scaring the shit out of me.	This is really scaring me.
this is a joke , are you all fucking retards?	This is a joke, are you all crazy?
everybody is such a fucking pussy.	Everybody is acting cowardly.
did you think i was going to sit back and let some imbecile spew hatred towards me?	Did you think I was going to let an unreasonable person hate me?
calm the fuck down, cnn.	Please calm your nerves, CNN.
why is the scale of that graph so fucked up?	Why is the scale of that graph not placed rightly.

http://ceur-ws.org/Vol-2932/paper2.pdf

Пример 2: ML и COVID-19

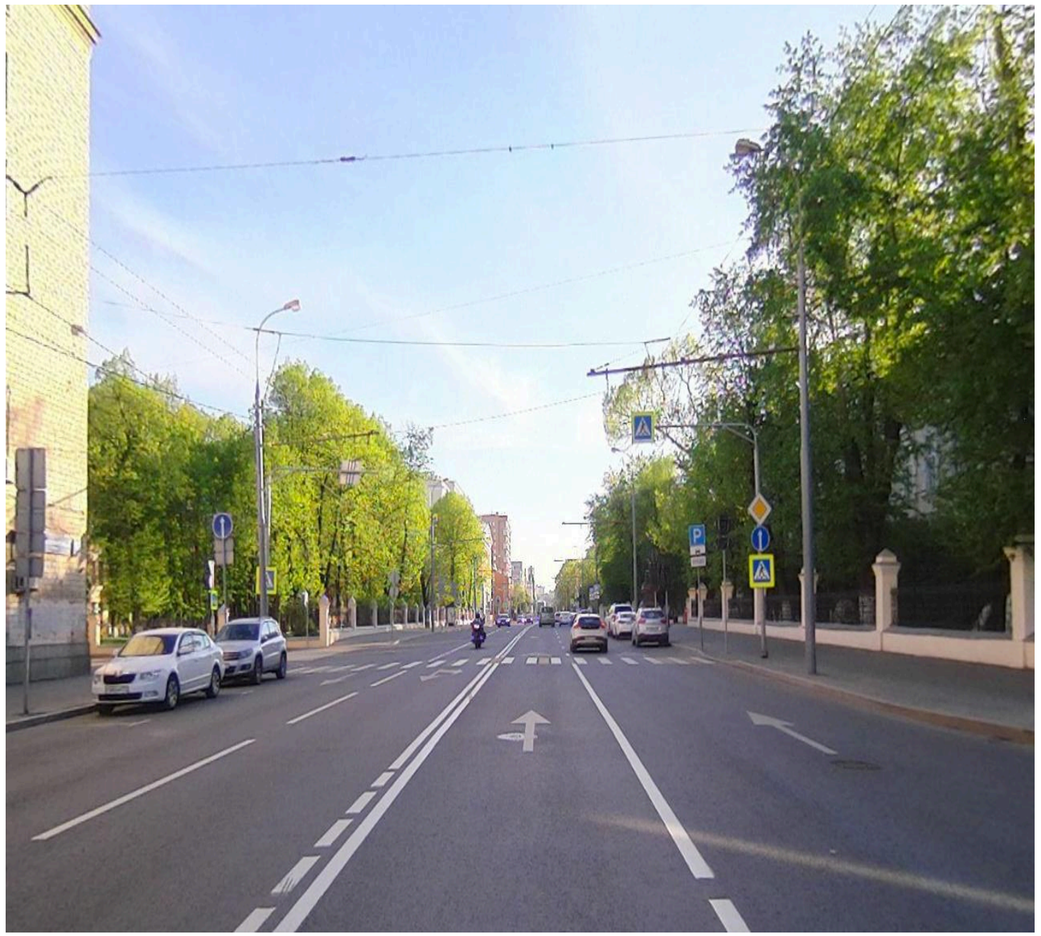


Были спроектированы сотни инструментов AI для распознавания covid-19. Они не сработали.

www.technologyreview.com

Изображение: https://www.bbc.com/news/business-52483082

Пример 3: Беспилотники



Тогда откуда брать датасеты?

На потребность в новой разметке влияют

Частота перезапуска модели

«Потолок» модели

Изменения в окружающем мире

Какие бывают подходы

- Синтетическая разметка
- Разметка внутри компании
- Аутсорсинг фрилансерам или вендорам
- Краудсорсинг

Синтетические данные

Генерирование данных с нужными параметрами

✓ Меньше ограничений на использование

✓ Экономит деньги и время

✗ Часто оказываются значительно хуже реальных данных

Разметка внутри компании

✓ Даёт предсказуемый результат

✓ Можно получить хорошую точность

✓ Прямое взаимодействие с разметчиками

✗ Долго

✗ Плохо масштабируется

Аутсорсинг

Найм временных сотрудников или фрилансеров

Заказ у внешней команды

✓ Не отвлекает на задачу команду data science

✓ Планирование ресурсов

✗ Меньше влияния на качество

✗ Может быть дорого или медленно

Краудсорсинг

Краудсорсинг - это что?

