



ДАТАСЕТЫ ДЛЯ МАШИННОГО ОБУЧЕНИЯ

Понятие датасета

- **Обработанный и структурированный массив данных.**
- У каждого объекта есть конкретные свойства: **признаки**, связи между объектами или определённое место в выборке данных.

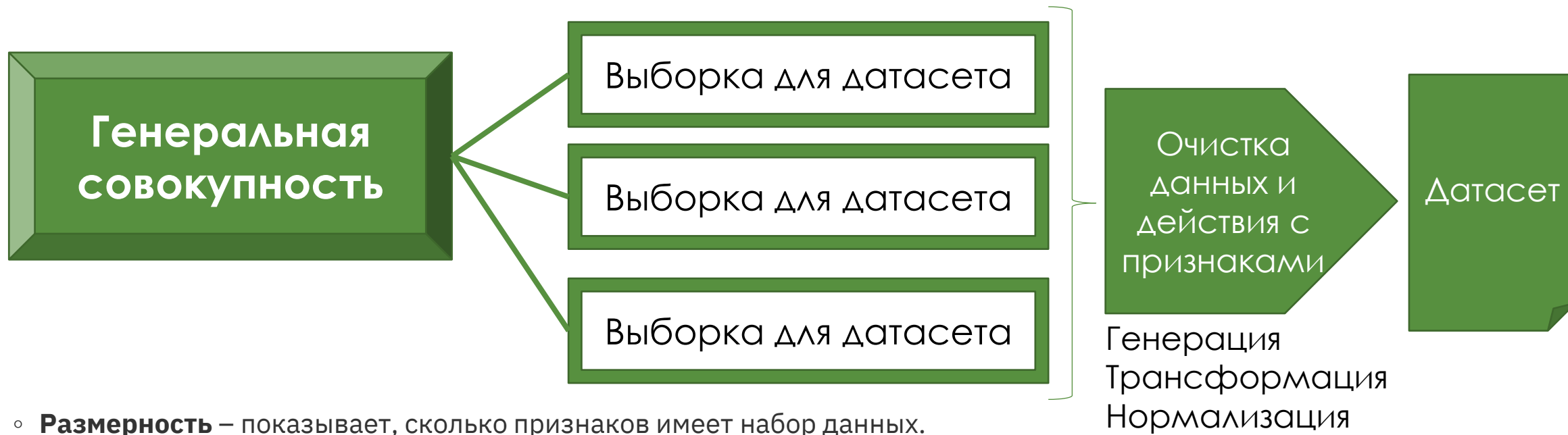
Примеры

- статистика покупок в магазине;
- географическое расположение офисов;
- демографические признаки населения;
- соответствие звуков к аудиотексту;
- заболевания с конкретными симптомами.

Способы сбора: вручную и автоматически



Характеристики датасета

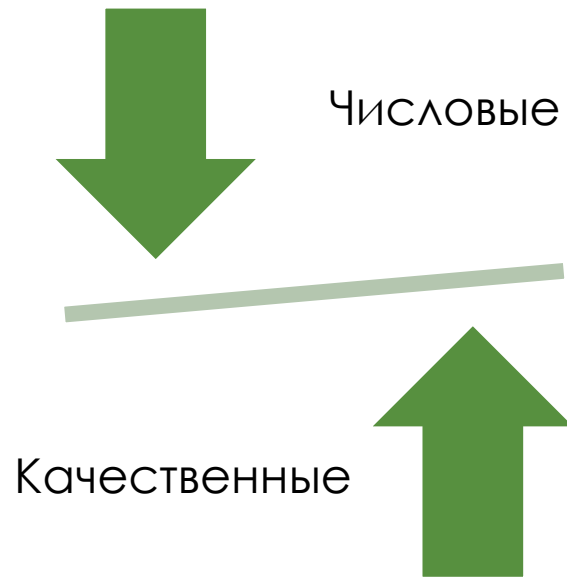


- **Размерность** – показывает, сколько признаков имеет набор данных.
- **Разреженность** – показатель, характеризующийся заполненностью датасета, т.е. те ячейки, которые заполнены ненулевыми значениями.
- **Разрешение.** Характеризуется возможностью обнаруживать какое-либо явление, при условии, что данные подробны ровно настолько, насколько это соответствует решению задачи.

