

Подготовка данных к обучению



Зачем нужны данные

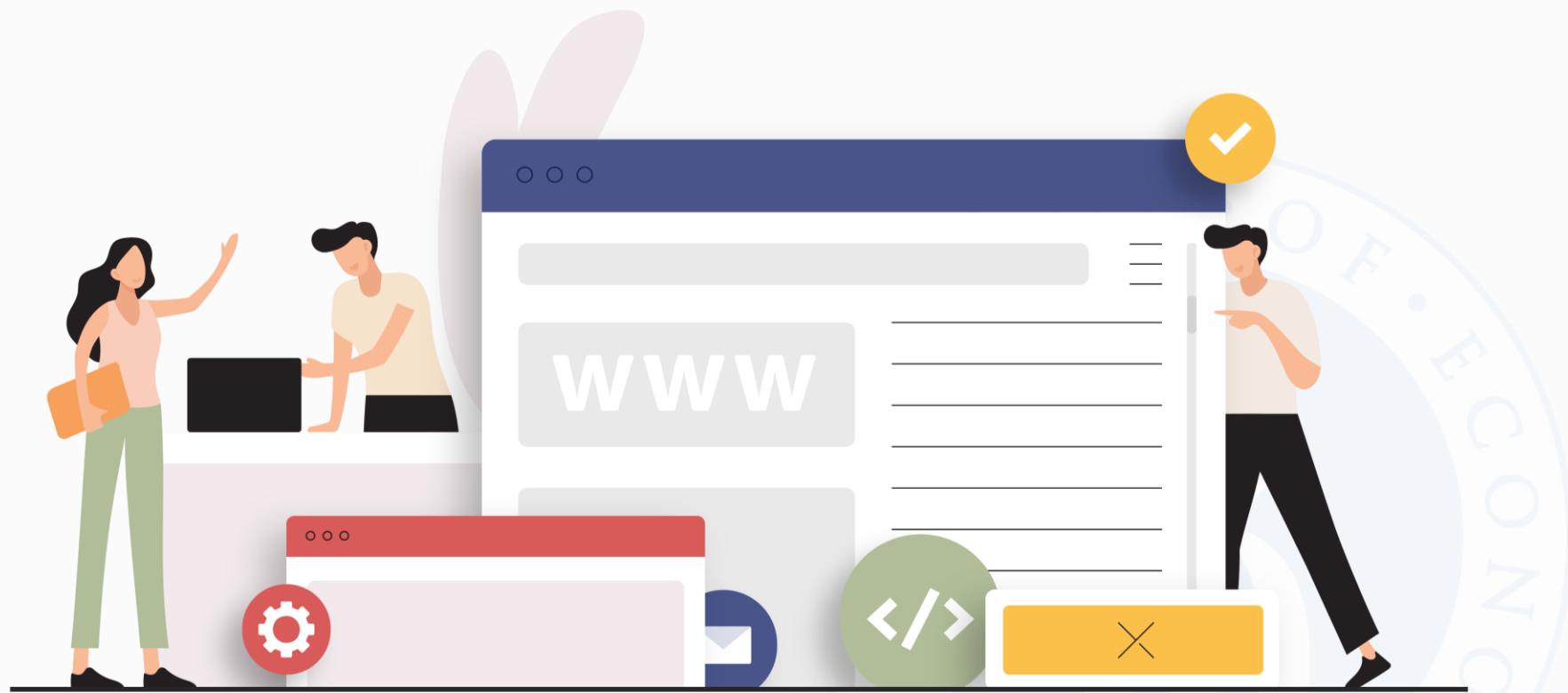
→ Большие данные — это новая нефть в XXI веке.
Но как и сырая нефть не пригодна для использования,
так и сырье данные не приносят большой пользы



Зачем нужны данные



На этой неделе мы расскажем, как подготовить данные, чтобы они приносили нам пользу и ценность



План



Откуда берутся данные



План



Откуда берутся данные



В каких форматах хранятся данные

План



Откуда берутся данные



В каких форматах хранятся данные



Как «почистить» данные



План



Откуда берутся данные



В каких форматах хранятся данные



Как «почистить» данные



Как вручную разметить данные



План



Откуда берутся данные



В каких форматах хранятся данные



Как «почистить» данные



Как вручную разметить данные

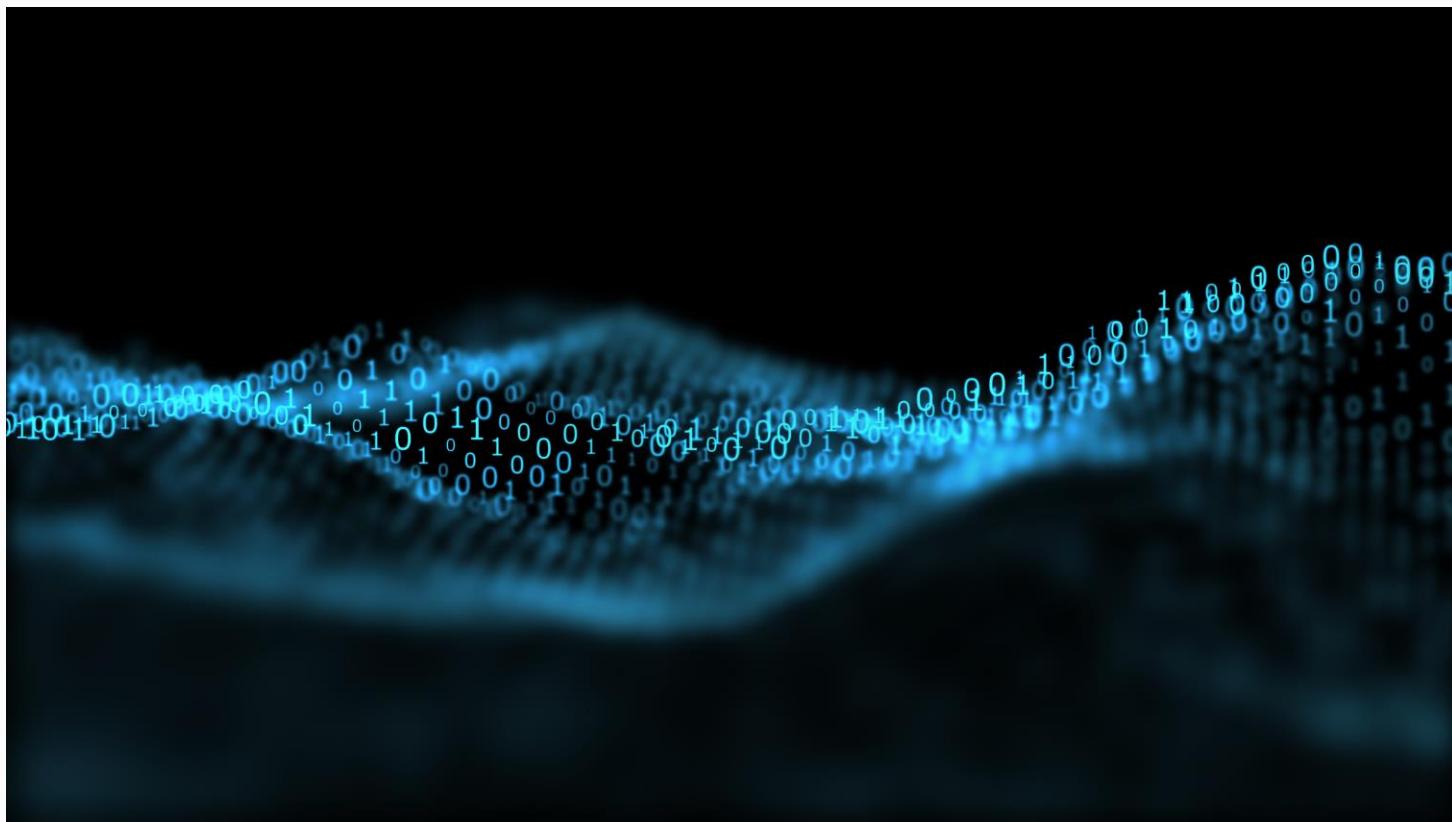


Что такое краудсорсинг и как он нам может помочь

Типы данных



Набор данных (**dataset**) — это обработанные и структурированные в определенном формате данные, на которых обучаются алгоритмы машинного обучения



Типы данных



Набор данных (dataset) — это обработанные и структурированные в определенном формате данные, на которых обучаются алгоритмы машинного обучения

Name	Sex	Age	Ticket	Cabin	Survived
Braund, Mr. Owen Harris	male	22.00	A/5 21171	NaN	0
John Bradley (Florence Briggs Th... Heikkinen, Miss. Laina	female	38.00 26.00	PC 17599 STON/O2. 3101282	C85 NaN	1 1
Mrs. Jacques Heath (Lily May Peel)	female	35.00	113803	C123	1
Allen, Mr. William Henry	male	35.00	373450	NaN	0

Типы данных



Набор данных (dataset) — это обработанные и структурированные в определенном формате данные, на которых обучаются алгоритмы машинного обучения

Целевая
переменная

Name	Sex	Age	Ticket	Cabin	Survived
Braund, Mr. Owen Harris	male	22.00	A/5 21171	NaN	0
John Bradley (Florence Briggs Th... Heikkinen, Miss. Laina	female	38.00	PC 17599	C85	1
Mrs. Jacques Heath (Lily May Peel)	female	26.00	STON/O2. 3101282	NaN	1
Allen, Mr. William Henry	male	35.00	113803	C123	1
			373450	NaN	0

Типы данных



Набор данных (dataset) — это обработанные и структурированные в определенном формате данные, на которых обучаются алгоритмы машинного обучения

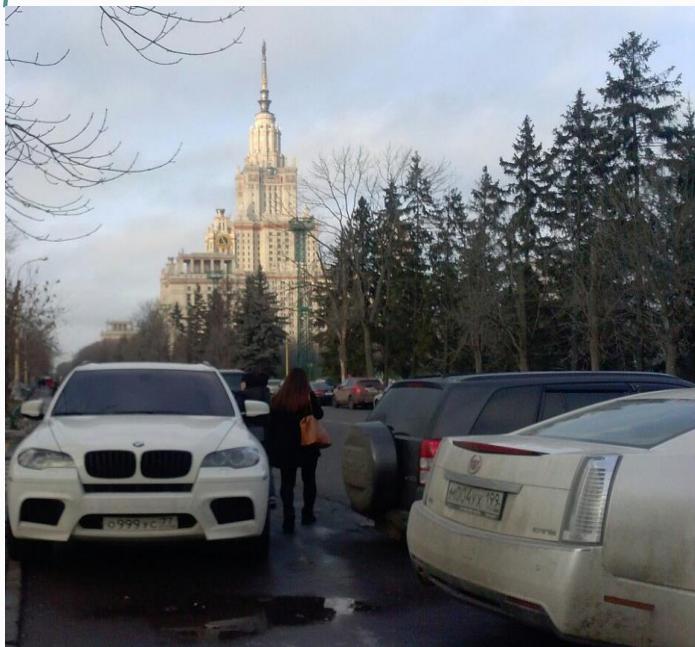
Признаки					Целевая переменная
Name	Sex	Age	Ticket	Cabin	Survived
Braund, Mr. Owen Harris	male	22.00	A/5 21171	NaN	0
John Bradley (Florence Briggs Th... Heikkinen, Miss. Laina	female	38.00	PC 17599 STON/O2. 3101282	C85 NaN	1 1
Mrs. Jacques Heath (Lily May Peel)	female	26.00	113803	C123	1
Allen, Mr. William Henry	male	35.00	373450	NaN	0

Типы данных



Набор данных (**dataset**) — это обработанные и структурированные в определенном формате данные, на которых обучаются алгоритмы машинного обучения

Признаки



Целевая
переменная



Способы получения разметки

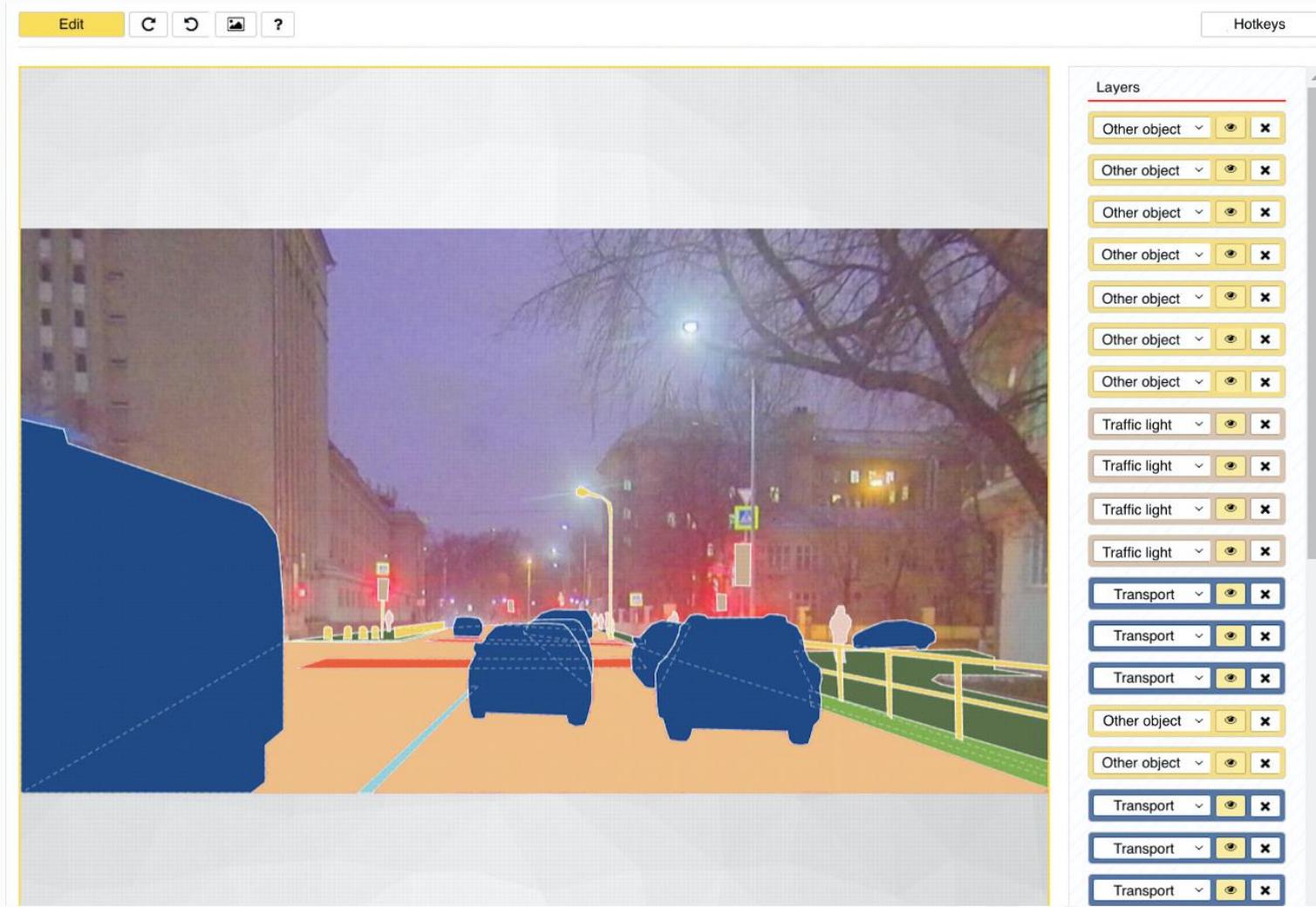


Разметка из логов

Years in current job	Credit Score	Annual Income	Maximum Open Credit	Loan Status
8 years	709.0	1167493.0	416746.0	Fully Paid
10+ years	NaN	NaN	850784.0	Fully Paid
8 years	741.0	2231892.0	750090.0	Fully Paid
3 years	721.0	806949.0	386958.0	Fully Paid
5 years	NaN	NaN	427174.0	Fully Paid
10+ years	7290.0	896857.0	272448.0	Charged Off

Способы получения разметки

→ Ручная разметка



Резюме



Большие данные играют огромную роль в XXI веке



Резюме



Большие данные играют огромную роль в ХХI веке



Чтобы использовать сырье данные с пользой,
их нужно предварительно обработать



Резюме



Большие данные играют огромную роль в ХХI веке



Чтобы использовать сырье данные с пользой,
их нужно предварительно обработать



Разметка данных бывает ручной и автоматической
из логов

Резюме



Большие данные играют огромную роль в ХХI веке



Чтобы использовать сырье данные с пользой,
их нужно предварительно обработать



Разметка данных бывает ручной и автоматической
из логов



Далее: форматы и способы хранения табличных данных

Форматы хранения табличных данных



План



В каких форматах хранятся табличные данные

План



В каких форматах хранятся табличные данные



Какие форматы наиболее популярны

План



В каких форматах хранятся табличные данные



Какие форматы наиболее популярны



Какие проблемы встречаются в табличных данных

Табличные данные



CSV — формат представления табличных данных

- Стока текста соответствует строке таблицы
- Поля таблицы разделяются запятыми

csv
Comma-Separated Values

fname, lname
nancy, davo
erin , bora
tony , rapha

:



names.csv

Расширение .csv

MIME-тип text/csv

Тип формата представление базы данных

Стандарт(ы) RFC 4180 ↗

Табличные данные



TSV — формат представления табличных данных

- Стока текста соответствует строке таблицы
- Поля таблицы разделяются символами табуляции

TSV
Tab-Separated Values

fname Inam
nancy davo
erin bora
tony rapha

:



names.TSV

Расширение .tsv

MIME-тип text/tsv

Тип формата представление базы данных

Стандарт(ы) RFC 4180

Табличные данные



XLSX — стандартный формат файлов Microsoft Excel, основанный на языке XML

→ Широко распространен

→ Используется для хранения небольших таблиц



Табличные данные



JSON — текстовый формат представления данных в нотации объекта JavaScript

- Легко читается человеком
- Понятен для машины

```
{  
  "firstName": "Иван",  
  "lastName": "Иванов",  
  "address": {  
    "streetAddress":  
      "Советская 1",  
    "city": "Ленинград",  
    "postalCode": 101101  
  },  
  "phoneNumbers": [  
    "888 123-1234",  
    "916 123-4567"  
  ]  
}
```