Абстрактное объемное задание

Облако исполнителей

Декомпозированная разметка

Краудсорсинг

✓ Скорость

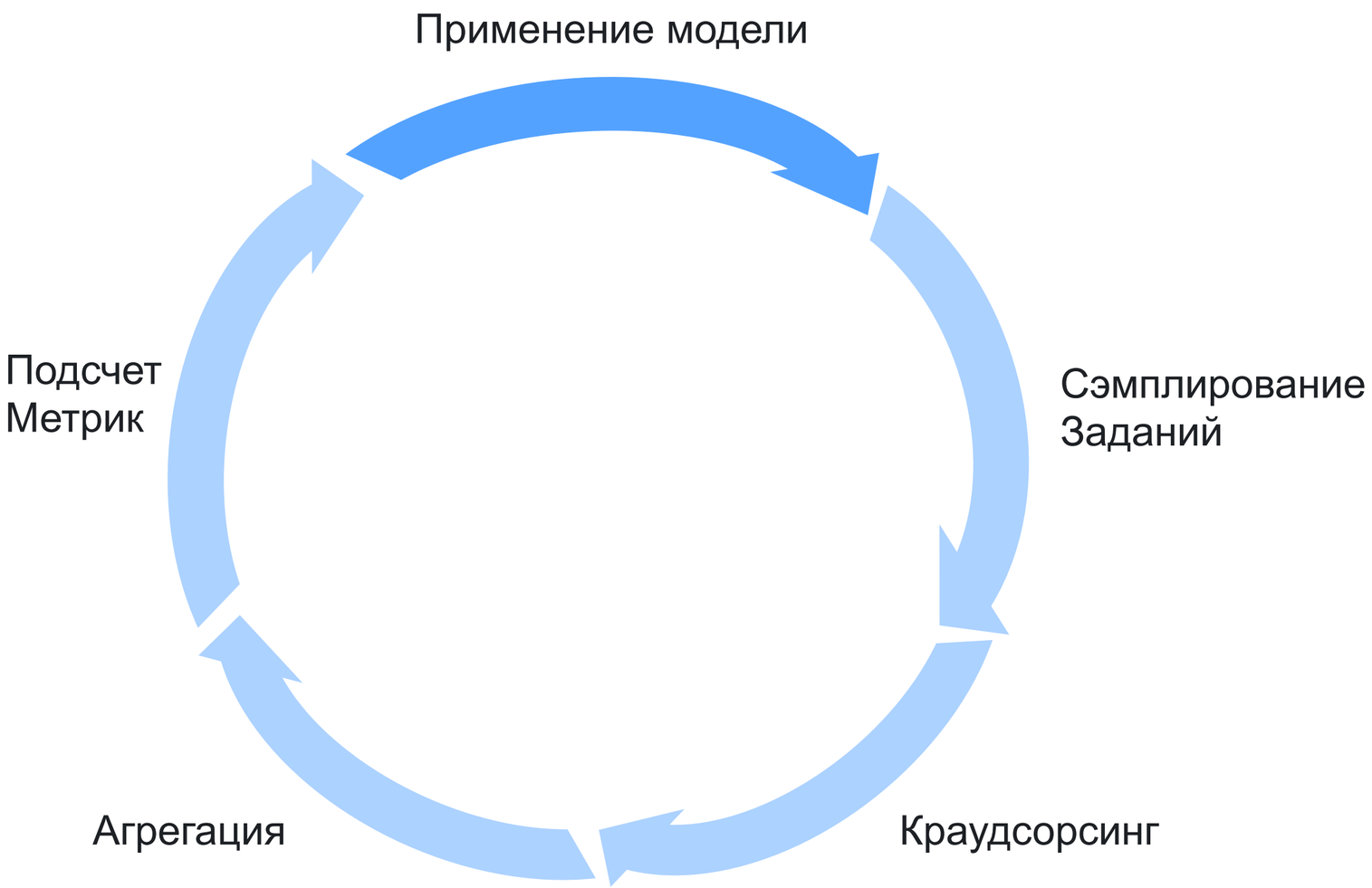
✓ Масштабируемость

✓ Настройка под задачу

✓ Экономичность

✗ Надо уметь добиваться «плюсов»