Проектирование признаков

- преобразования **исходной** информации в **набор данных**.
- **Набор данных** — это коллекция размеченных образцов

$\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, где \mathbf{x}_i — вектор **информативных** признаков



Кодирование – преобразование качественного в числовой

Биннинг – преобразование числового в качественный

Нормализация - процесс преобразования фактического диапазона значений числового признака в стандартный диапазон значений

$$\bar{x}^{(j)} = \frac{x^{(j)} - \min^{(j)}}{\max^{(j)} - \min^{(j)}}$$

Отсутствующие значения признаков

- удалить данные с отсутствующими значениями признаков из набора данных (набор данных достаточно велик);
- использовать алгоритм обучения, умеющий работать с отсутствующими значениями (зависит от библиотеки и конкретной реализации алгоритма);
- ИСПОЛЬЗОВАТЬ ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ ВОССТАНОВЛЕНИЯ ДАННЫХ.

Методы восстановления данных

- Замена отсутствующего значения средним значением
- Замена значением, выходящим за пределы диапазона нормальных значений
- Решить задачу регрессии, используя отсутствующее значение в качестве целевой переменной.

Виды датасетов

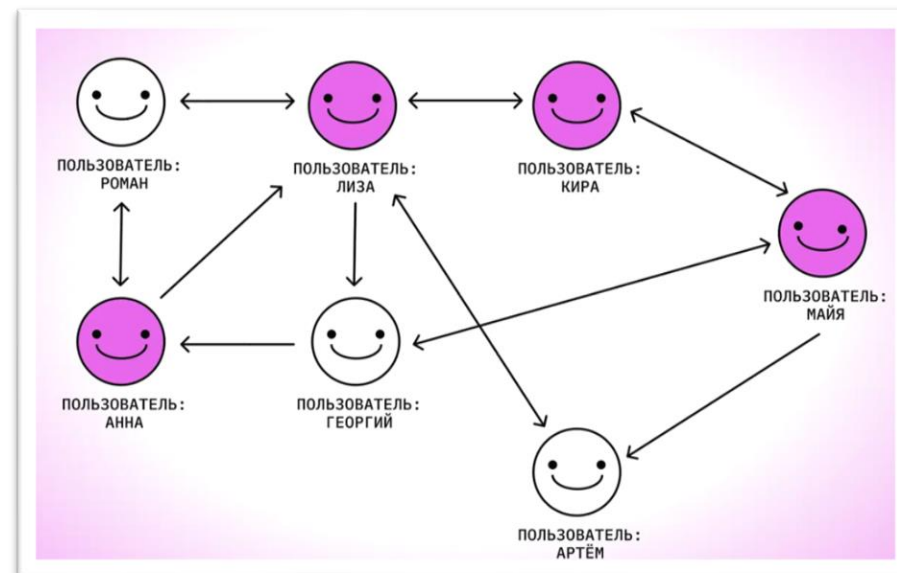
	пол м	пол ж	возраст	родной город москва
иван	1	0	30	1
пётр	1	0	25	0
ольга	0	1	24	0