Табличные данные



Базы данных

→ MySQL

→ PostgreSQL

→ MongoDB

→ Redis

Используются
для хранения больших
объемов данных



Проблемы в табличных данных

→ Отсутствие структурированности в информации

images	
123	image_id
123	user_id
ABC	url_vk
123	status
✓	no_person
ABC	impression
ABC	url_storage
ABC	id_toloka_task
ABC	id_toloka_no_person
ABC	validation_info
ABC	toloker_user_id

users	
123	user_id
ABC	user_name
✓	personal_data
123	reviews_available
123	count
ABC	user_surname
123	sex
123	age
ABC	birth_date
ABC	time

payments	
123	id
123	user_id
ABC	snippet_type
123	amount
123	app_id
ABC	attach1
ABC	text

allusers	
123	user_id
123	access
ABC	time

Проблемы в табличных данных

→ Разные единицы записи

Например:

-  Мили, километры, метры, сантиметры, футы
-  Минуты, часы, дни, года
-  Килограммы, унции, фунты, тонны

Проблемы в табличных данных

→ Различные кодировки

→ ASCII, Windows-1251, Unicode

Д'Ñ« Ð¼Ð¾Ð¶ÐµÑ, Ðµ Ð½Ð°Ñ□Ñ, Ñ€Ð¾Ð_Ñ, ÑŒ Ñ□Ð²Ð¾ÑŽ ÐºÐ»ÐºÐ²Ð_Ð°Ñ, ÑfÑ€Ñf
Ð’Ð»Ñ□ Ñ€ÑfÑ□Ñ□Ð°Ð¾Ð¹ Ñ€Ð°Ñ□Ð°Ð»ÐºÐ_Ð°Ð_ Windows XP - Ñ□Ñ, Ð°Ð½Ð’Ð°
Ñ€Ñ, Ð½Ð°Ñ□ Ñ€ÑfÑ□Ñ□Ð°Ð°Ñ□ ÐºÐ»ÐºÐ²Ð_Ð°Ñ, ÑfÑ€Ð°

Ð’Ð°Ð¶Ð½Ð¾! Ð§Ñ, Ð¾Ð±Ñ< Ð’Ð¾Ð±Ð°Ð²Ð_Ñ, ÑŒ Ñ€Ð°Ñ□Ð°Ð»ÐºÐ_Ð°Ñf ÐºÐ»Ð°
Ð²Ð_Ð°Ñ, ÑfÑ€Ñ< Ð¿Ð¾Ð_ Windows XP/2003, Ð¿Ð¾Ð»ÑŒÐ·Ð¾Ð²Ð°Ñ, ÐµÐ»ÑŒ
Ð’Ð¾Ð»Ð¶ÐµÐ½Ð²Ð¾Ð¹Ñ, Ð_Ð²Ñ□Ð_Ñ□Ñ, ÐµÐ¼Ñf ÐºÐ°Ð°
Ð□Ð_Ð¼Ð_Ð½Ð_Ñ□Ñ, Ñ€Ð°Ñ, Ð¾Ñ€, Ð_Ð½Ð°Ñ‡ÐµÐ²Ñ□Ð»ÐµÐ_ÑfÑŽÑ‰ÐµÐ¹
Ñ□ÐµÑ□Ñ□Ð_Ð_Ñ□Ñ, Ð_Ñ... ÑfÑ□Ñ, Ð°Ð½Ð¾Ð²Ð¾Ð° Ð½Ðµ Ð±ÑfÐ_ÐµÑ, - Ð°Ð°Ð°
Ð_Ð½Ðµ ÑfÑ□Ñ, Ð°Ð½Ð°Ð²Ð»Ð_Ð²Ð°Ð»Ð_!

ÐšÐ°Ð° Ð°Ð°Ñ, Ð_Ð²Ð_Ñ€Ð¾Ð²Ð°Ñ, ÑŒ Ñ□Ñ, Ð°Ð½Ð_Ð°Ñ€Ñ, Ð½ÑfÑŽ Ñ€Ð°
Ñ□Ð°Ð»Ð°Ð_Ð°Ñf Ñ€ÑfÑ□Ñ□Ð°Ð¾Ð¹ ÐºÐ»ÐºÐ²Ð_Ð°Ñ, ÑfÑ€Ñ< Ð¿Ð¾Ð_ XP / 2003:

Проблемы в табличных данных



Пропуски в данных

Years in current job	Credit Score	Annual Income	Maximum Open Credit	Loan Status
8 years	709.0	1167493.0	416746.0	Fully Paid
10+ years	NaN	NaN	850784.0	Fully Paid
8 years	741.0	2231892.0	750090.0	Fully Paid
3 years	721.0	806949.0	386958.0	Fully Paid
5 years	NaN	NaN	427174.0	Fully Paid
10+ years	7290.0	896857.0	272448.0	Charged Off

Проблемы в табличных данных



Изменчивость данных со временем

Например:

- Добавление новых полей в базу данных
- Изменение формата записи информации
- Изменение единиц записи

Резюме



Существует множество форматов хранения табличных данных



Резюме



Существует множество форматов хранения табличных данных



Форматы CSV, TSV, XLSX, JSON удобны для хранения небольших таблиц

Резюме



Существует множество форматов хранения табличных данных



Форматы CSV, TSV, XLSX, JSON удобны для хранения небольших таблиц



Большие таблицы хранятся в базах данных: MySQL, PostgreSQL, Mongo DB, и др.

Резюме

В табличных данных могут возникать проблемы:



Резюме

В табличных данных могут возникать проблемы:



Отсутствие структурированности информации

Резюме

В табличных данных могут возникать проблемы:



Отсутствие структурированности информации



Разные единицы записи

Резюме

В табличных данных могут возникать проблемы:



Отсутствие структурированности информации



Разные единицы записи



Различные кодировки

Резюме

В табличных данных могут возникать проблемы:



Отсутствие структурированности информации



Разные единицы записи



Различные кодировки



Пропуски в данных



Резюме

В табличных данных могут возникать проблемы:

-  Отсутствие структурированности информации
-  Разные единицы записи
-  Различные кодировки
-  Пропуски в данных
-  Изменчивость данных со временем

Резюме

В табличных данных могут возникать проблемы:

-  Отсутствие структурированности информации
-  Разные единицы записи
-  Различные кодировки
-  Пропуски в данных
-  Изменчивость данных со временем
-  Далее: рассмотрим реальные табличные данные и произведем их предобработку

Форматы хранения разметки для задач компьютерного зрения

План



Какие задачи компьютерного зрения существуют

План



Какие задачи компьютерного зрения существуют



В каких форматах хранятся данные

План



Какие задачи компьютерного зрения существуют



В каких форматах хранятся данные



Способы хранения разметки для задачи сегментации

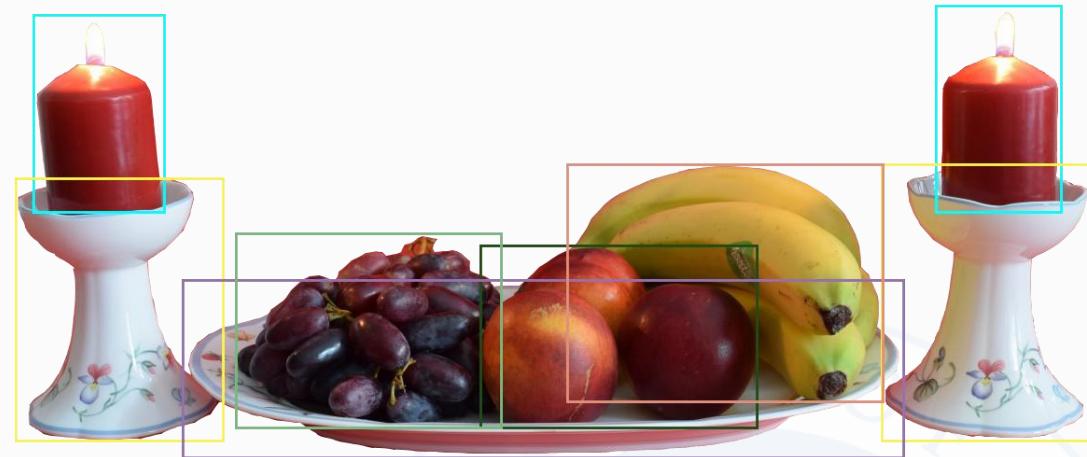
Виды задач

→ Обнаружение объектов

→ Сегментация

→ Обнаружение ключевых точек

→ Классификация, теггинг



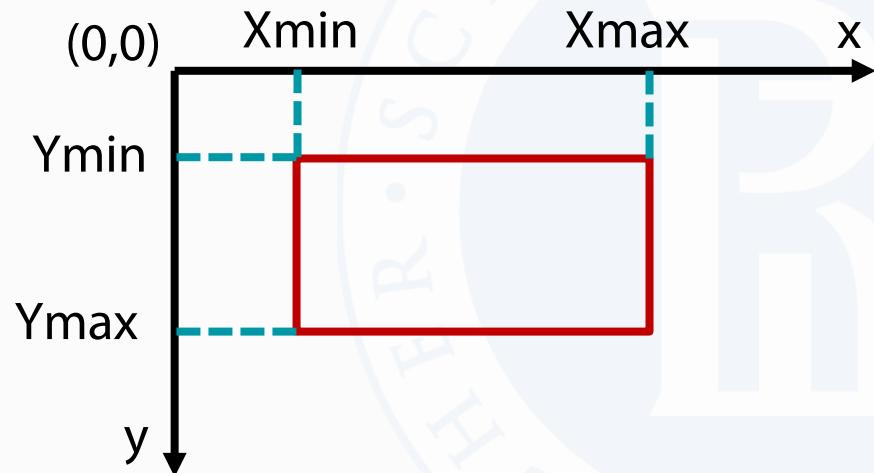
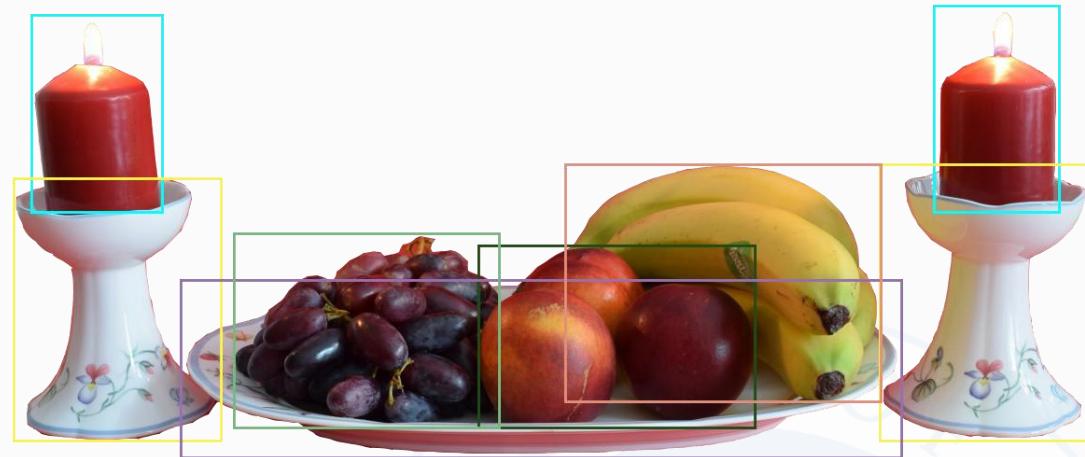
Виды задач

→ Обнаружение объектов

→ Сегментация

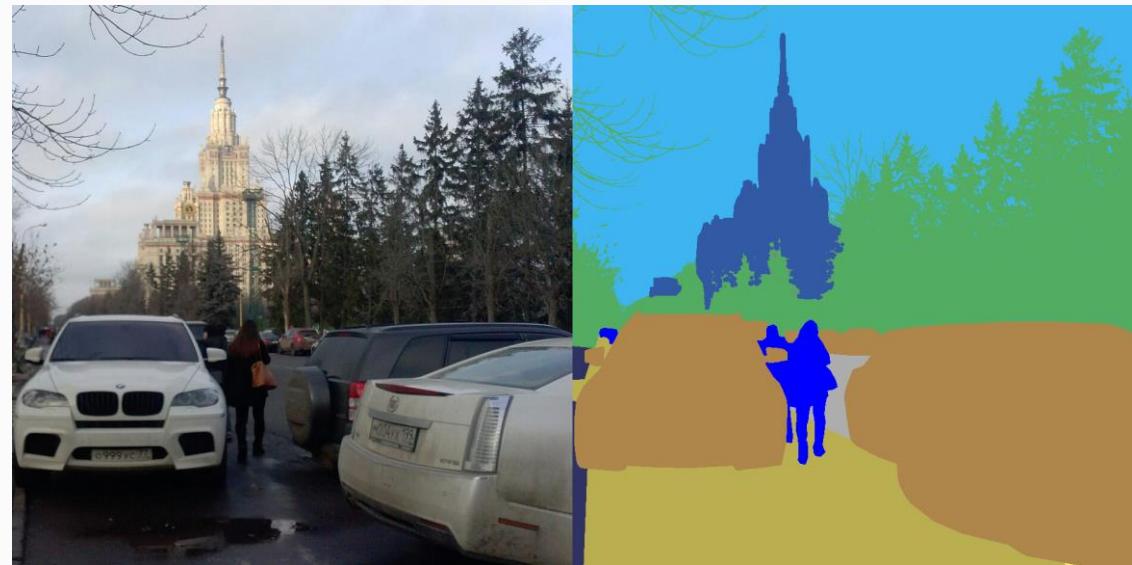
→ Обнаружение ключевых точек

→ Классификация, теггинг



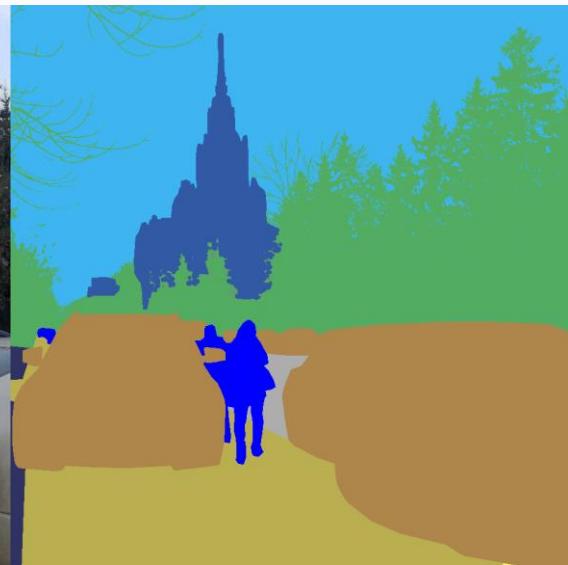
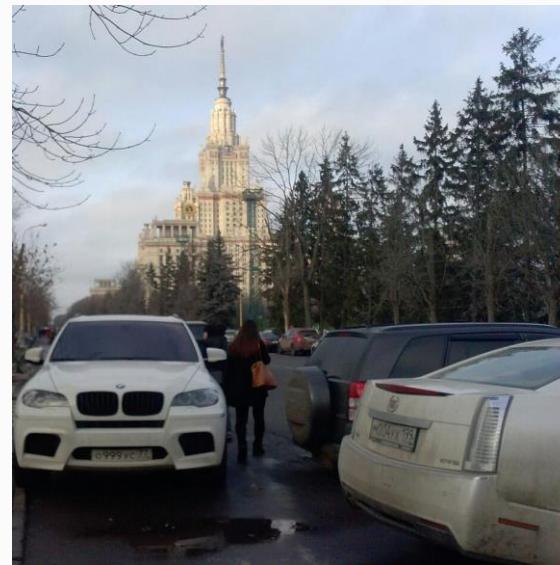
Виды задач

- Обнаружение объектов
- Сегментация
- Обнаружение ключевых точек
- Классификация, теггинг



Виды задач

→ Обнаружение объектов



→ Сегментация

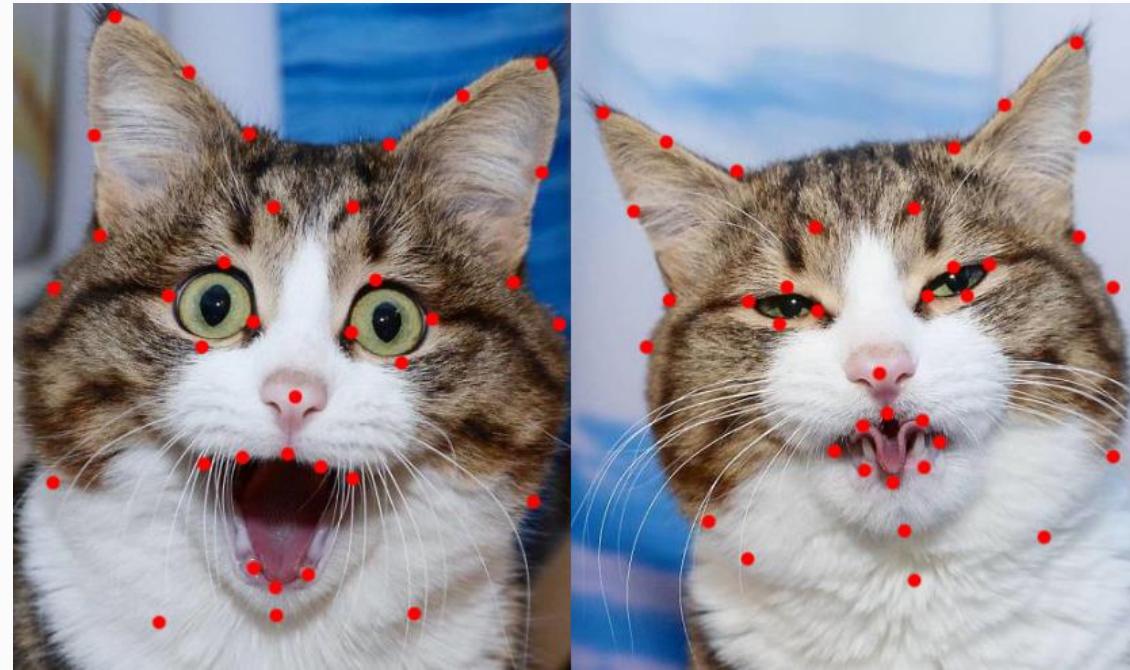
→ Обнаружение ключевых точек

→ Классификация, теггинг

```
dataset
| images
| 00000000139.jpg
| 00000000285.jpg
| ...
| masks
| 00000000139.png
| 00000000285.png
| ...
```