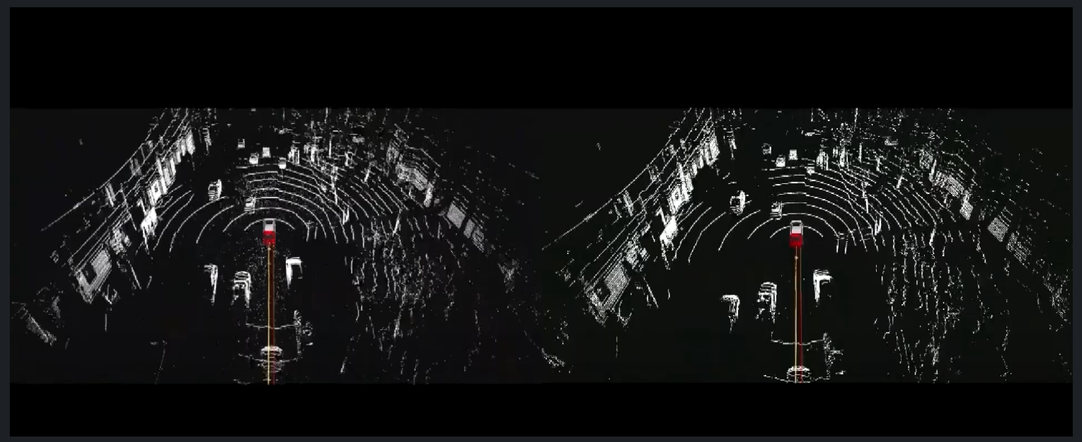
✗ Нужны инструменты контроля качества

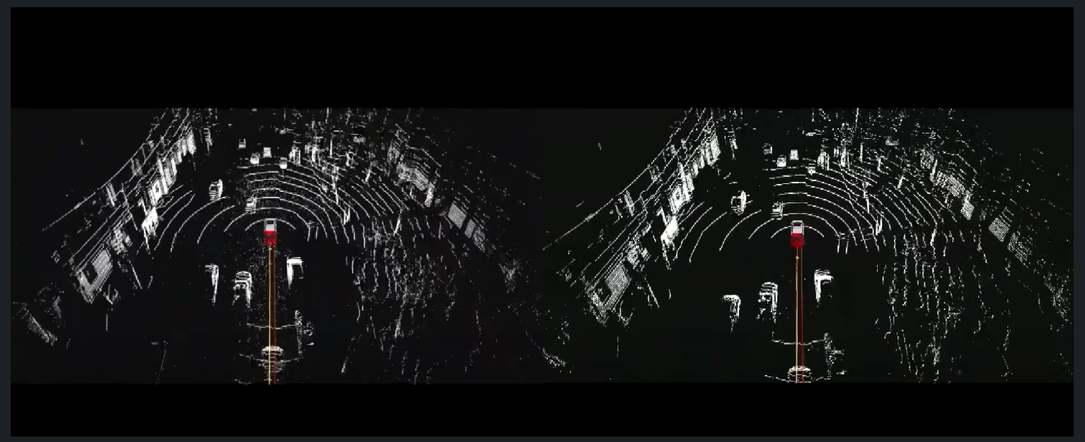

Применение модели
Сэмплирование
Заданий
Краудсорсинг
Агрегация
Подсчет
Метрик

Реальные кейсы

Беспилотники

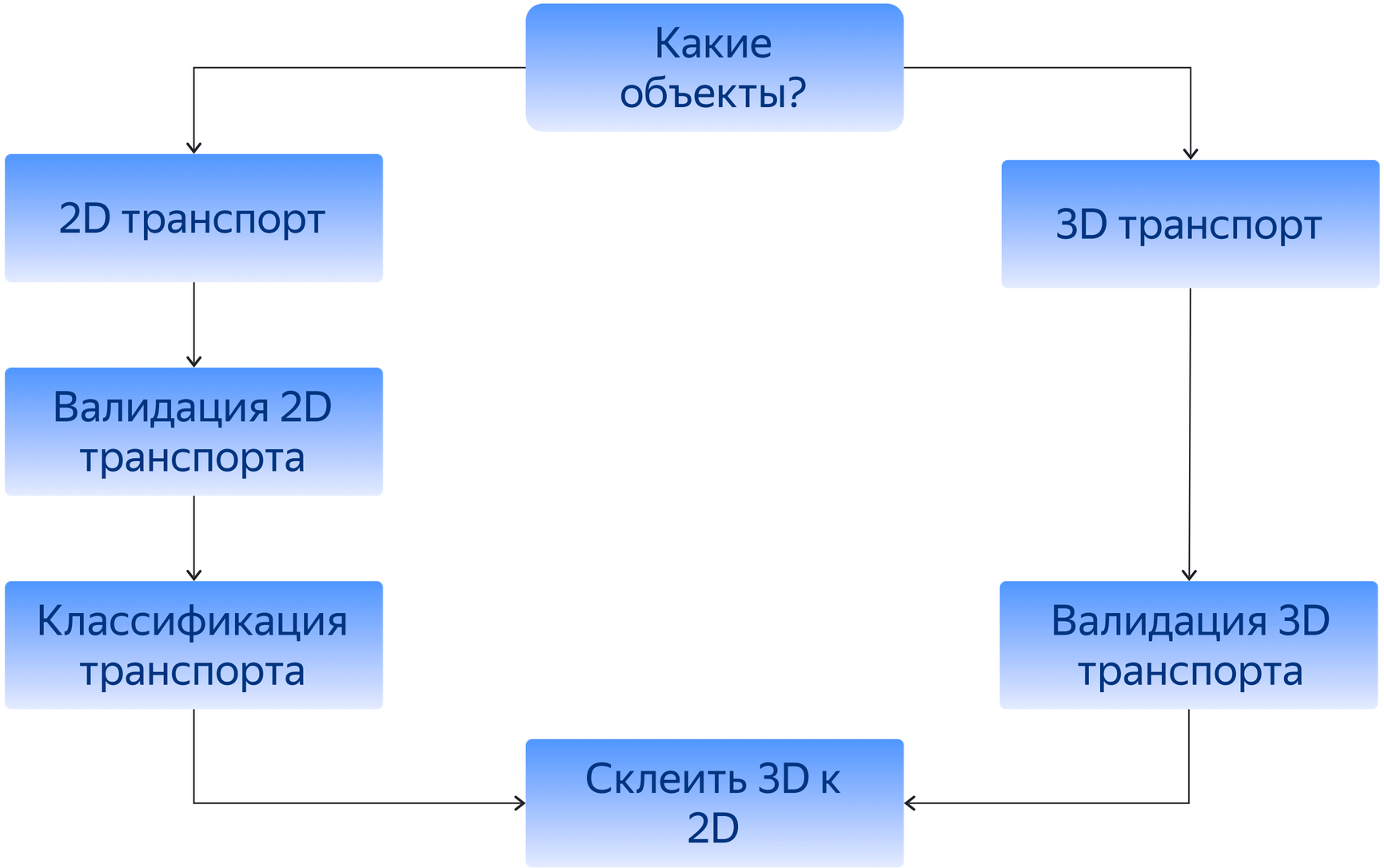
Active learning





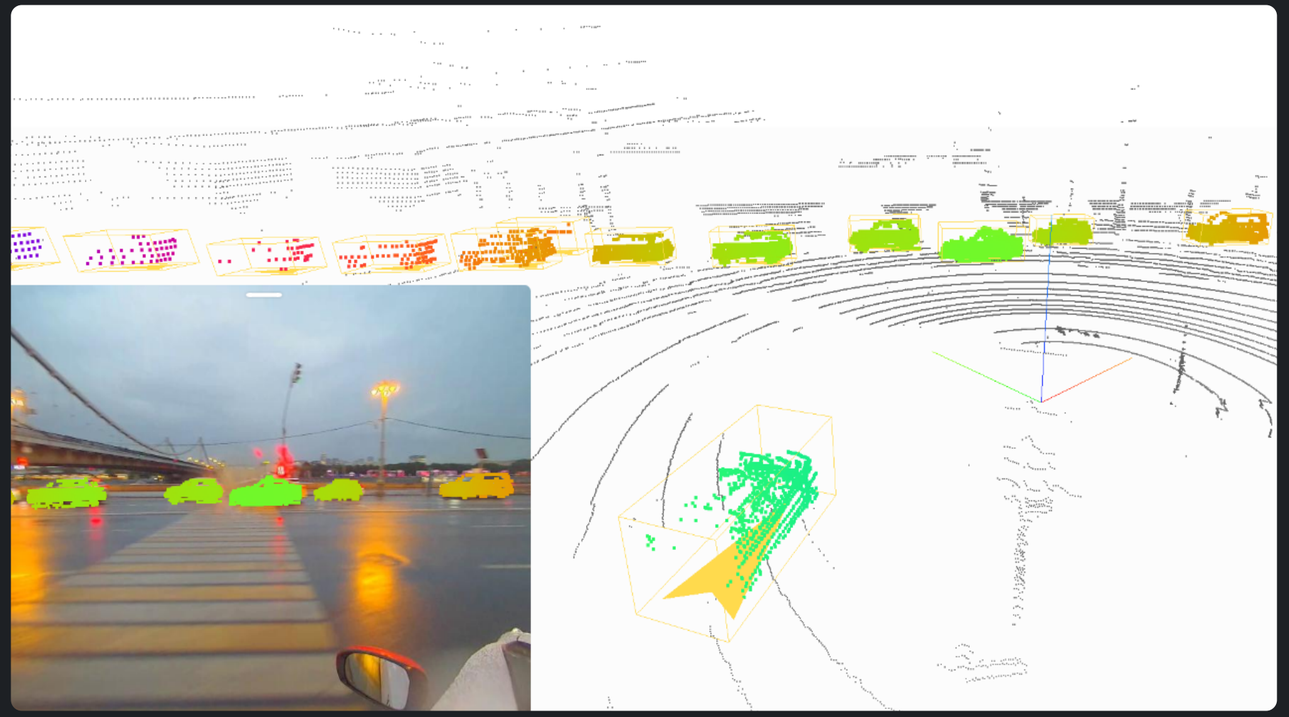
Беспилотники: разметка 3D транспорта