Простая запись

Упорядоченная запись

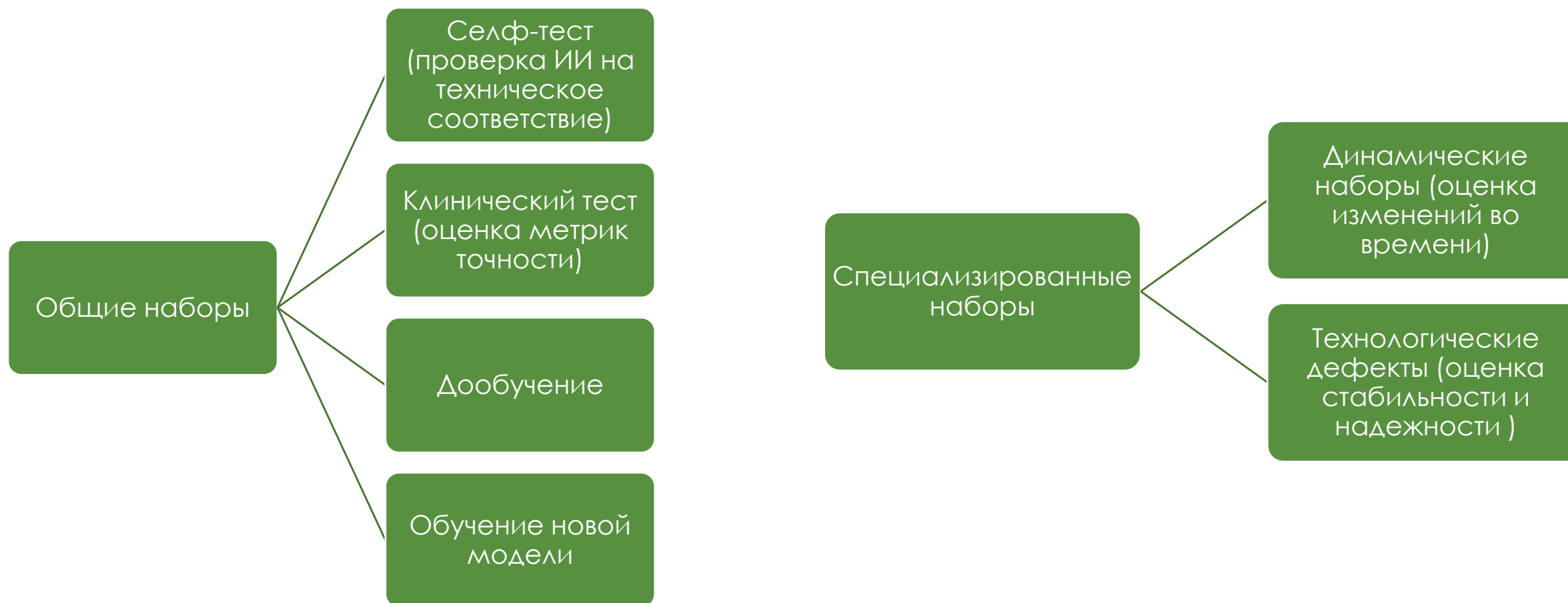
- Временной ряд
- Пространственные
- Последовательные

В **Российской Федерации датасет** приравнивается к **базе данных** и подлежит добровольной государственной регистрации в качестве результата интеллектуальной деятельности.