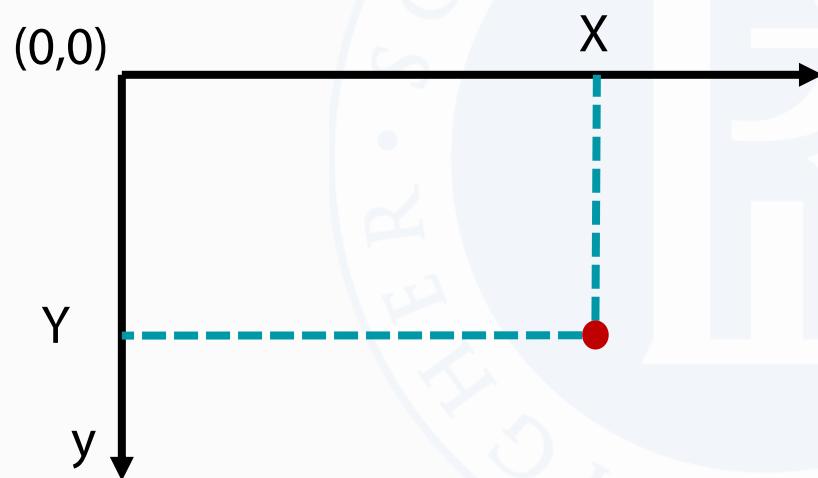
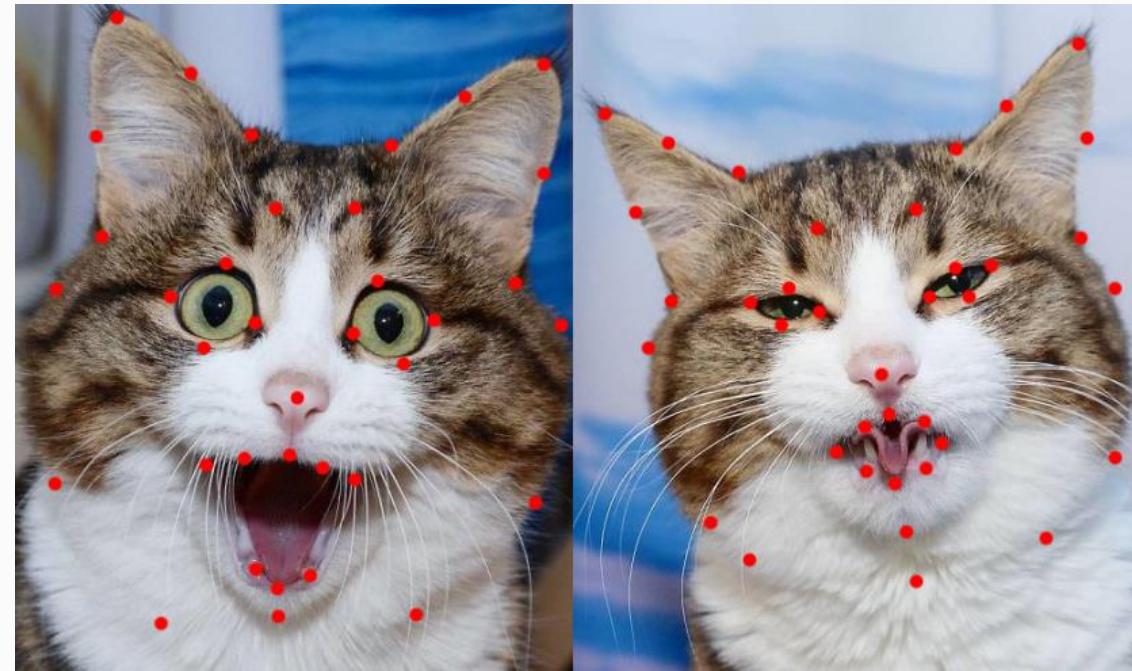
Виды задач

- Обнаружение объектов
- Сегментация
- Обнаружение ключевых точек
- Классификация, теггинг



Виды задач

- Обнаружение объектов
- Сегментация
- Обнаружение ключевых точек
- Классификация, теггинг



Виды задач

- Обнаружение объектов
- Сегментация
- Обнаружение ключевых точек
- Классификация, теггинг



На изображении присутствуют животные?

- Да Нет

Какие животные присутствуют на изображении?

- Кот
 Собака
 Носорог
 Жираф
 Медведь

Форматы хранения разметки

ImageNet — огромный набор аннотированных изображений, предназначенных для тестирования методов распознавания образов и компьютерного зрения

- Более 14 миллионов изображений
- 22 000 классов
- Изображения разбиты по папкам
- В каждой папке находятся изображения, принадлежащие одному классу

The logo for ImageNet features the word "IMAGENET" in a bold, sans-serif font. The letter "I" is large and black, while the other letters are smaller and gray. To the left of the "I", there are three small colored squares: green, orange, and red.

Форматы хранения разметки

Pascal VOC — огромный набор аннотированных данных для задачи обнаружения объектов на изображении

- Более 10 000 изображений
- 20 классов
- 27 450 аннотированных объектов
- XML файл для каждого изображения



Форматы хранения разметки

```
<annotation>
    <folder>Kangaroo</folder>
    <filename>00001.jpg</filename>
    <path>./Kangaroo/stock-12.jpg</path>
<size>
    <width>450</width>
    <height>319</height>
    <depth>3</depth>
    </size>
<object>...</object>
</annotation>
```

Форматы хранения разметки

```
<annotation>
    <folder>Kangaroo</folder>
    <filename>00001.jpg</filename>
    <path>./Kangaroo/stock-12.jpg</path>
<size>
    <width>450</width>
    <height>319</height>
    <depth>3</depth>
    </size>
<object>...</object>
</annotation>
```

Форматы хранения разметки

```
<annotation>
    <folder>Kangaroo</folder>
    <filename>00001.jpg</filename>
    <path>./Kangaroo/stock-12.jpg</path>
<size>
    <width>450</width>
    <height>319</height>
    <depth>3</depth>
</size>
<object>...</object>
</annotation>
```

Форматы хранения разметки

```
<annotation>
    <folder>Kangaroo</folder>
    <filename>00001.jpg</filename>
    <path>./Kangaroo/stock-12.jpg</path>
<size>...</size>
<object>
    <bndbox>
        <xmin>233</xmin>
        <ymin>89</ymin>
        <xmax>386</xmax>
        <ymax>262</ymax>
    </bndbox>
</object>
</annotation>
```

Форматы хранения разметки

COCO — набор изображений для задач обнаружения объектов и ключевых точек, сегментации, описания объекта на изображении

- Более 330 тысяч изображений
- 1.5 миллиона выделенных объектов
- Один JSON для всего набора данных



Форматы хранения разметки

```
{"annotations": [  
    {  
        "segmentation": [ [ 510.66,  
                            423.01,  
                            511.72,  
                            420.03,  
                            ...,  
                            510.45,  
                            423.01 ] ],  
        "area": 702.10,  
        "bbox": [ 433.07, 355.93, 138.6, 228.6 ],  
        "image_id": 397133,  
        "category_id": 18,  
        "id": 1768  
    } ] }
```

Форматы хранения разметки

```
{"annotations": [  
    {  
        "segmentation": [ [ 510.66,  
                            423.01,  
                            511.72,  
                            420.03,  
                            ...,  
                            510.45,  
                            423.01 ] ],  
        "area": 702.10,  
        "bbox": [433.07,355.93,138.6,228.6],  
        "image_id": 397133,  
        "category_id": 18,  
        "id": 1768  
    } ] }
```

Форматы хранения разметки

```
{"annotations": [  
    {  
        "segmentation": [ [ 510.66,  
                            423.01,  
                            511.72,  
                            420.03,  
                            ...,  
                            510.45,  
                            423.01 ] ],  
        "area": 702.10,  
        "bbox": [ 433.07, 355.93, 138.6, 228.6 ],  
        "image_id": 397133,  
        "category_id": 18,  
        "id": 1768  
    } ] }
```

Способы хранения разметки сегментации

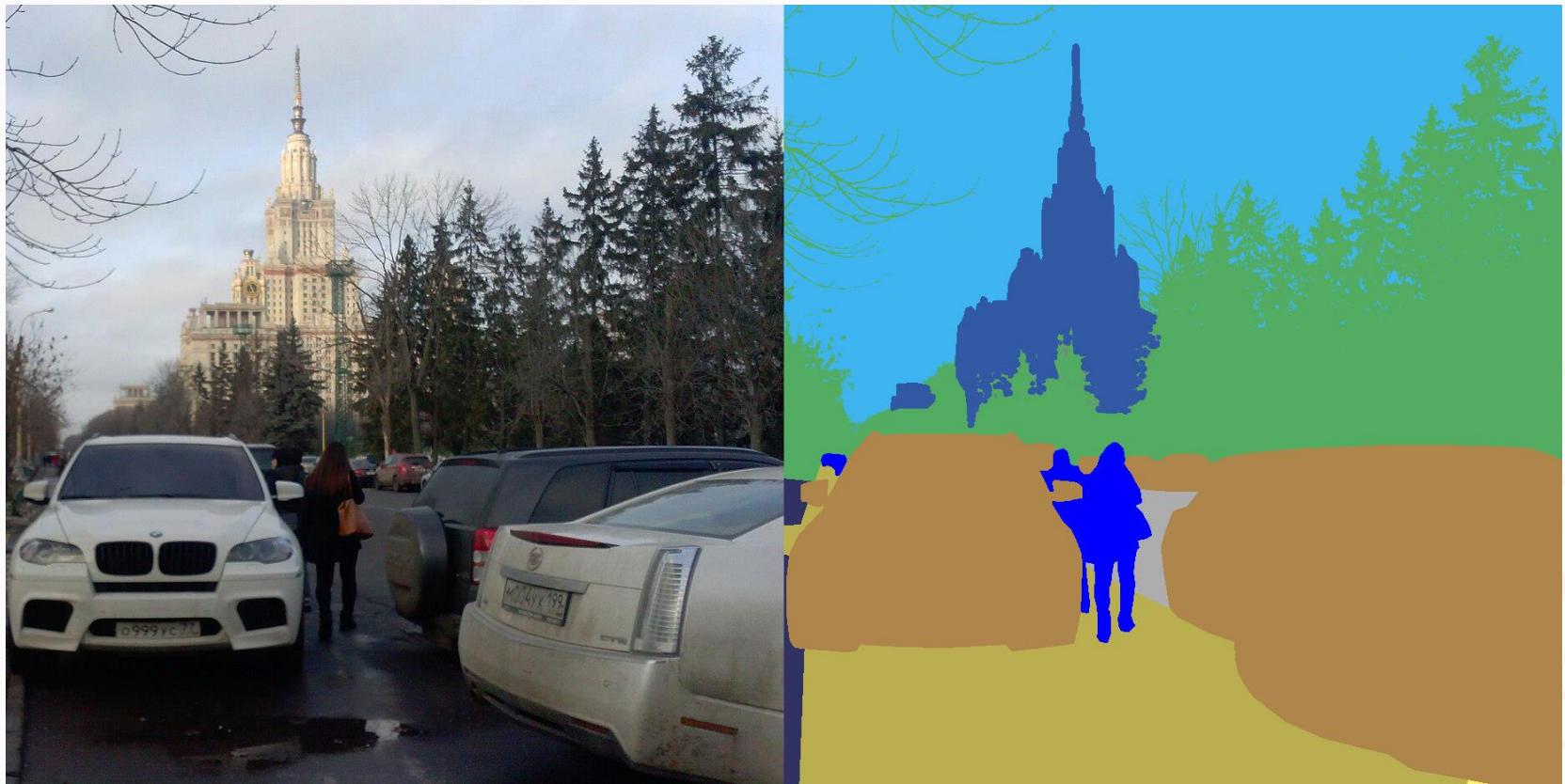
→ Массив полигонов

```
"segmentation": [  
    [  
        331.8, //x1  
        402.2, //y1 }  
        337.9, //x2 }  
        403.7, //y2 }  
        362.4,  
        .....  
        362.4,  
        422.0,  
    ]  
]
```



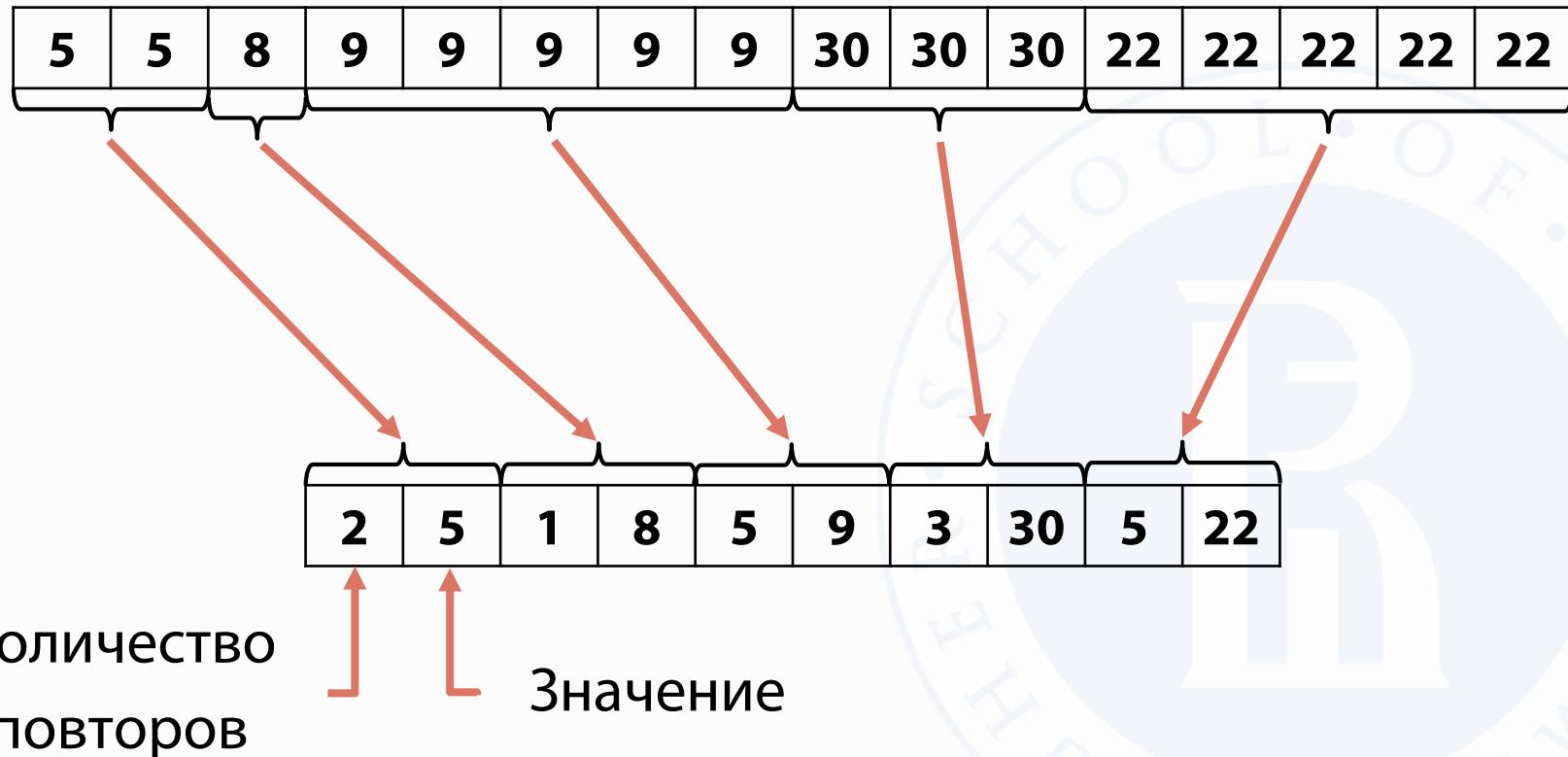
Способы хранения разметки сегментации

→ Маски



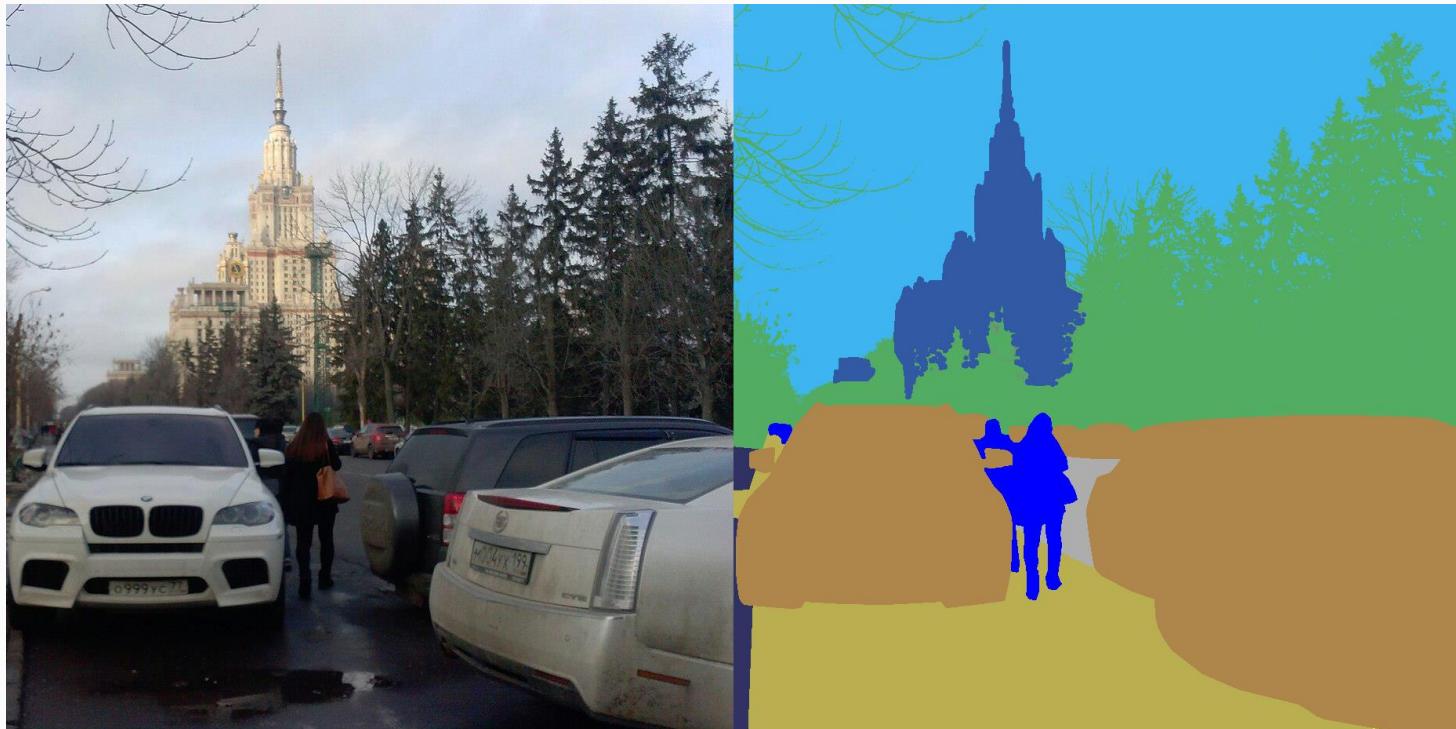
Способы хранения разметки сегментации

→ Кодирование длин серий (run-length encoding) — алгоритм сжатия данных, заменяющий повторяющиеся символы (серии) на один символ и число его повторов