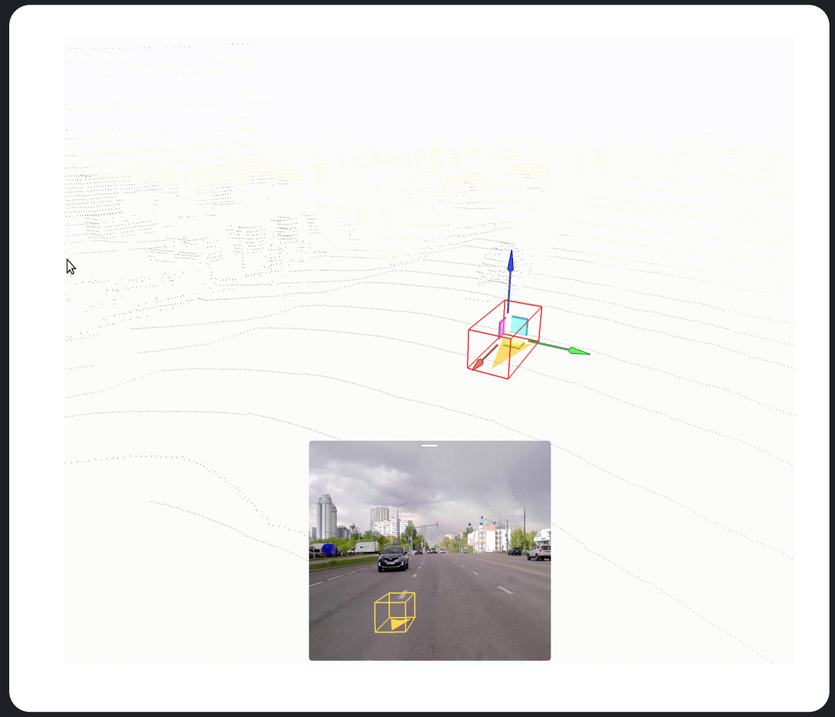
Декомпозиция задачи



Разметка 2D

Разметка 3D-объектов



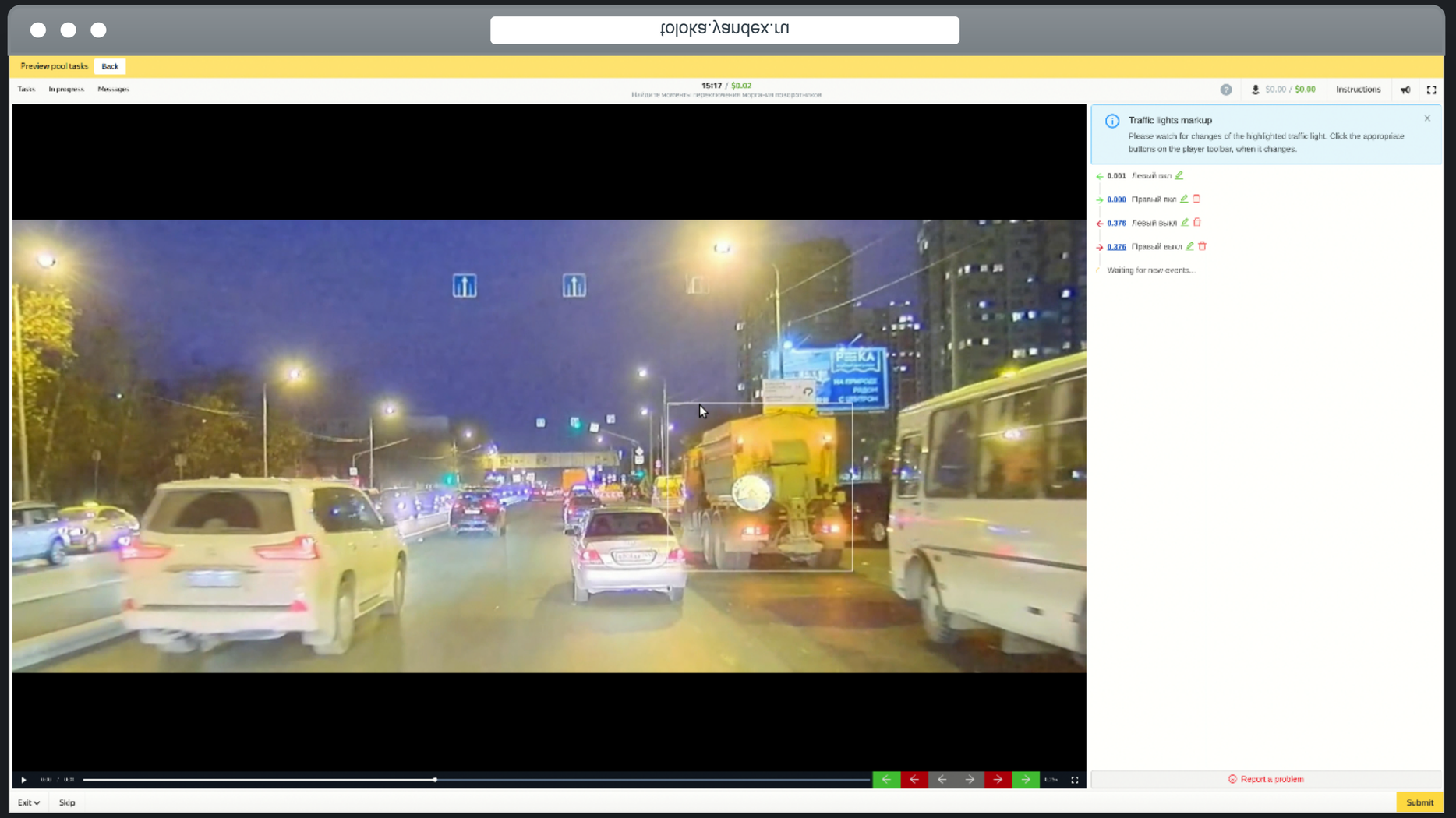


Беспилотники: разметка видео

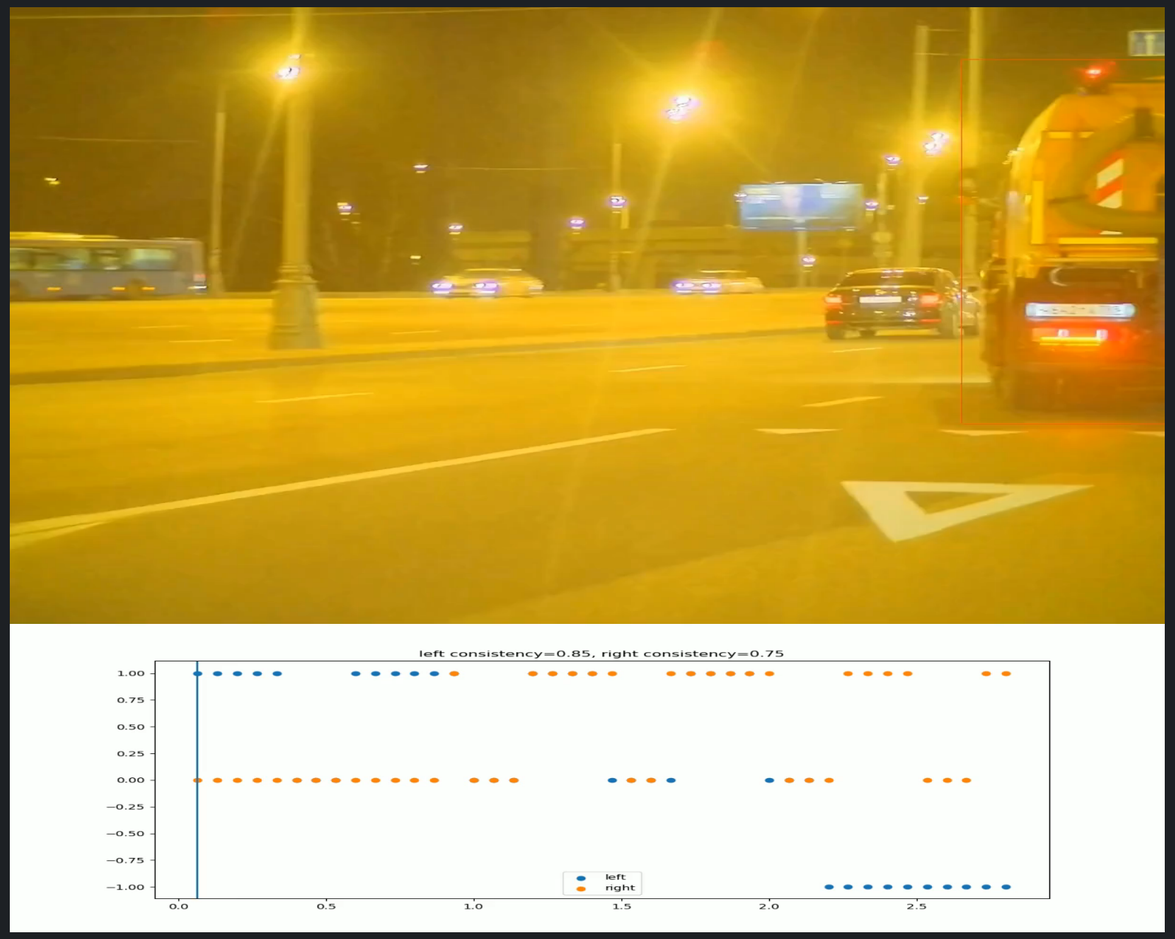
Для некоторых задач нужна информация о последовательности сцен, не об отдельных сценах

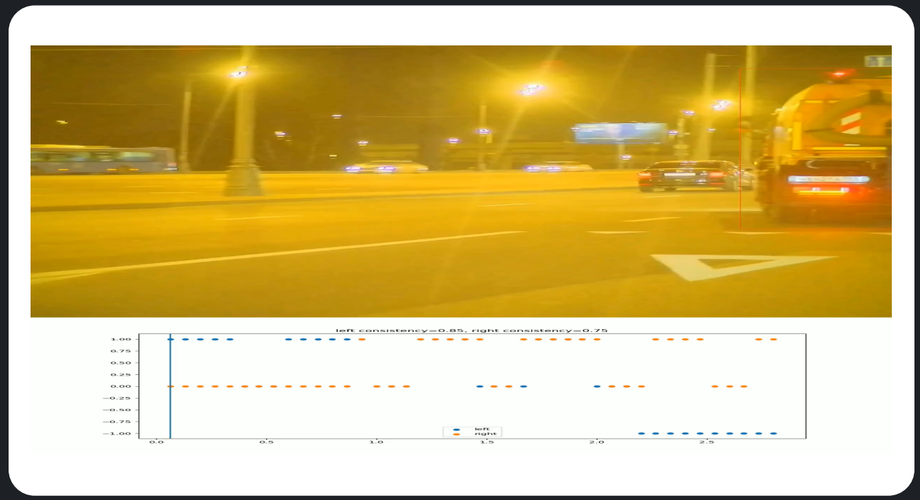
1. Трекинг объектов
2. Повторяющиеся процессы

Интерфейс


toloka.yandex.ru
Preview pool tasks
Back
Tasks
In progress
Messages
15:17 / $0.02
$0.00 / $0.00
Instructions
Traffic lights markup
Please watch for changes of the highlighted traffic light. Click the appropriate buttons on the player toolbar, when it changes.
0.001 Левый вкл
0.000 Правый вкл
0.376 Левый выкл
0.376 Правый выкл
Waiting for new events...
Report a problem
Exit
Skip
Submit