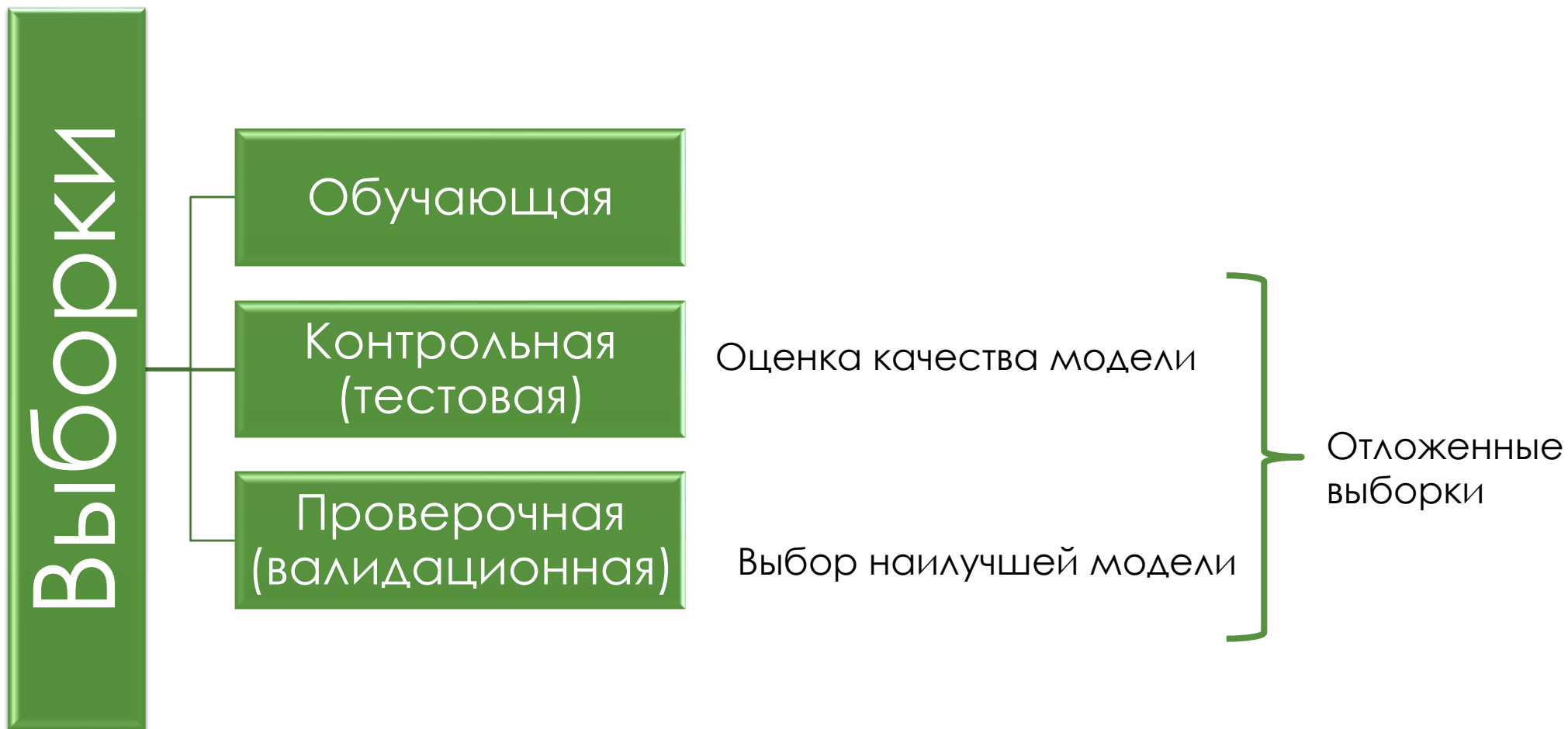


Граф

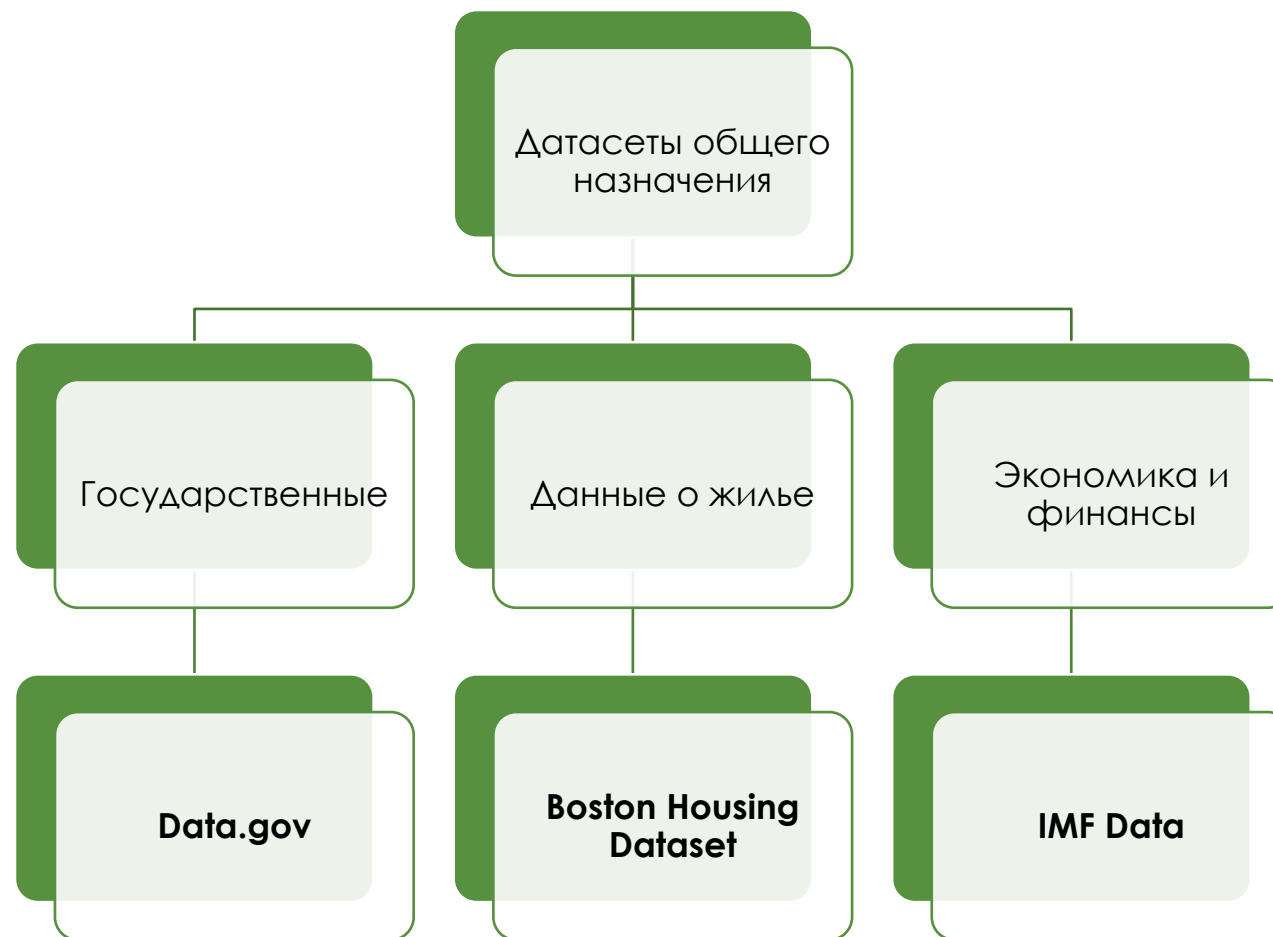
Классификация датасетов



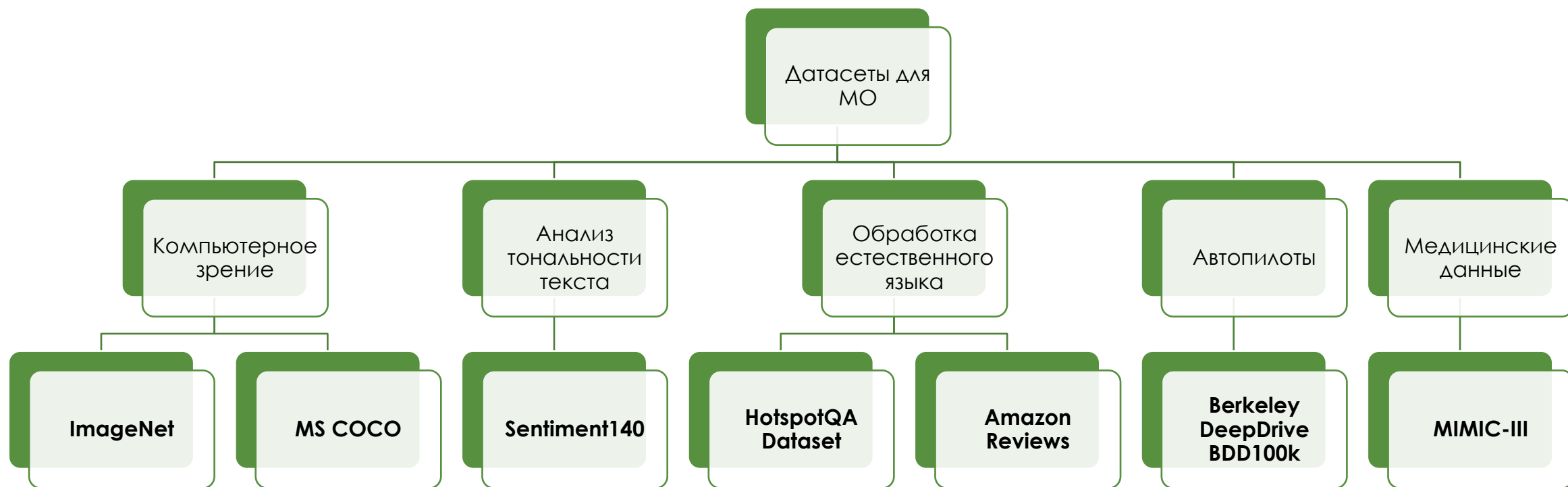
Типы выборок



Примеры датасетов



Примеры датасетов



Примеры датасетов, открытых для использования

- [World Bank Open Data](#). Наборы данных о демографии и экономических показателях.
- [IMF Data](#). Датасеты о финансах и ценах на товары.
- [Google Trends](#). Данные о поисковой статистике и трендовых запросах.
- [xView](#). Большой набор воздушных снимков Земли с аннотациями.
- [Labelme](#). Большой датасет с уже размеченными изображениями.
- [Labelled Faces in the Wild](#). 13 тысяч размеченных изображений лиц.
- [HotspotQA Dataset](#). Датасет с вопросами-ответами для генерации ответов на часто задаваемые простые вопросы.
- [Berkeley DeepDrive BDD100k](#). Тысячи часов вождения для обучения автопилотов.
- [MIMIC-III](#). Обезличенные медицинские данные пациентов.
- [CREMA-D](#) — датасет для распознавания эмоций по записи голоса.

Поиск датасетов

Google Dataset Search



Dataset Search

Поиск наборов данных