Способы хранения разметки сегментации

→ Кодирование длин серий (run-length encoding) — алгоритм сжатия данных, заменяющий повторяющиеся символы (серии) на один символ и число его повторов



Способы хранения разметки сегментации

→ Кодирование длин серий (run-length encoding) — алгоритм сжатия данных, заменяющий повторяющиеся символы (серии) на один символ и число его повторов

Матрица

1000

1100

1110

1111



Сжатая запись

11031202130114

Резюме



Для каждой задачи используется свой формат хранения данных

Резюме



Для каждой задачи используется свой формат хранения данных



Нет универсального и стандартизированного формата хранения разметки

Резюме



Для каждой задачи используется свой формат хранения данных



Нет универсального и стандартизированного формата хранения разметки



Разметку сегментации можно хранить несколькими способами: масками, полигонами, сжатыми масками при помощи кодирования длин серий

Резюме



Для каждой задачи используется свой формат хранения данных



Нет универсального и стандартизированного формата хранения разметки



Разметку сегментации можно хранить несколькими способами: масками, полигонами, сжатыми масками при помощи кодирования длин серий



Далее: рассмотрим набор данных COCO более детально

Инструменты для разметки данных



План



Рассмотрим самые популярные программы
для разметки

План



Рассмотрим самые популярные программы
для разметки



Определим, какие программы для каких задач
подходят

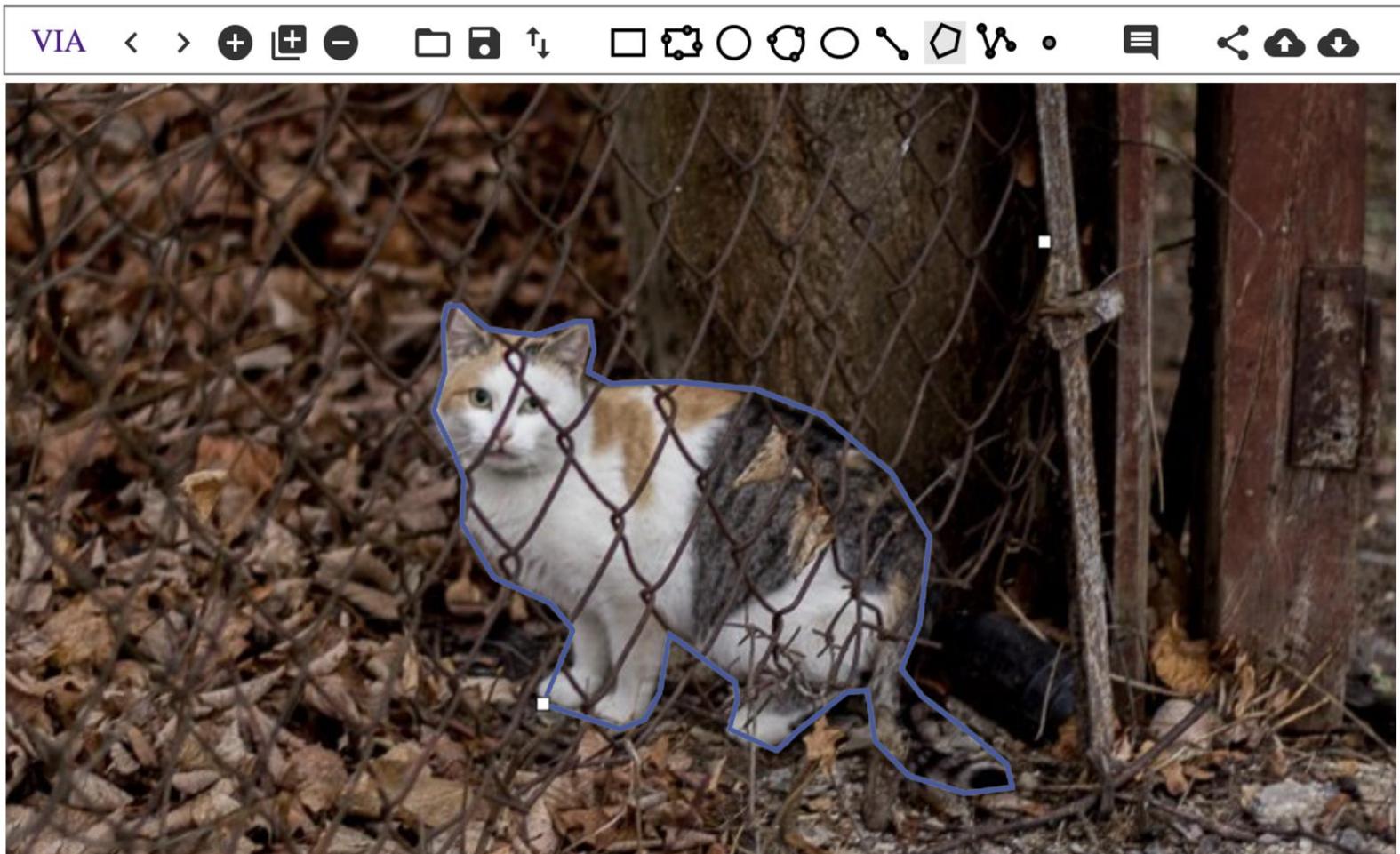
Ipyannotate

```
from glob import glob
from PIL import Image

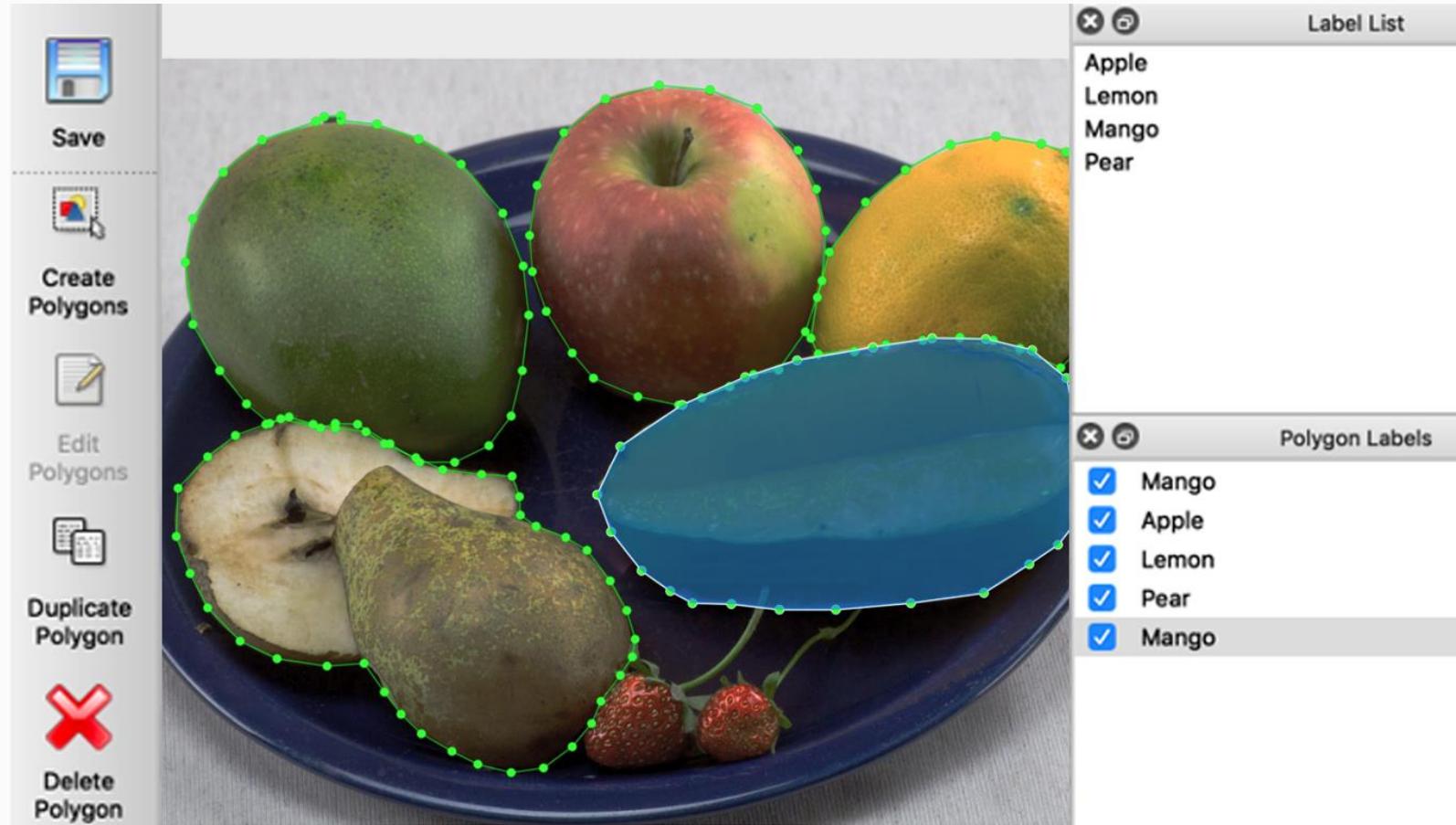
data = [Image.open(_) for _ in glob('i/dogs_cats/*.jpg')]
annotation = annotate(data, buttons=[Button('dog'), Button('cat'), Back(), Next()])
annotation
```



VGG Image Annotator (VIA)



Labelme



Supervise.ly

The screenshot displays the Supervise.ly web application interface. On the left, a vertical toolbar contains icons for navigation, zooming, and labeling. The main area shows a photograph of a pink MacBook Pro 2016 on a wooden desk. A yellow desk lamp is positioned behind it, and a small plant sits on a log slice next to the laptop. A red dot is highlighted on the lamp's base, with a callout bubble labeled "lamp_color: black". Another callout bubble labeled "macbook pro 2016" points to the laptop screen. To the right, a sidebar titled "IMAGES 1" shows a thumbnail of the image and its source: "chang-duong-480253-unsplash". Below this, a section titled "FIGURES 3 (68)" displays a classification table:

#	Class	Percentage
1	table-lamp	4%
2	laptop	11%
3	plant	19%

Computer Vision Annotation Tool (CVAT)



Adobe Photoshop



Person p

Loc l

Org o

Event e

Date d

Other z

Barack Hussein Obama II ✖ (born August 4, 1961 ✖) is an American ✖ attorney

and politician who served as the 44th President of the United States ✖ from

January 20, 2009 ✖, to January 20, 2017 ✖. A member of the

Democratic Party ✖, he was the first African American ✖ to serve as president.

He was previously a United States Senator ✖ from Illinois ✖ and a member of the

Illinois State Senate ✖.

Label Studio

Person^[1]

Organization^[2]

Fact^[3]

Money^[4]

Date^[5]

Time^[6]

Ordinal^[7]

Percent^[8]

Product^[9]

Language^[0]

Location^[q]

To have faith Organization is to trust yourself Language to the water

Резюме



Существует множество инструментов для разметки данных

Резюме



Существует множество инструментов для разметки данных



Каждый инструмент имеет свои плюсы и свои минусы

Резюме



Существует множество инструментов для разметки данных

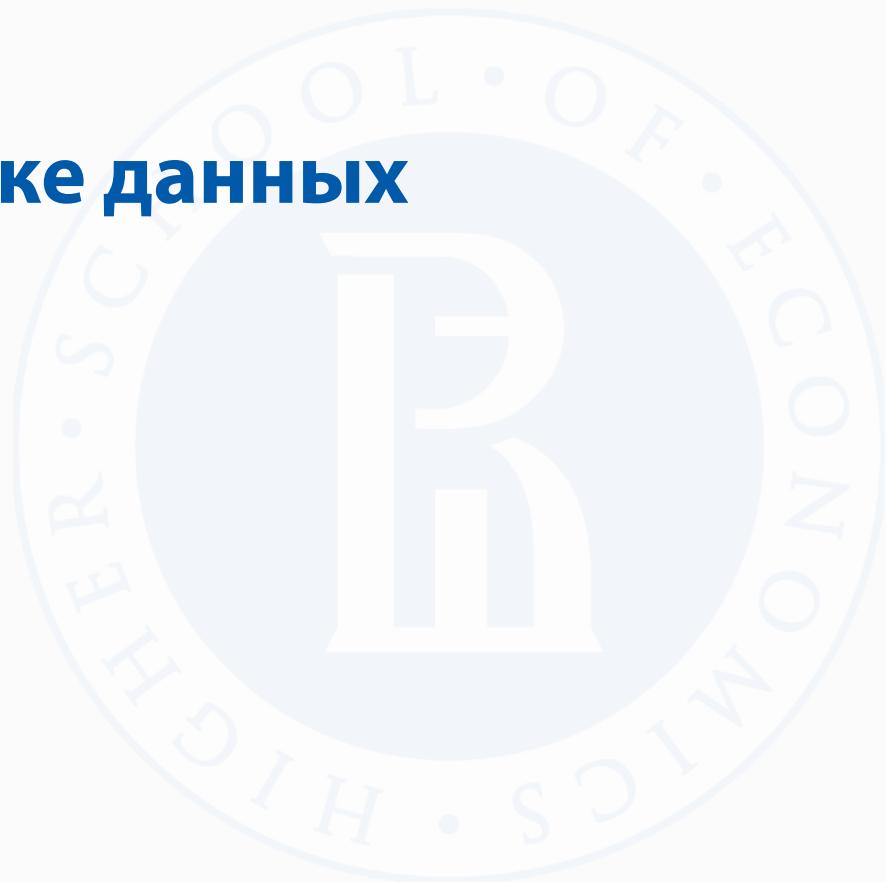


Каждый инструмент имеет свои плюсы и свои минусы



Далее: рассмотрим более детально инструменты
IPyannotate и Label Studio

Тренды в разметке данных



Цели



Увеличение производительности разметчиков

Цели



Увеличение производительности разметчиков



Увеличение качества разметки



Цели



Увеличение производительности разметчиков



Увеличение качества разметки



Уменьшение стоимости разметки

Цели



Увеличение производительности разметчиков



Увеличение качества разметки



Уменьшение стоимости разметки



Уменьшение времени разметки

Цели



Увеличение производительности разметчиков



Увеличение качества разметки



Уменьшение стоимости разметки



Уменьшение времени разметки



Интеграция разметки в ML процесс

Active Learning

Данные перед разметкой прогоняются через алгоритм
машиинного обучения



Active Learning

Данные перед разметкой прогоняются через алгоритм машинного обучения

В первую очередь размечаются данные с низкой уверенностью в ответе. Как правило, они оказываются сложными, и алгоритм с ними плохо справляется

Active Learning

Данные перед разметкой прогоняются **через алгоритм машинного обучения**

В первую очередь размечаются данные **с низкой уверенностью в ответе**. Как правило, они оказываются сложными, и алгоритм с ними **плохо справляется**

→ Алгоритм обучается быстрее на меньшем объеме данных

Active Learning

Данные перед разметкой прогоняются **через алгоритм машинного обучения**

В первую очередь размечаются данные **с низкой уверенностью в ответе**. Как правило, они оказываются сложными, и алгоритм с ними **плохо справляется**

- Алгоритм обучается быстрее на меньшем объеме данных
- Необходимо иметь хоть как-то **работающую модель**

Active Learning

Данные перед разметкой прогоняются **через алгоритм машинного обучения**



В первую очередь размечаются данные с **низкой уверенностью в ответе**. Как правило, они оказываются **сложными**, и алгоритм с ними **плохо справляется**

- Алгоритм обучается быстрее на меньшем объеме данных
- Необходимо иметь хоть как-то **работающую модель**

Pre-labeling

Данные перед разметкой прогоняются через алгоритм машинного обучения

Pre-labeling

Данные перед разметкой прогоняются **через алгоритм машинного обучения**

Разметчики поправляют ошибки алгоритма

Pre-labeling

Данные перед разметкой прогоняются через алгоритм машинного обучения

Разметчики поправляют ошибки алгоритма

→ Ускорение разметки до x10 раз

Pre-labeling

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-labeling

Данные перед разметкой прогоняются **через алгоритм машинного обучения**



Разметчики поправляют ошибки алгоритма

- Ускорение разметки **до x10 раз**
- Необходимо иметь хоть как-то **работающую модель**