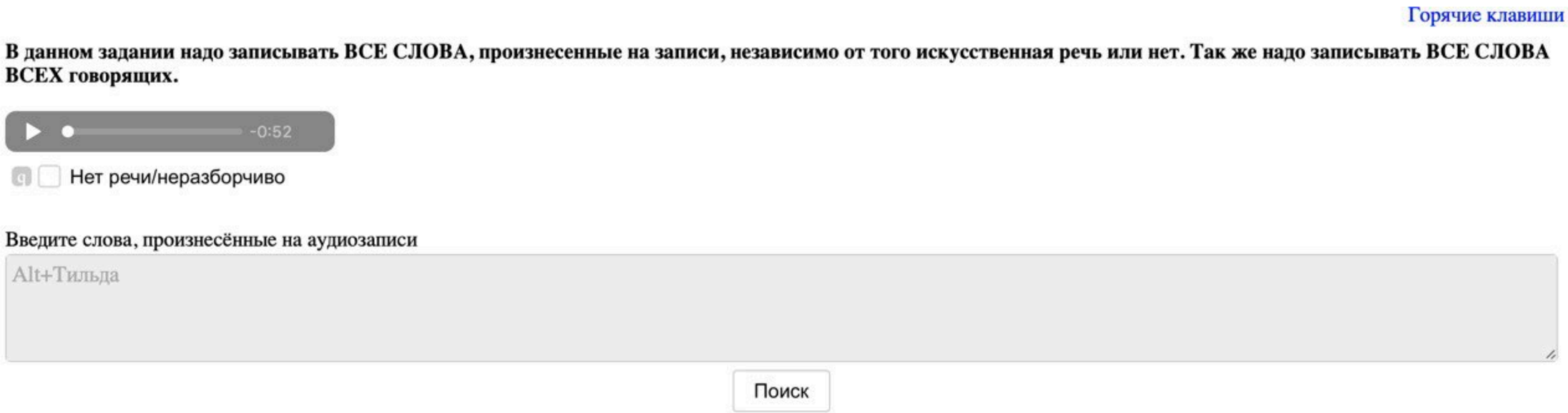
Результаты разметки видео





Алиса

Аннотирование



Инструкция

В данном задании требуется прослушать звуковой отрывок и записать произнесенные слова в том порядке, в котором они прозвучали. Задание лучше выполнять в тишине, с наушниками. В случае необходимости прослушать запись необходимо несколько раз.

Порядок выполнения задания

ОБЩАЯ СТРАТЕГИЯ

Сначала записываем текст в соответствии с правилами ниже, потом выбираем нужную radio button из первых трех, а потом - проставляем галочки в оставшихся пяти чекбоксах.

Если на записи **есть речь**, определите количество говорящих, отметьте соответствующий вариант и запишите то, что вы услышали в поле "Текст аннотации". В тексте аннотации при этом должна содержаться только речь, обращенная к Алисе. Если есть сомнения, по умолчанию речь считаем обращенной к Алисе. Если есть речь, не направленная к Алисе, то выбираем вариант **"Есть речь, НЕ обращенная к Алисе"**.

Пример

Если на аудиозаписи **нет речи**, говорящие издают нечленораздельные звуки, напевают мелодию без слов, либо вы не можете разобрать ни одного слова - отметьте вариант "Нет речи/неразборчиво".

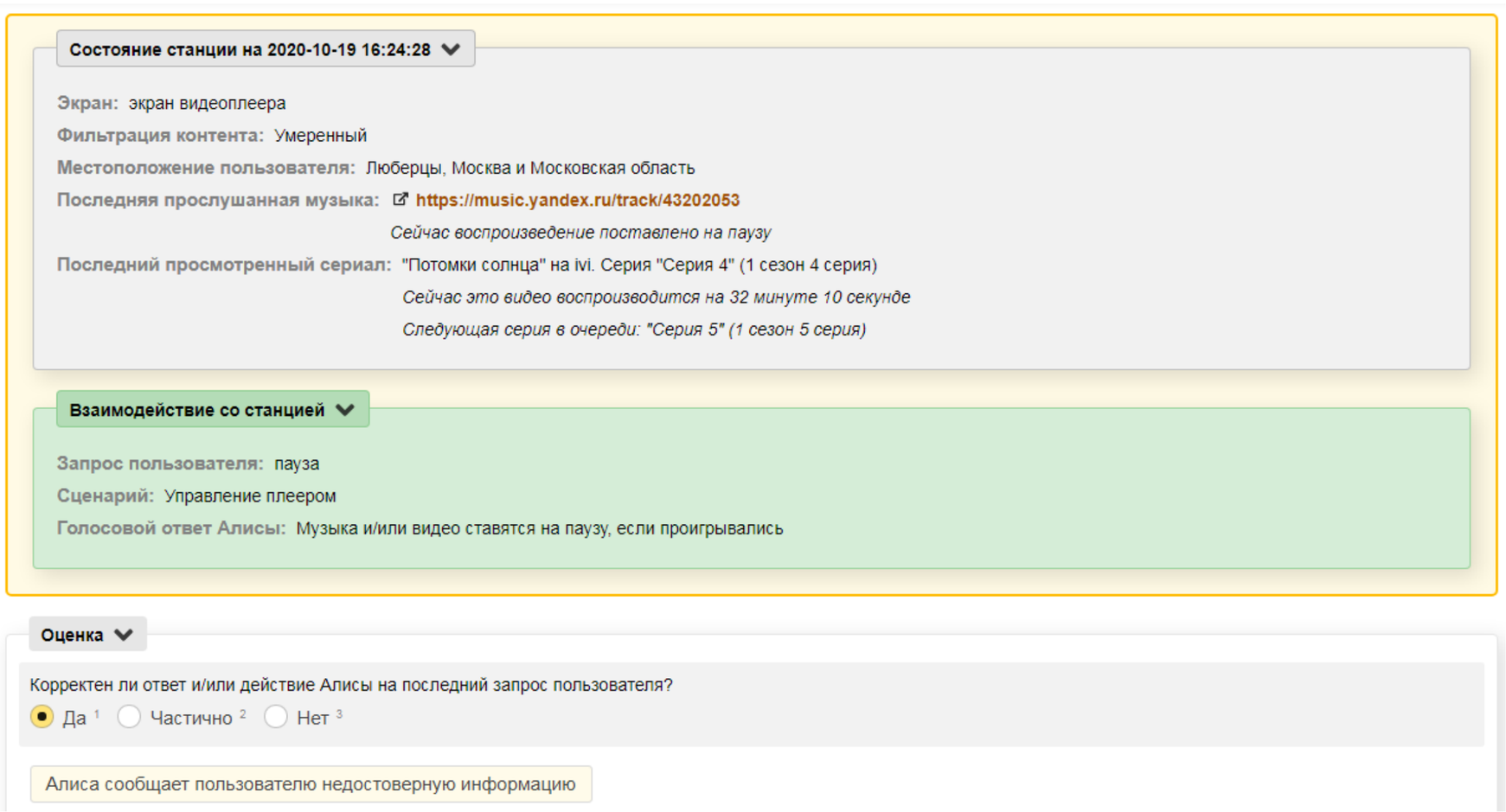
Если часть речи можно расшифровать, **то расшифровываем понятные слова и неразборчивые заменяем символом "?" (см. ниже)**.