Online Learning

- Постоянное переучивание и обновление модели при аннотации новых данных

Online Learning

- Постоянное переучивание и обновление модели при аннотации новых данных

Набор
данных



Online Learning

→ Постоянное переучивание и обновление модели при аннотации новых данных

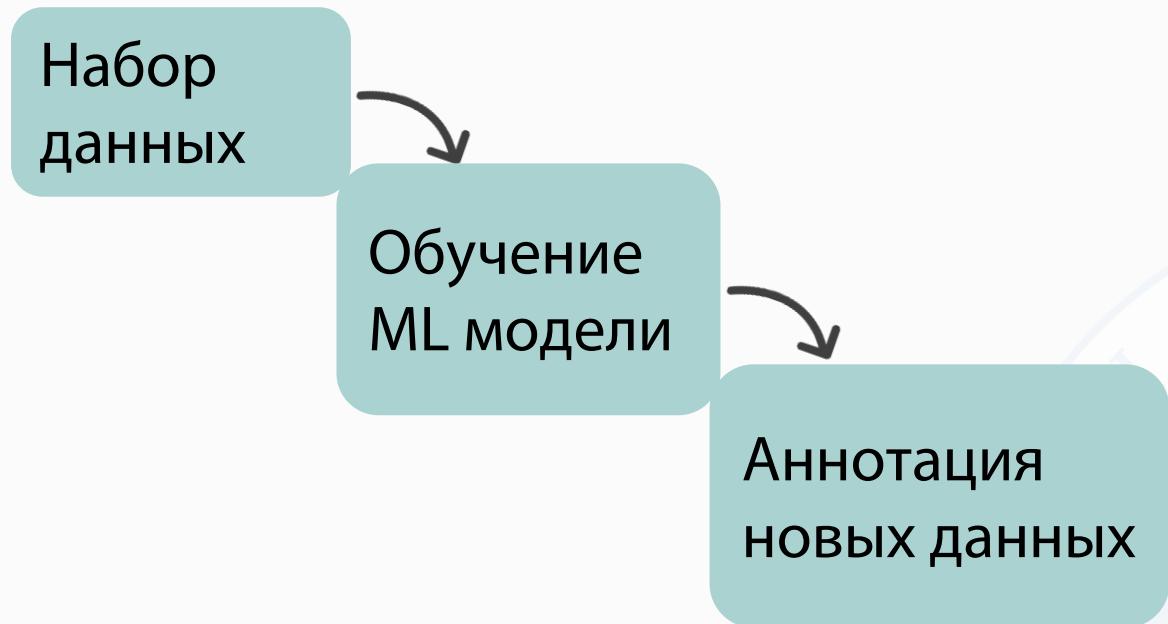
Набор
данных

Обучение
ML модели



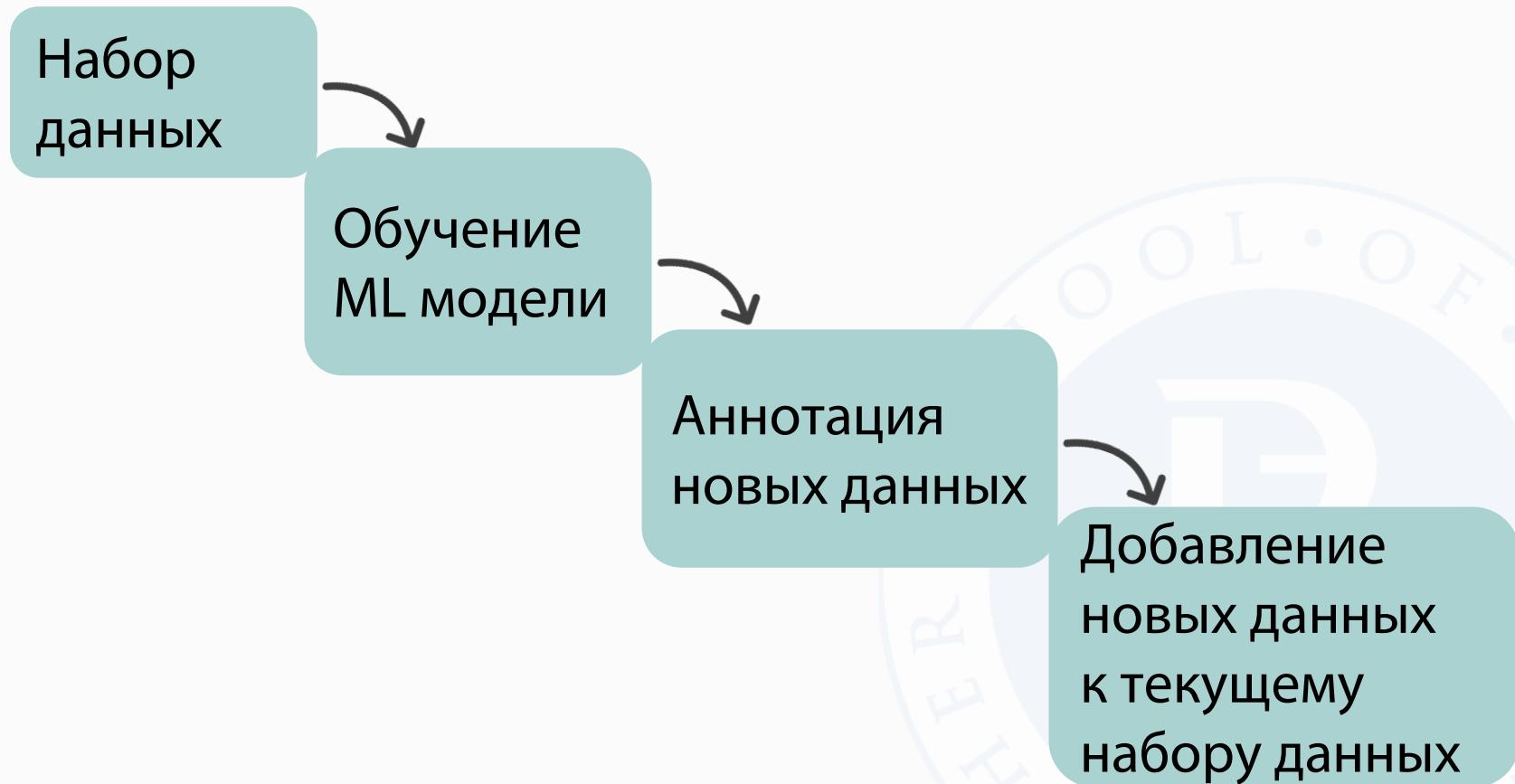
Online Learning

→ Постоянное переучивание и обновление модели при аннотации новых данных



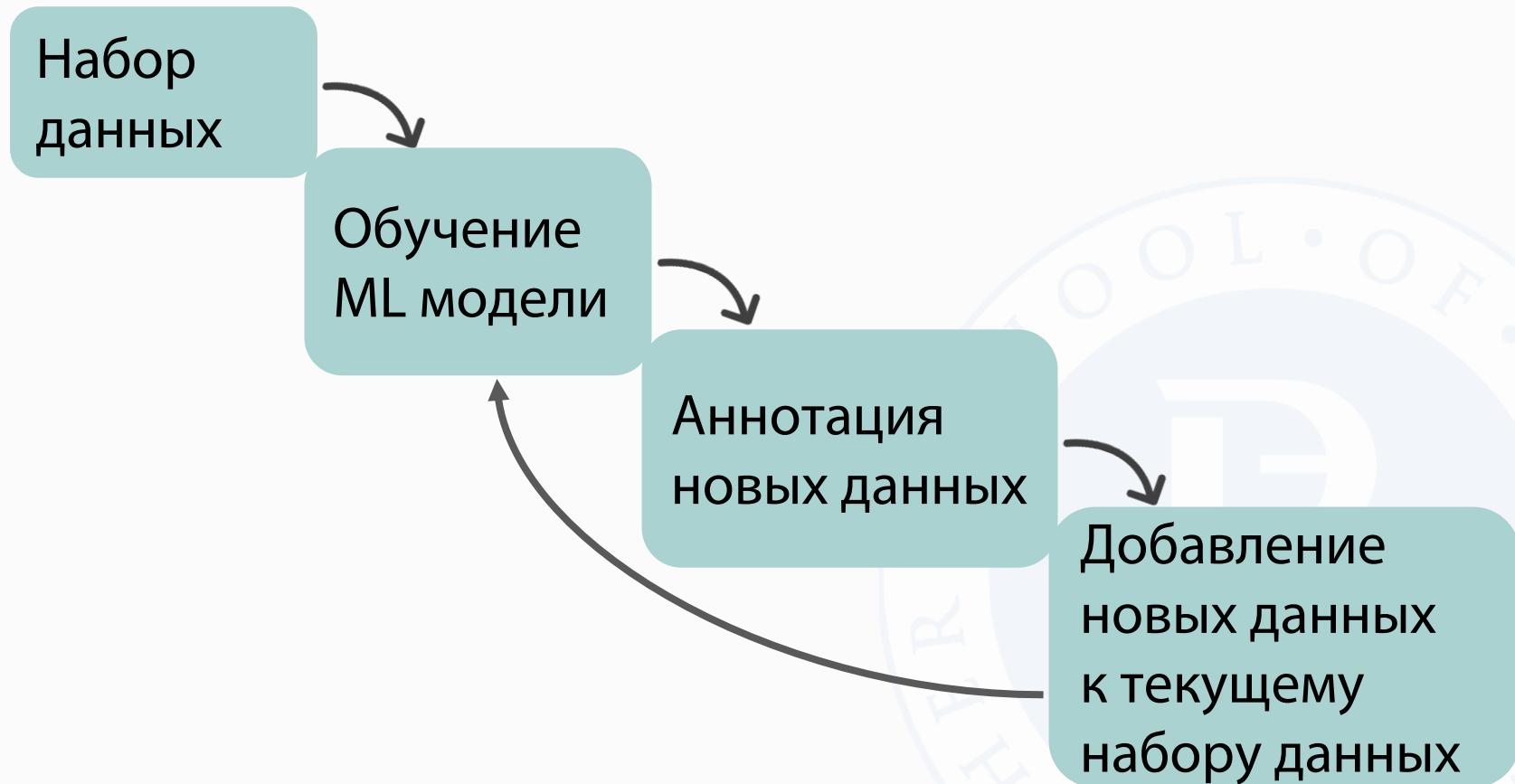
Online Learning

→ Постоянное переучивание и обновление модели при аннотации новых данных



Online Learning

→ Постоянное переучивание и обновление модели при аннотации новых данных



Human in the Loop



Концепция, которая объединяет человеческий
и машинный интеллекты



Human in the Loop

- Концепция, которая объединяет человеческий и машинный интеллекты
- Основную работу выполняет ML модель, люди размечают только самые сложные случаи

Human in the Loop

- Концепция, которая объединяет человеческий и машинный интеллекты
- Основную работу выполняет ML модель, люди размечают только самые сложные случаи
- Такой подход позволяет достичнуть точности 99.9%

Human in the Loop

- Концепция, которая объединяет человеческий и машинный интеллекты
- Основную работу выполняет ML модель, люди размечают только самые сложные случаи
- Такой подход позволяет достичнуть точности 99.9%
- Применяется в задачах, где стоимость ошибки очень высока

Human in the Loop

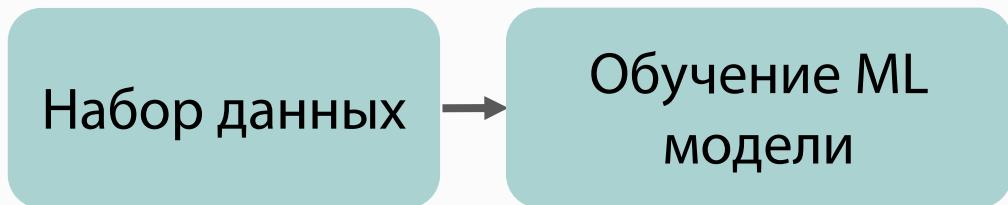


Human in the Loop

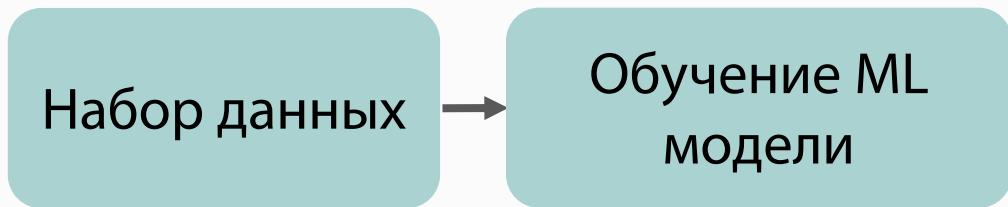
Набор данных



Human in the Loop

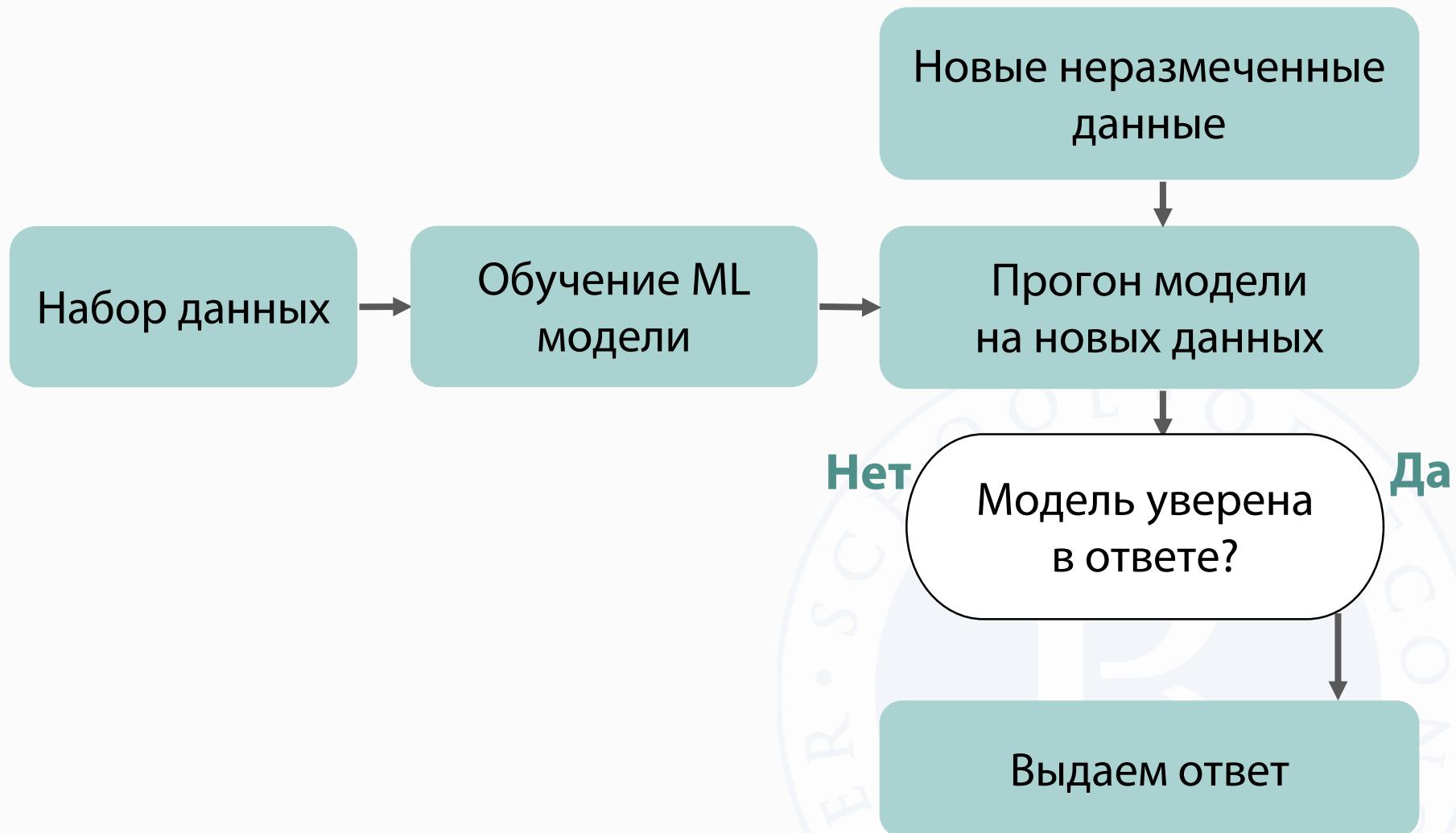


Human in the Loop

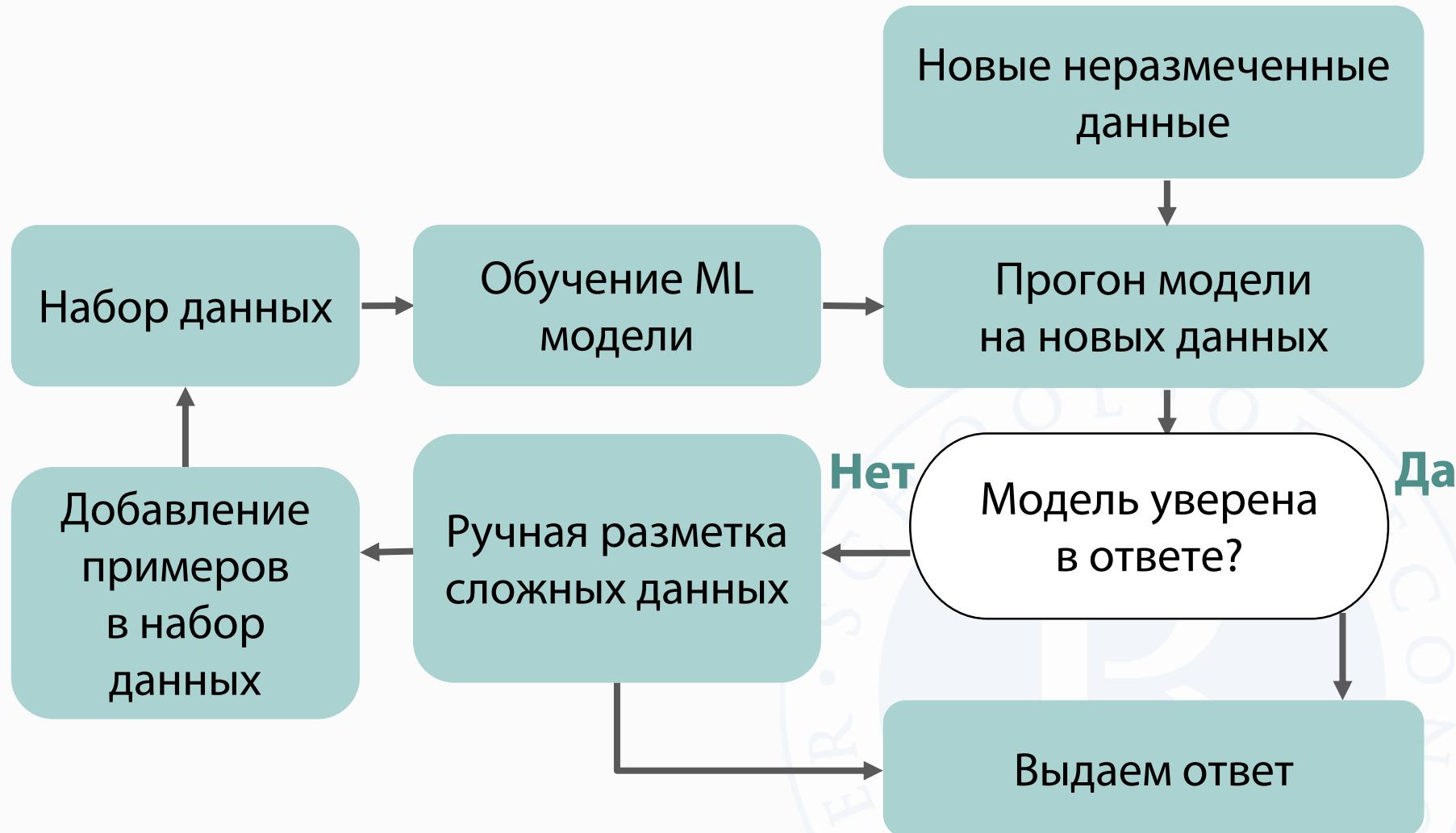


Новые неразмеченные
данные

Human in the Loop



Human in the Loop



Резюме



Существует множество методов и техник, которые позволяют увеличить скорость и качество разметки

Резюме



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

Резюме



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



Далее: введение в краудсорсинг

Краудсорсинг



Краудсорсинг

→ Краудсорсинг (crowdsourcing) — это способ организации рабочего процесса, в котором большая высокоуровневая задача **делится на множество типовых подзадач**

Краудсорсинг

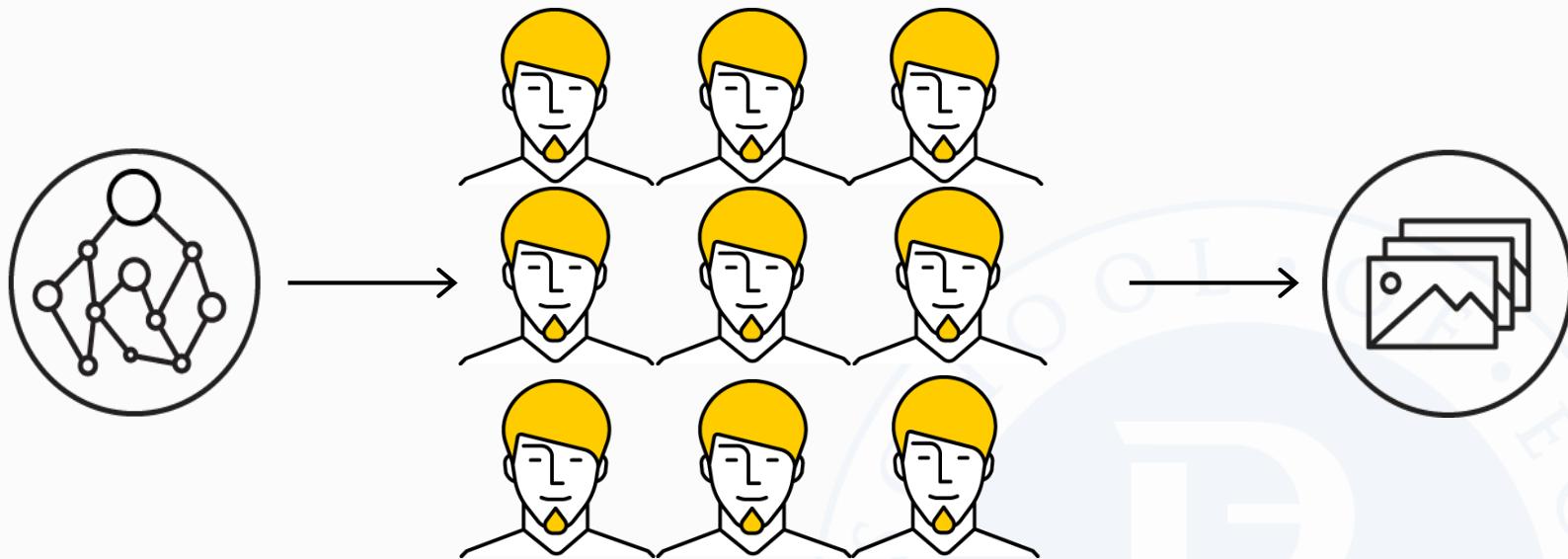
- Краудсорсинг (crowdsourcing) — это способ организации рабочего процесса, в котором большая высокоуровневая задача **делится на множество типовых подзадач**
- Их выполняет большое количество независимых друг от друга исполнителей. Каждый делает одно или несколько простых заданий, и это приводит к **решению высокоуровневой задачи**

Краудсорсинг

- Краудсорсинг (crowdsourcing) — это способ организации рабочего процесса, в котором большая высокоуровневая задача **делится на множество типовых подзадач**
- Их выполняет большое количество независимых друг от друга исполнителей. Каждый делает одно или несколько простых заданий, и это приводит к **решению высокоуровневой задачи**
- Иначе говоря, **краудсорсинг** — это замена экспертизы одного специалиста на «мудрость толпы»

Краудсорсинг

→ Особый способ построения бизнес-процессов



Большая задача

Облако исполнителей

Результат

Пример: бинарная классификация

Эта кошка белая?

Да

Нет



Пример: мульти-классификация

“

If you are a gourmand, I can recommend you the "Real French restaurant", located in the historic cellar, with elements of antique design and quite interesting cuisine. The restaurant is small, but very cozy and romantic. The restaurant is very suitable for romance and even for business meetings.

Is it a feedback?

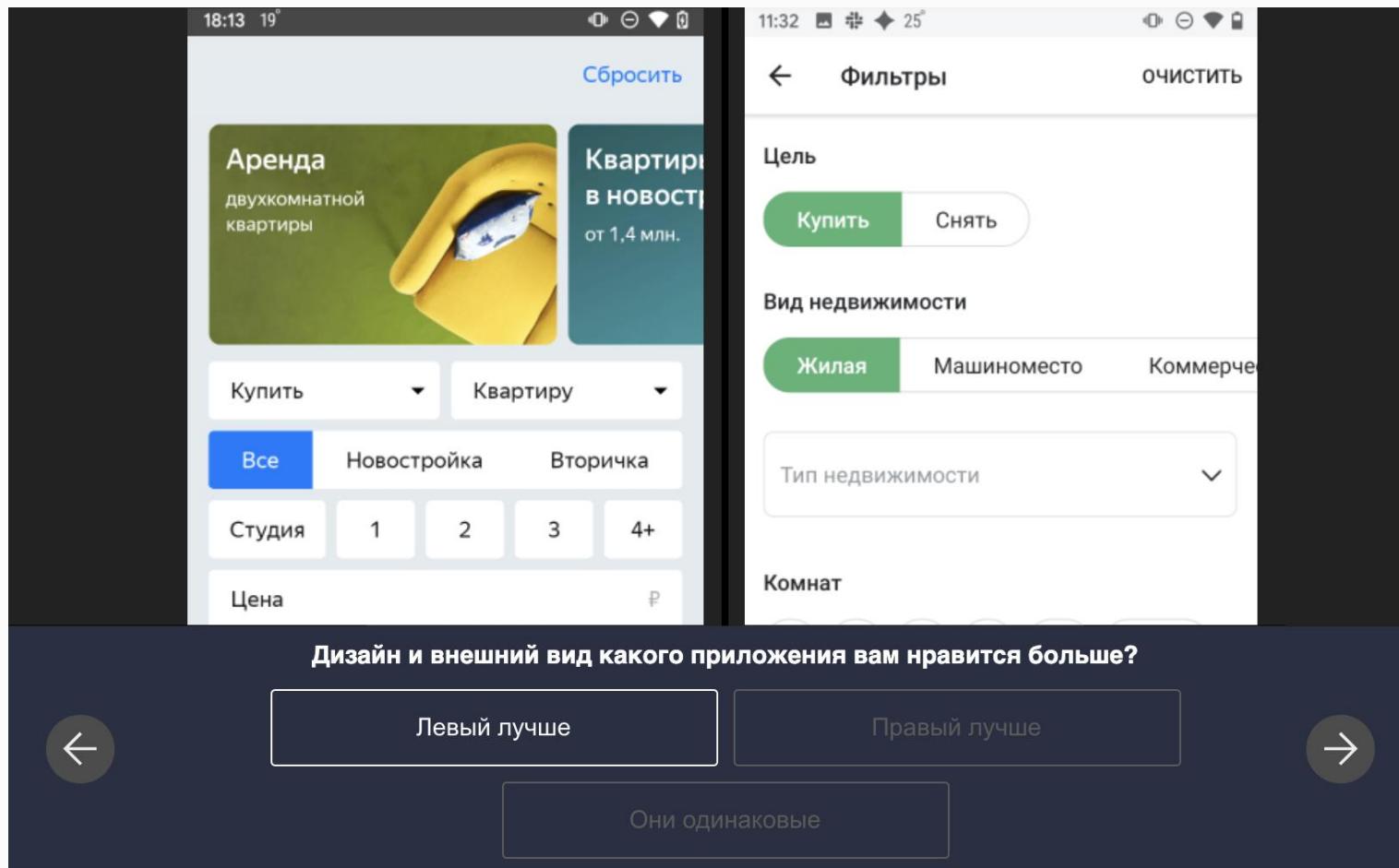
Yes, it is No, it's other comment

Personal information ?

Swearing, vulgarity, insults, aggressive statements ?

Spam, advertising ?

Пример: попарное сравнение



Пример: расшифровка в текст

▶ 0:00 / 0:09 — 🔊 ⋮

1 Один говорящий

2 Несколько говорящих

3 Нет речи / неразборчиво

Текст аннотации

4

Пример: выделение сущностей

Набор зелени	НАЗВАНИЕ Лук	НАЗВАНИЕ укроп	ОБЪЕМ 200г	Q Название
				W Бренд
				E Объем
				R Ошибка
				T Производитель

Пример: запись аудио/фото/видео

Карл у Клары украл кораллы, а Клара у Карла украла кларнет.



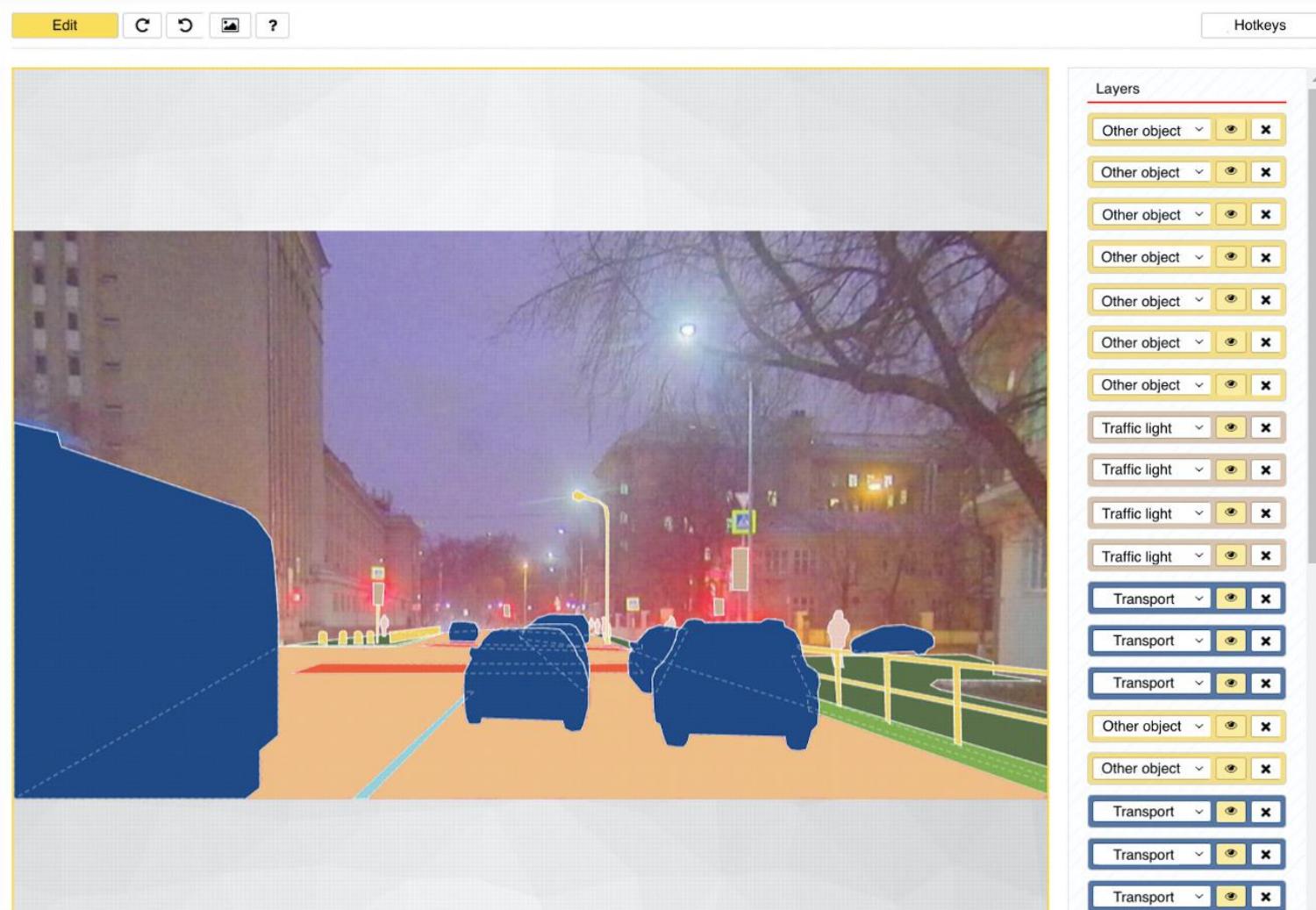
Открыть диктофон

Пример: проведение опросов

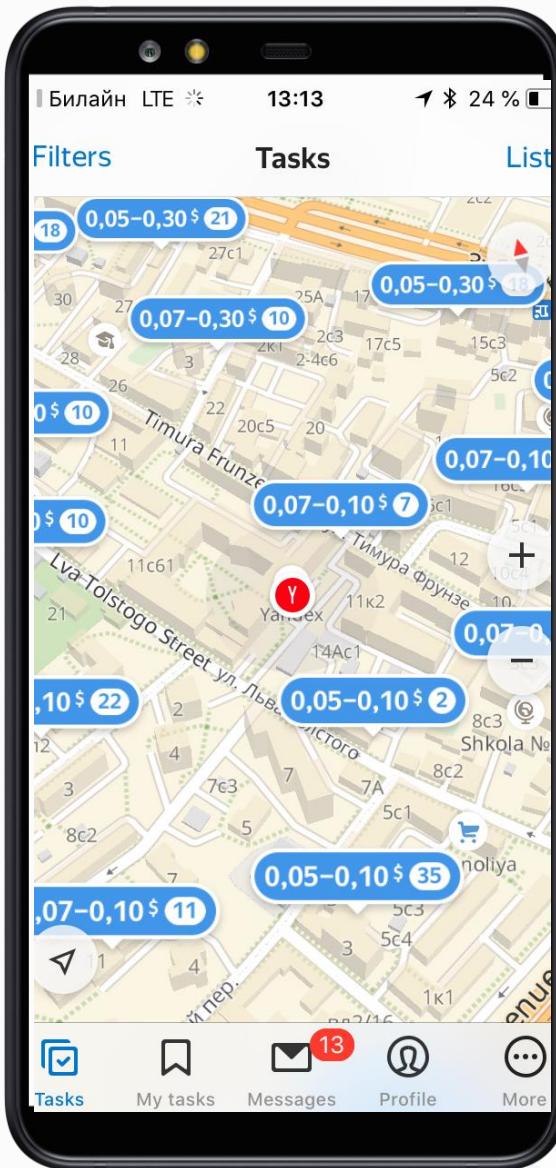
Какое у вас образование?

- Высшее
(университет, институт, академия)
- Среднее специальное
(колледж, училище, ПТУ)
- Среднее
(школа)
- Нет образования

Пример: сегментация изображений



Пример: полевые задания



Двусторонний рынок



Исполнители

Toloka.yandex.ru

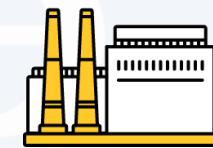


Двусторонний рынок



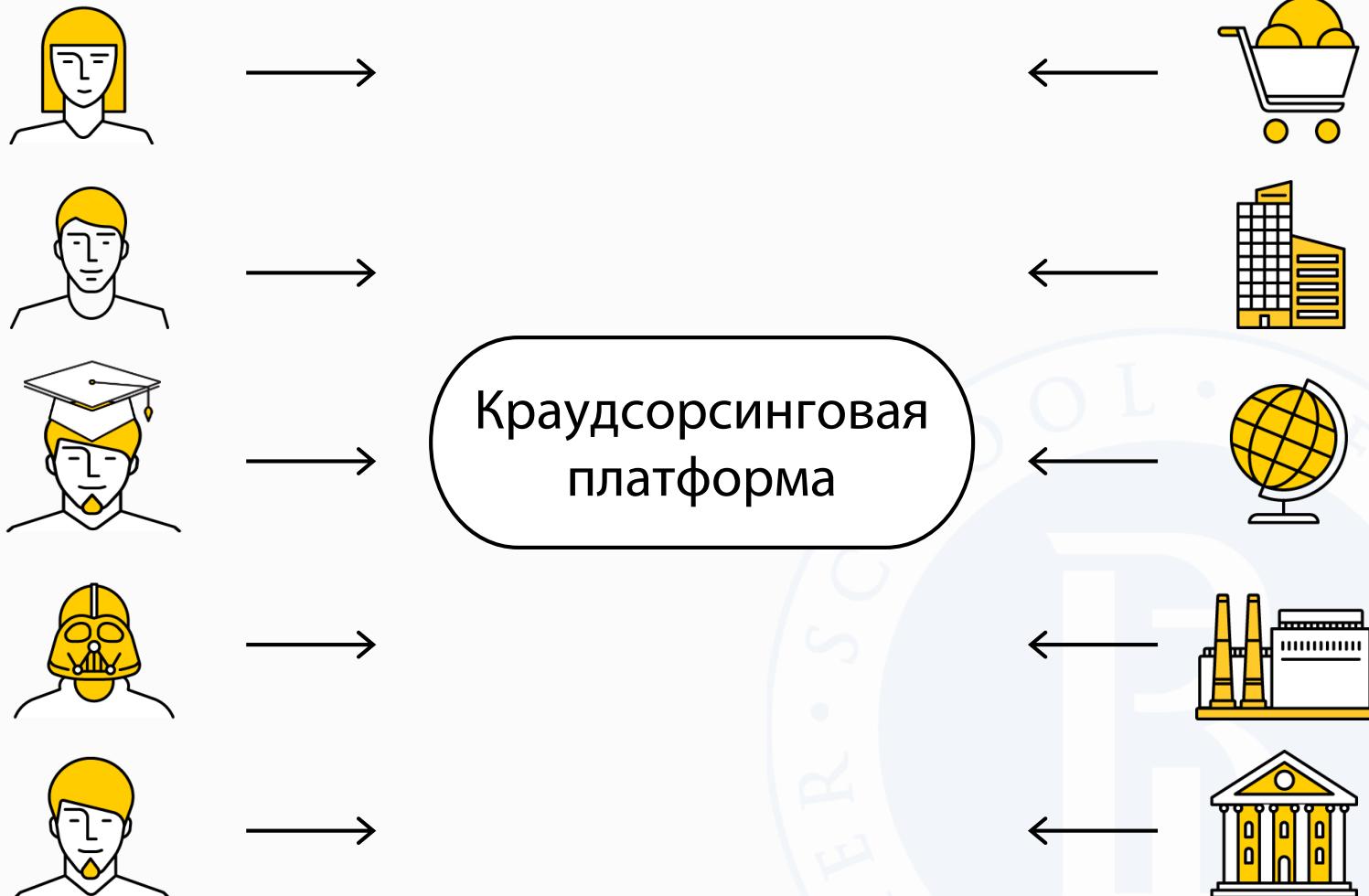
Исполнители

Toloka.yandex.ru



Заказчики

Двусторонний рынок



Исполнители

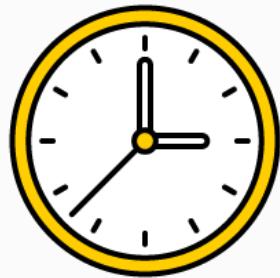
Заказчики

Краудсорсинговые платформы: примеры

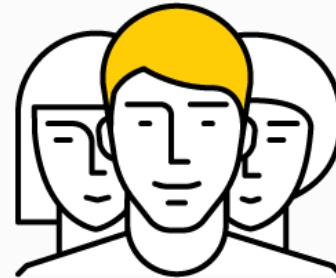
- Yandex.Toloka
- Amazon Mechanical Turk
- Microworkers.com
- Gigwalk
- ClickWorker
- CloudFactory
- Figure Eight
- CrowdSource
- DefinedCrowd