Даже если запрос вопросительный - в конце не ставим символ "?"

Если на записи присутствуют голоса нескольких человек, слова которых можно разобрать, либо один человек говорит на фоне различимой речи из телевизора, радио или на фоне речи "Алисы", необходимо записывать ТОЛЬКО основного говорящего. Обратите внимание: играющая из телевизора или радио музыка не считается речью, если Вы не можете различить отдельные слова.

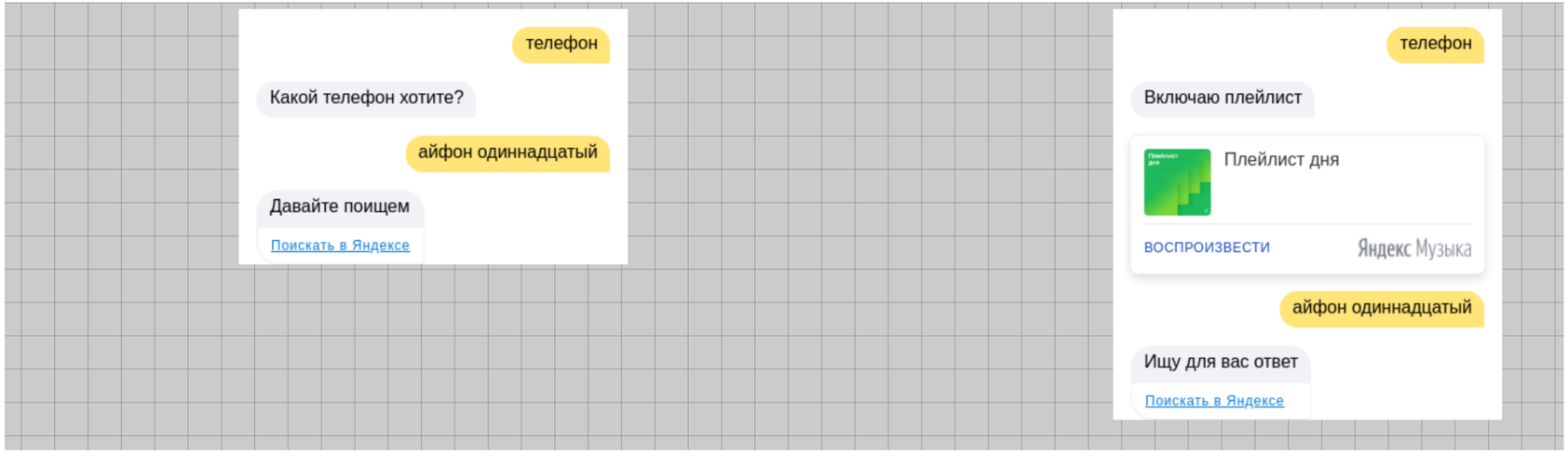
Алиса: релевантность

UE2E (Uber End-To-End)



Попарное сравнение сессий

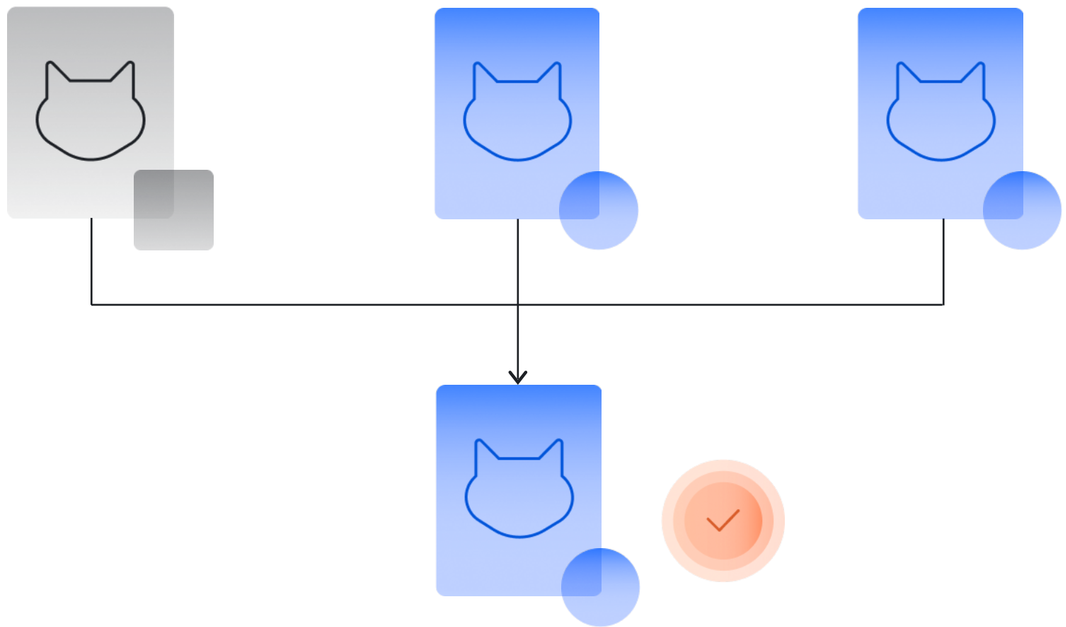
Текущая версия 56.9% screen_09bf-479b21d0... 43.1% Новая версия



Комментарии выбравших “Текущая версия”	Комментарии выбравших “Новая версия”
Левый лучше потому-что человек может хочет купить телефон,а в правом он включает музыку,которая возможно и не нужна человеку в данный момент	Фраза ищу для вас ответ в этом варианте не так режет слух, вариант справа не понятн главное с кем собирается искать.
Ответ Алисы слева лучше, так как чётко отвечает на запрос пользователя по поводу поиска информации о телефоне айфоне 11	Мне больше нравится диалог в правом варианте.
Алиса уточняет модель телефона, не включает плейлист. После уточнения начинает поиск.	Более содержательный диалог
Правый скриншот лучше, т.к. Алиса поняла тему, на которую интересуется пользователь.	Интереснее вариант справа.
	нравится наличие плейлиста