Плюсы краудсорсинговых платформ



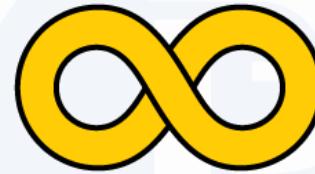
24/7



Разнообразие экспертиз
исполнителей



Широкое покрытие
регионов



Непрерывные
процессы

Резюме



Краудсорсинг позволяет заменить экспертизу одного человека мудростью толпы

Резюме



Краудсорсинг позволяет заменить экспертизу одного человека мудростью толпы



С помощью краудсорсинга можно решить огромное количество задач

Резюме



Краудсорсинг позволяет заменить экспертизу одного человека мудростью толпы



С помощью краудсорсинга можно решить огромное количество задач



Краудсорсинг имеет множество плюсов, но для его успешного использования требуются определенные навыки и умения

Резюме



Краудсорсинг позволяет заменить экспертизу одного человека мудростью толпы



С помощью краудсорсинга можно решить огромное количество задач



Краудсорсинг имеет множество плюсов, но для его успешного использования требуются определенные навыки и умения



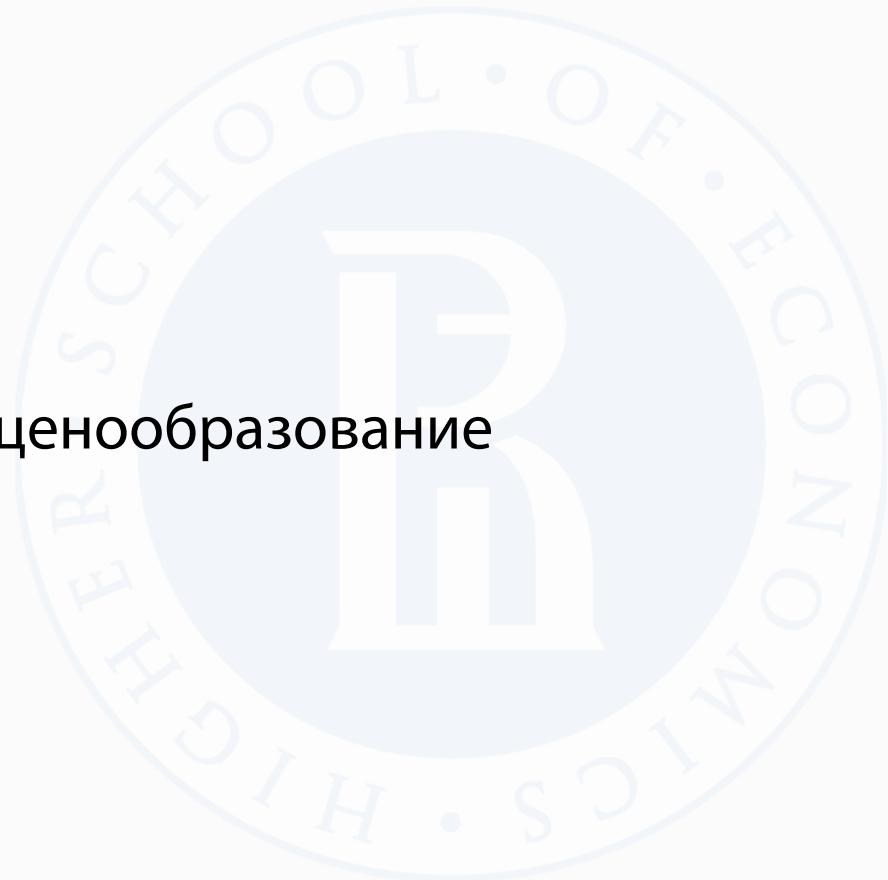
Далее: основные компоненты краудсорсинга

Основные компоненты эффективного и масштабируемого краудсорсинга

Часть 1

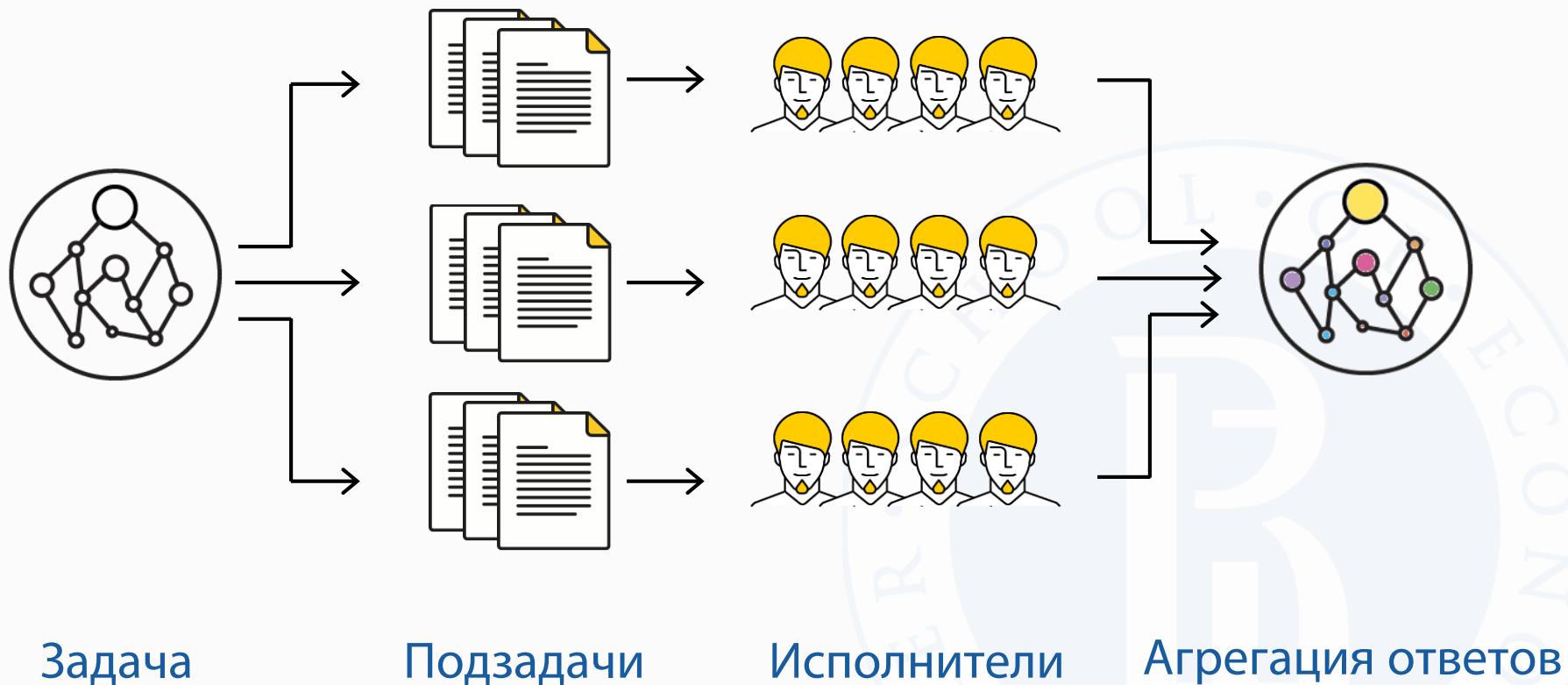
Основные компоненты

- Декомпозиция
- Инструкция
- Интерфейс задания
- Контроль качества
- Агрегация результатов
- Динамическое перекрытие и ценообразование



Декомпозиция

→ Декомпозиция — разбиение сложной задачи на несколько более простых подзадач



Декомпозиция: зачем?

- Исполнители, как правило, не специалисты в вашей задаче

Чем проще задание, тем:

- Проще инструкция
- Легче его выполнить
- Лучше качество исполнения
- Больше исполнителей могут его выполнять

Декомпозиция: когда?

Задание нуждается в декомпозиции, если оно:

- Требует более чем 3-5 действий от исполнителя
- Имеет длинную тяжело читаемую инструкцию

Пример декомпозиции: много вопросов

Все вопросы в одном задании
Плохая практика



Какое животное на фото?

- > Кошка
- > Собака
- > Кролик
- > Коала

Какого оно цвета?

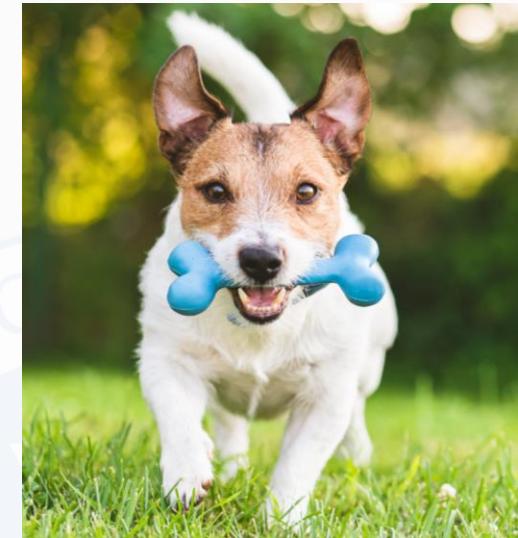
- > Белое
- > Черное
- > Другое

Виден ли его хвост? Оно бежит?

- > Да
- > Нет
- > Да
- > Нет

Пример декомпозиции: много вопросов

Каждый вопрос в отдельном задании
Хорошая практика



Какое животное на фото?

- › Кошка
- › Собака
- › Кролик
- › Коала

Какого цвета животное?

- › Белое
- › Черное
- › Другое

Виден ли хвост?

- › Да
- › Нет

Пример декомпозиции: проверка ответа



Задание: Выделите всех хомяков на фото



Пример декомпозиции: проверка ответа

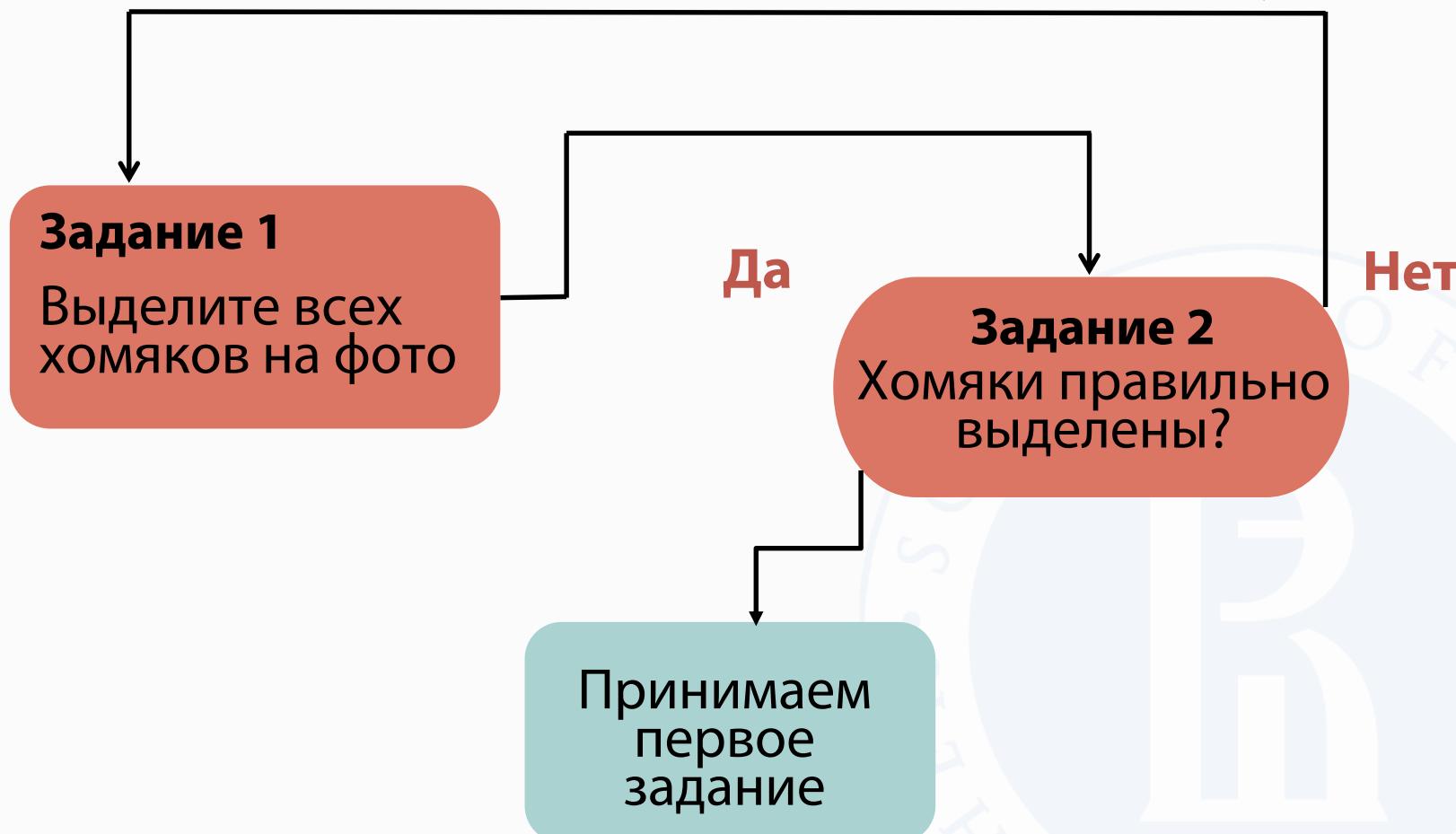
- Задание: Выделите всех хомяков на фото
- Проблема: Выделение может быть сделано по-разному
- Следовательно, сложно сравнить с ответами в контрольных заданиях и агрегировать ответы от разных исполнителей

Хорошее решение: выдать ответ на проверку другому исполнителю



Пример декомпозиции: проверка ответа

Отклоняем задание и отправляем на переразметку



Декомпозиция для полевых исследований

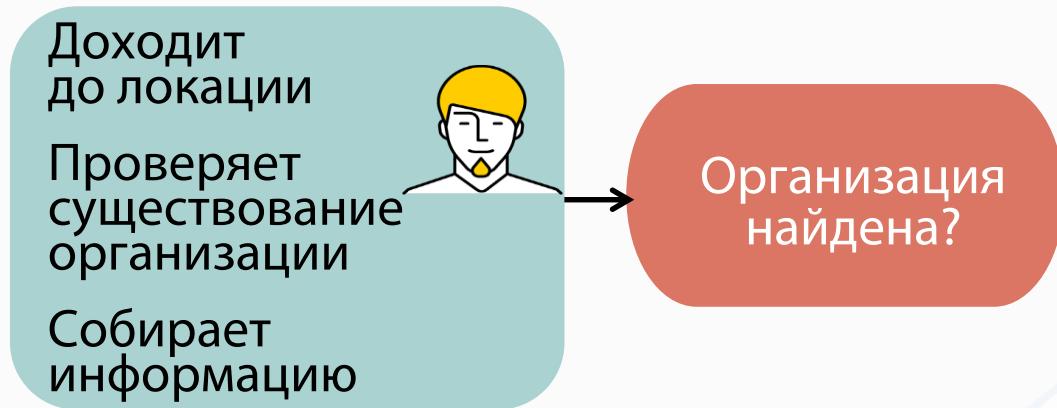
Доходит
до локации



Проверяет
существование
организации

Собирает
информацию

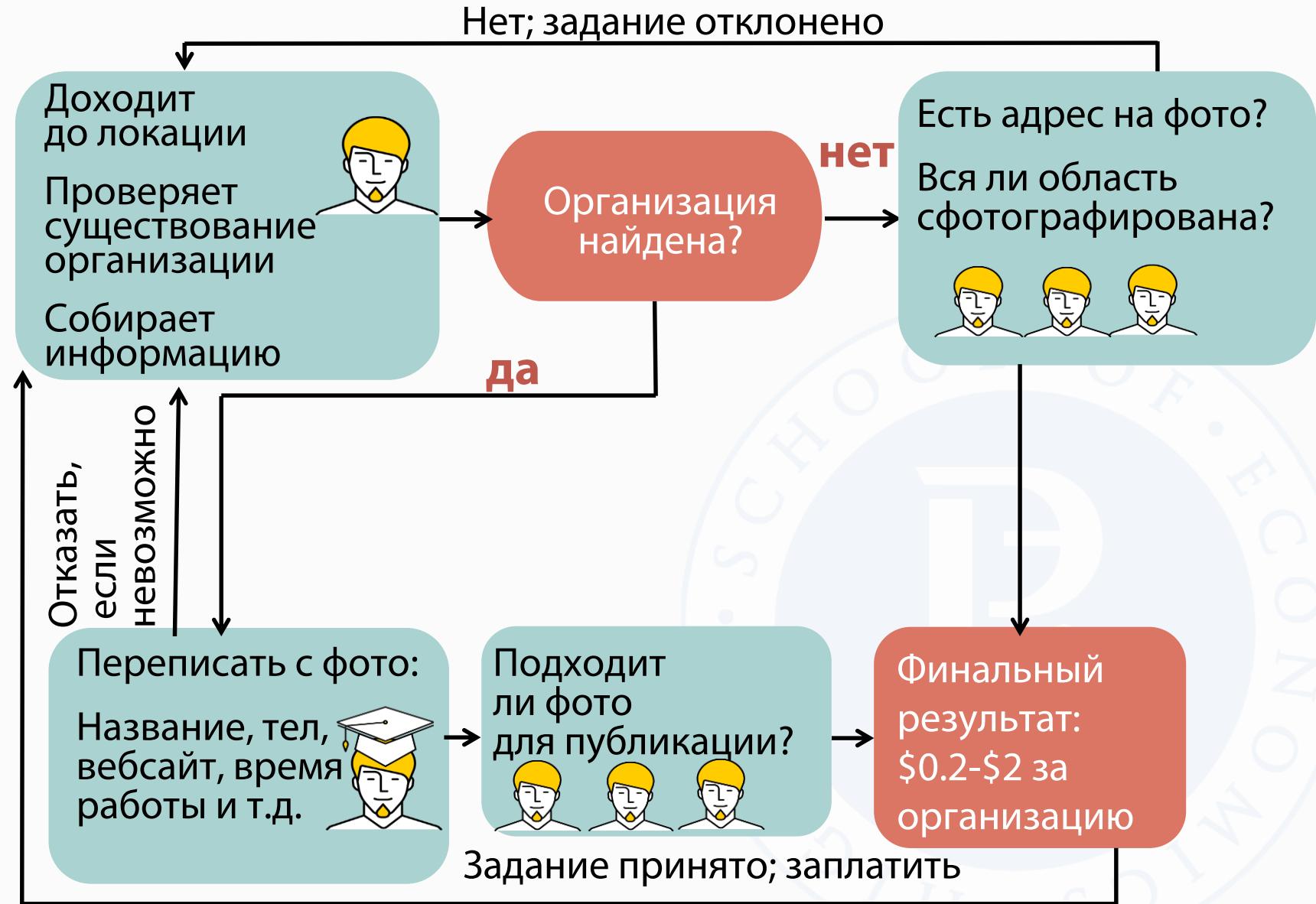
Декомпозиция для полевых исследований



Декомпозиция для полевых исследований



Декомпозиция для полевых исследований



Инструкция

→ Инструкция — указания, свод правил, устанавливающий порядок и способ осуществления, выполнения чего-либо

Инструкция

- Инструкция — указания, свод правил, устанавливающий порядок и способ осуществления, выполнения чего-либо
- Инструкция — основной способ коммуникации между заказчиком и исполнителями

Инструкция

- Инструкция — указания, свод правил, устанавливающий порядок и способ осуществления, выполнения чего-либо
- Инструкция — основной способ коммуникации между заказчиком и исполнителями
- Грамотно написанная инструкция необходима для качественно выполняемого задания

Инструкция: типичная структура

- Цель задания, предложенного к выполнению
- Описание интерфейса задания
- Алгоритм требуемых действий
- Примеры хороших и плохих ответов
- Алгоритм действий и примеры для редких случаев
- Ссылки на дополнительные материалы

Инструкция: типичная структура

- Цель задания, предложенного к выполнению
- Описание интерфейса задания
- Алгоритм требуемых действий
- Примеры хороших и плохих ответов
- Алгоритм действий и примеры для редких случаев
- Ссылки на дополнительные материалы

Большинство
подводных
камней здесь

Неоднозначность инструкции для редких случаев: пример

Эта кошка белая?

Да

Нет



OK: ответ и задание кажутся очевидными

Неоднозначность инструкции для редких случаев: пример

Эта кошка белая?

Да

Нет



Какой корректный ответ?

Неоднозначность инструкции для редких случаев: пример

Эта кошка белая?

Да

Нет



Как исправить:

В инструкции: уточните, что вы имеете в виду под «белая кошка»

В интерфейсе: добавьте кнопку «Не знаю», чтобы отловить редкий случай

Неоднозначность инструкции для редких случаев: пример

Эта кошка белая?

Да

Нет



Редкий случай: много кошек

Неоднозначность инструкции для редких случаев: пример

Эта кошка белая?

Да

Нет



Редкий случай: не кошка

Неоднозначность инструкции для редких случаев: пример

Эта кошка белая?

Да

Нет

404: Cannot download
the image

Редкий случай: картинка не была
показана

Неоднозначность инструкции для редких случаев: пример

**Тяжело продумать все возможные
ситуации, но вы можете:**

В инструкции: объяснить, что делать
в нестандартной ситуации

В интерфейсе: добавить текстовое поле,
дав возможность исполнителю
сообщить о ситуации



Интерфейс задания

- Реализуется с помощью JS, HTML, CSS
- Есть готовые шаблоны под множество задач

От удобства интерфейса зависит:

- Скорость выполнения заданий
- Стоимость задания
- Процент ошибок при выполнении заданий

Интерфейс задания: лучшие практики



Для более быстрого исполнения:



«Горячие клавиши» для чекбоксов / кнопок / и т.п.



Снижение количества переходов на сторонние сайты



Оптимальное расположение заданий на странице

Интерфейс задания: лучшие практики



Для лучшего качества и снижения числа ошибок:



Динамический интерфейс
(скрывать/показывать поля ввода)



Адаптивный интерфейс (хороший на любом устройстве
и в любом разрешении экрана)



Динамическая проверка введенных данных



Всегда проверяйте ваши интерфейсы
(предпросмотр шаблона)

Резюме



Декомпозиция необходима, если вы хотите решить сложную задачу с помощью краудсорсинга

Резюме



Декомпозиция необходима, если вы хотите решить сложную задачу с помощью краудсорсинга



Грамотно написанная инструкция — залог взаимопонимания между заказчиком и исполнителями

Резюме



Декомпозиция необходима, если вы хотите решить сложную задачу с помощью краудсорсинга



Грамотно написанная инструкция — залог взаимопонимания между заказчиком и исполнителями



Удобный и проработанный интерфейс задания позволяет исполнителям быстрее выполнять задания и меньше ошибаться

Резюме



Декомпозиция необходима, если вы хотите решить сложную задачу с помощью краудсорсинга



Грамотно написанная инструкция — залог взаимопонимания между заказчиком и исполнителями



Удобный и проработанный интерфейс задания позволяет исполнителям быстрее выполнять задания и меньше ошибаться



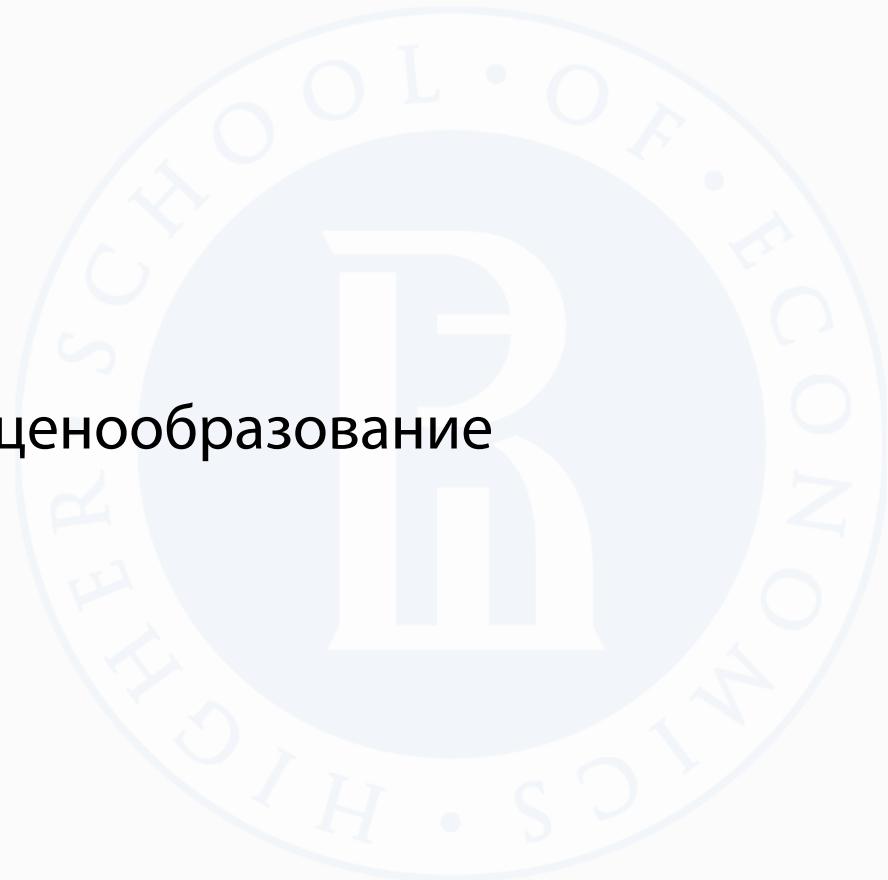
Далее: Основные компоненты эффективного и масштабируемого краудсорсинга. Часть 2

Основные компоненты эффективного и масштабируемого краудсорсинга

Часть 2

Основные компоненты

- Декомпозиция
- Инструкция
- Интерфейс задания
- Контроль качества
- Агрегация результатов
- Динамическое перекрытие и ценообразование



Контроль качества

→ Контроль качества — набор техник, правил, концепций, позволяющих поддерживать качество размечаемых данных на нужном вам уровне



Контроль качества

«До» выполнения задания

- Отбор исполнителей
- Хорошо написанная инструкция



Контроль качества

«Во время» выполнения задания

- Проверочные задания (т.н. golden set и honey pots)
- Хорошо построенный интерфейс
- Мотивация
(например, стимулирующее ценообразование)
- Техники выявления роботов и мошенников
(быстрые ответы)

Контроль качества

«После» выполнения задания

- Пост-проверка (отложенная приемка)
- Консенсус между исполнителями и агрегация ответов

Отбор исполнителей

Фильтруйте по:

- Статичным свойствам (образование, язык и т.п.)
- Вычисляемым свойствам (браузер, регион и т.п.)
- Навыкам (для контроля уровня качества на ваших заданиях и для получения исполнителей с лучшим качеством по прошлым проектам)

Отбор исполнителей

Обучайте исполнять ваши задания:

- Используйте тренировочные задания, чтобы показать, как их выполнять
- Используйте экзаменационные задания для оценки уровня обучения



Проверочные задания

→ Проверочные задания — это задания с заранее известным корректным ответом, показываемые исполнителям для оценки их качества



Проверочные задания

Лучшие практики:

- Распределение ответов в golden set равно распределению во всех заданиях
- Но редких вариантов ответа должно быть больше
- Регулярно обновляйте ваше множество проверочных заданий в целях безопасности
- Автоматическая генерация проверочных заданий исполнителями: задания с ответами, в которых сильно уверены (например, агрегация ответов от большого числа исполнителей)

Мотивация

- Бонусы за регулярное хорошее качество работы
- Геймификация (например, «ачивки», соревнования)
- Цена, зависящая от качества



Техники исключения мошенников

- Контролировать скорость ответов (т. н. «быстрые ответы»)
- Проверять, был ли переход по ссылке до ответа
- Проверять, был ли воспроизведен видео-файл до ответа



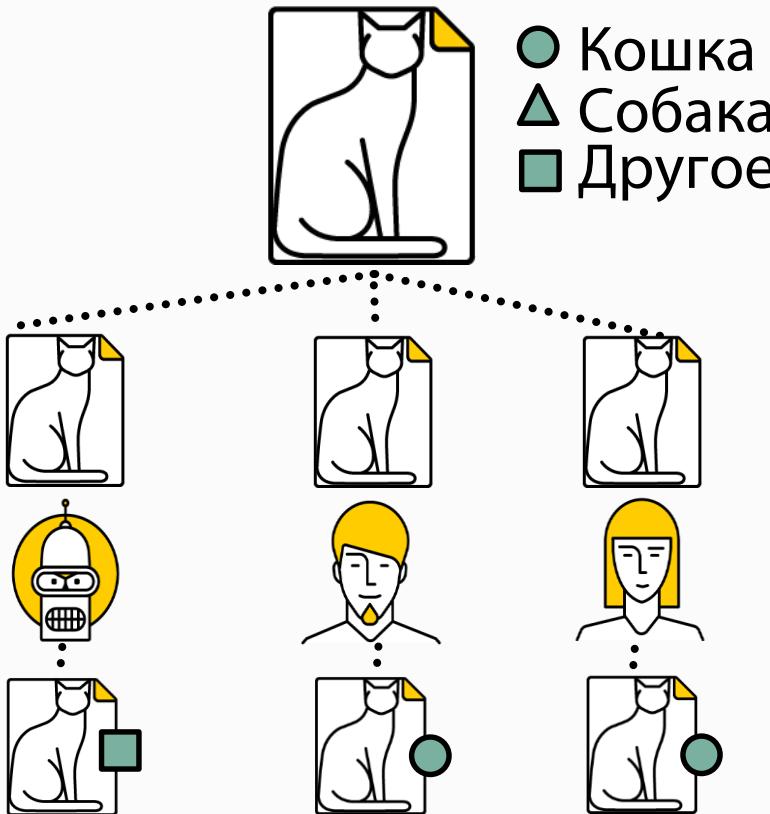
Пост-проверка (отложенная приемка)

Исполнитель получает деньги, только если ответ принят

- Используется, когда задание сложное (ни проверочные задания, ни модели консенсуса не работают)
- Может выполняться вами самостоятельно, но можно использовать других исполнителей в задании другого типа

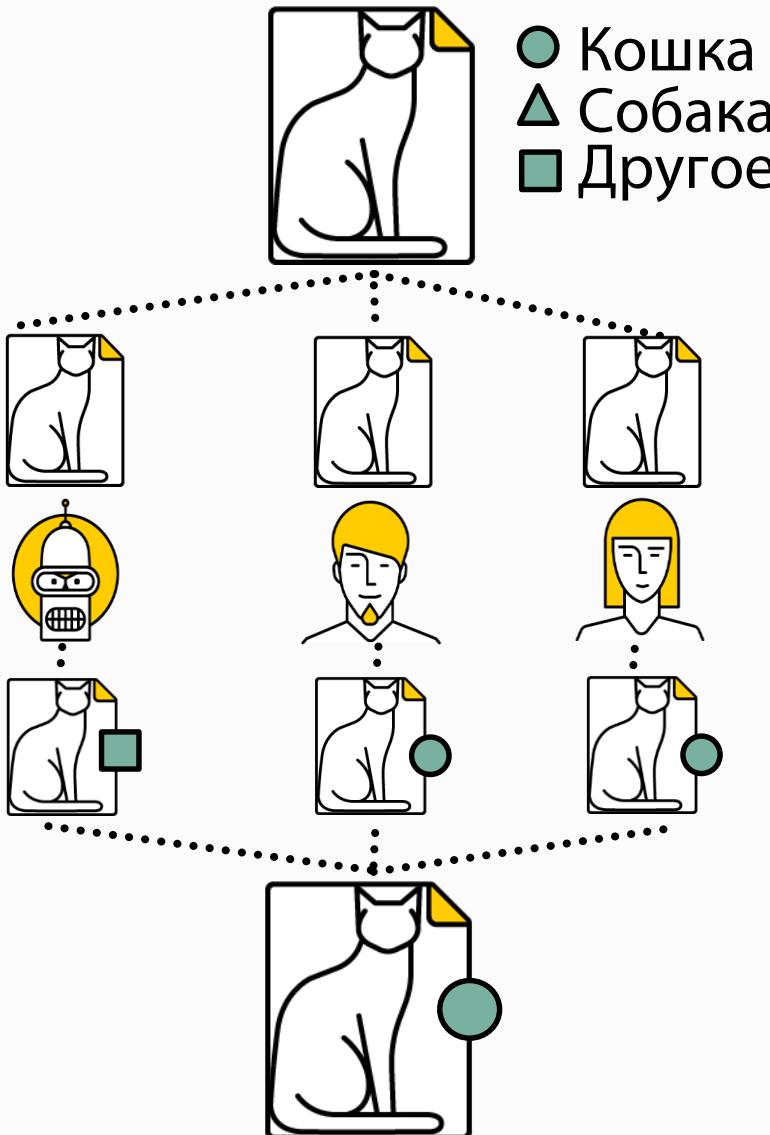
Таким образом, мы имеем дело с иерархией проектов
(применяем декомпозицию)

Консенсус между исполнителями



- Выдаем одно задание нескольким исполнителям
- Исполнители ставят свою оценку объекту
- Выводим информацию о качестве из консенсуса

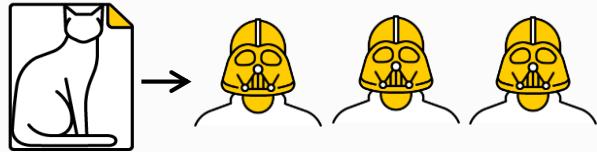
Агрегация



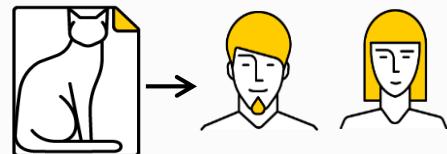
- Загружаем несколько копий каждого объекта
- Исполнители ставят свою оценку объекту
- Агрегируем несколько оценок в одну более точную оценку для каждого изображения
- Простейший способ: выбрать самый популярный ответ (Majority Vote)

Динамическое перекрытие

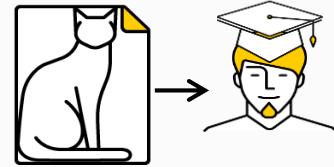
→ Получить агрегированные оценки заданного качества, используя наименьшее число шумных оценок



Несколько неизвестных исполнителей



Пара исполнителей с известным хорошим качеством



Один эксперт с высоким качеством

Цены зависят от

Структуры задания

- Оплата за страницу заданий (т.н. task suite)
 - Время, необходимое для выполнения задания: контроль дохода в час
- Экономические аспекты рынка:
- Чем меньше предложение исполнителей (e.g. из-за редких навыков), тем выше цена
 - Как быстро вам нужен результат (latency)?

Цены зависят от

Качество результата

- Стимулируйте лучшее качество ценой, зависящей от качества



Резюме



Получить хорошее качество можно с помощью:

- Отбора исполнителей
- Хорошо написанной инструкции
- Добавления проверочных заданий
- Грамотно построенного интерфейса

Резюме



Получить хорошее качество можно с помощью:

- Мотивации исполнителей
- Правил контроля качества для выявления мошенников
- Пост-проверки (отложенной приемки)
- Консенсуса между исполнителями
- Агрегации ответов

Резюме



Динамическое перекрытие позволяет снизить стоимость, при этом не потерять в качестве

Резюме



Динамическое перекрытие позволяет снизить стоимость, при этом не потерять в качестве



Методы агрегации позволяют объединять ответы исполнителей, за счет этого достигается высокое качество

Резюме



Динамическое перекрытие позволяет снизить стоимость, при этом не потерять в качестве



Методы агрегации позволяют объединять ответы исполнителей, за счет этого достигается высокое качество



Цену на задание выставляет сам заказчик. Она зависит от сложности задачи, требуемого качества, а также от того, какие цены у «конкурентов»

**Хорошая декомпозиция —
ключ к успеху**



		Простая инструкция
IF	Простой в использовании интерфейс задания	
Хорошая декомпозиция	Исполнители делают задания с более хорошим качеством	Стандартные методы агрегации работают хорошо
THEN	Качество проще контролировать	Цены проще контролировать и оптимизировать