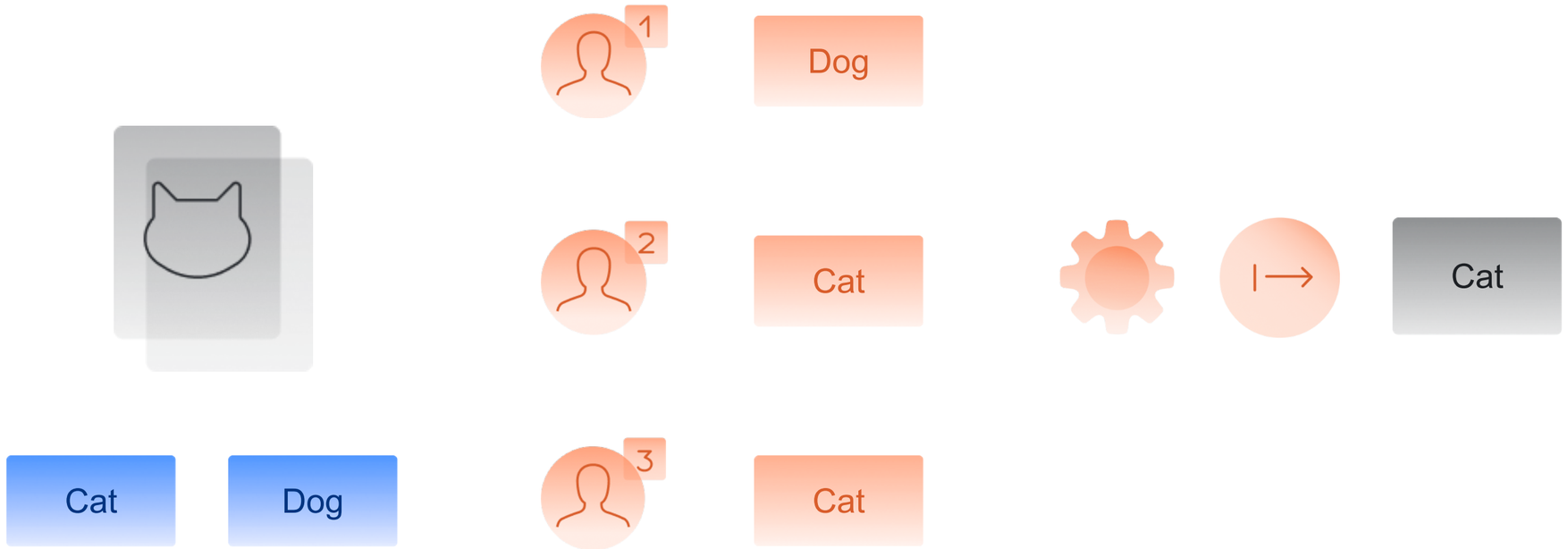
Как получить качественную разметку?

Перекрытие



Как агрегировать краудсорсинговые аннотации?

Мнение большинства



Взвешенное мнение большинства

Cat Dog

Cat

Качество в среднем:
0.20

Cat

Качество в среднем:
0.60

Dog

Качество в среднем:
0.90

Мнение большинства

Cat	Dog
1 1	1
2	1

Cat

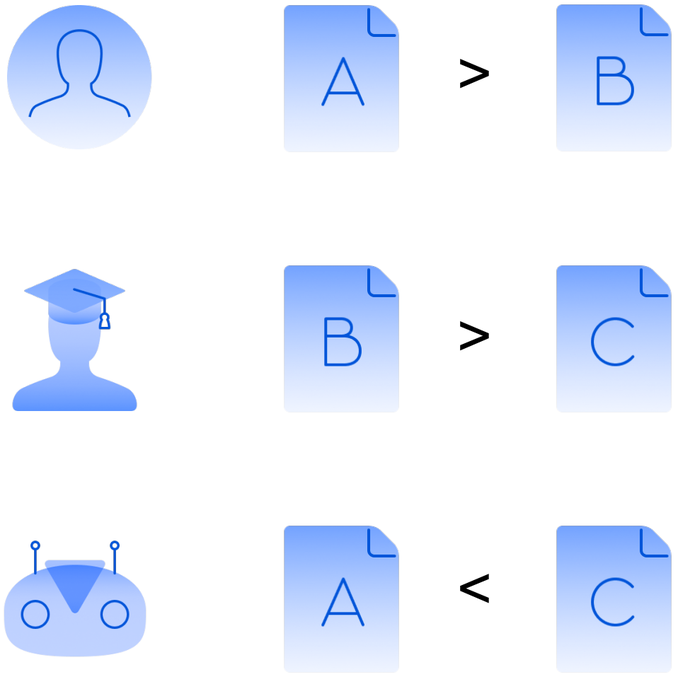
Взвешенное мнение большинства

Cat	Dog
0.2 × 1 0.6 × 1	0.9 × 1
0.8	0.9

Dog

Если данные не категориальные

Попарные сравнения



Текст

{’a b c d’, ‘b z d e’, ‘b c d e f’}

Сегментации

Crowd-Kit

Интерфейсы & методы

Crowd-Kit

- BaseClassificationAggregator
- BaseEmbeddingsAggregator
- BaseTextsAggregator
- BasePairwiseAggregator
- BaseImageSegmentationAggregator

fit()

predict()

predict_proba()

predict_scores()

fit_predict()

Входные данные

Crowd-Kit

- Ответ
- Исполнитель
- Задание

Методы агрегации

Crowd-Kit

Категориальные данные

Method
Majority Vote
Dawid-Skene
Gold Majority Vote
M-MSR
Wawa
Zero-Based Skill
GLAD

Текст

Method
RASA
HRRASA
ROVER

Сегментация

Method
Segmentation MV
Segmentation RASA
Segmentation EM

Попарные сравнения

Method
Bradley-Terry
Noisy Bradley-Terry

Датасеты

Crowd-Kit

Relevance 2

Relevance 5

Crowdspeech

IMBD-WIKI-SbS

```
from crowdkit.datasets import get_datasets_list

[name for name, description in get_datasets_list()]
```

```
['relevance-2',
 'relevance-5',
 'mscoco',
 'mscoco_small',
 'crowdspeech-dev-clean',
 'crowdspeech-test-clean',
 'crowdspeech-dev-other',
 'crowdspeech-test-other',
 'imdb-wiki-sbs',
 'nist-trec-relevance']
```

Метрики

Crowd-Kit

- Данные ⟶ alpha_krippendorff
  consistency
  uncertainty
- Исполнители ⟶ accuracy_on_aggregates

Какую задачу вы бы хотели попробовать решить с помощью краудсорсинга?

Спасибо!

Евгения Суходольская

Data Evangelist

sukhodolskaya@toloka.ai

https://toloka.ai/