Попробовать [коронавирус COVID-19](#) или [water quality site:canada.ca](#).

Подробнее [о Поиске наборов данных...](#)

▼ Последнее обновление

▼ Формат скачивания

▼ Права использования

▼ Тема

Бесплатно

Сохраненные наборы данных

do, xls, xml

Обновлено Jun 23, 2018

W

База кадастровой стоимости
объектов недвижимости (за...
data.wu.ac.at

html, xlsx

Обновлено Nov 22, 2013



Текущая рыночная стоимость
жилых зданий

knoema.com

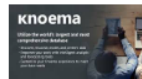
csv, json, sdmx, xls

Обновлено May 10, 2020

S

Ремонт квартир цены,
строительство домов цена,...
stroitel-kharkov.net

pdf



Текущая рыночная стоимость жилых зданий

Открыть сайт

[Knoema | knoema.com](https://knoema.com)

json, xls, csv, sdmx

Набор данных обновлен

May 10, 2020

Набор данных предоставлен

[Knoema](https://knoema.com)

Авторы

Federal State Statistics Service, Russia

Период

2011 - 2017

Географическая область

Russian Federation, Ulyanovsk Region, Kaluga Region, Russian Federation, Rostov Region, Russian Federation, Republic of, Karelia, Russian Federation, Krasnoyarsk Krai, Russian Federation, Sakhalin Region, Russian Federation, Orel Region, Russian Federation, Komi, Republic of, Russian Federation,



ru.knoema.com

csv, json, sdmx, xls

Обновлено Jan 22, 2018

Палау Ожидаемая
продолжительность обучени...

ru.knoema.com

csv, json, sdmx, xls

Обновлено Jan 22, 2018

Российская Федерация
Ожидаемая...

ru.knoema.com

csv, json, sdmx, xls

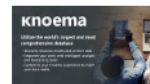
Обновлено Jan 22, 2018

Колумбия Ожидаемая
продолжительность обучени...

ru.knoema.com

csv, json, sdmx, xls

Обновлено Jan 22, 2018



Российская Федерация Ожидаемая продолжительность обучения (высшее образование)

Открыть сайт

[Мировой Атлас Данных | ru.k...](#)

csv, sdmx, xls, json

Набор данных обновлен

Jan 22, 2018

Набор данных создан и предоставлен следующими организациями:

[Knoema](#)

Период

2003 - 2018

Географическая область

Россия

Измеряемые переменные

Ожидаемая продолжительность обучения (высшее образование)

Описание

Площадка для соревнований по машинному обучению

kaggle

Competitions

Datasets

Code

Discussions

Courses

...

Search

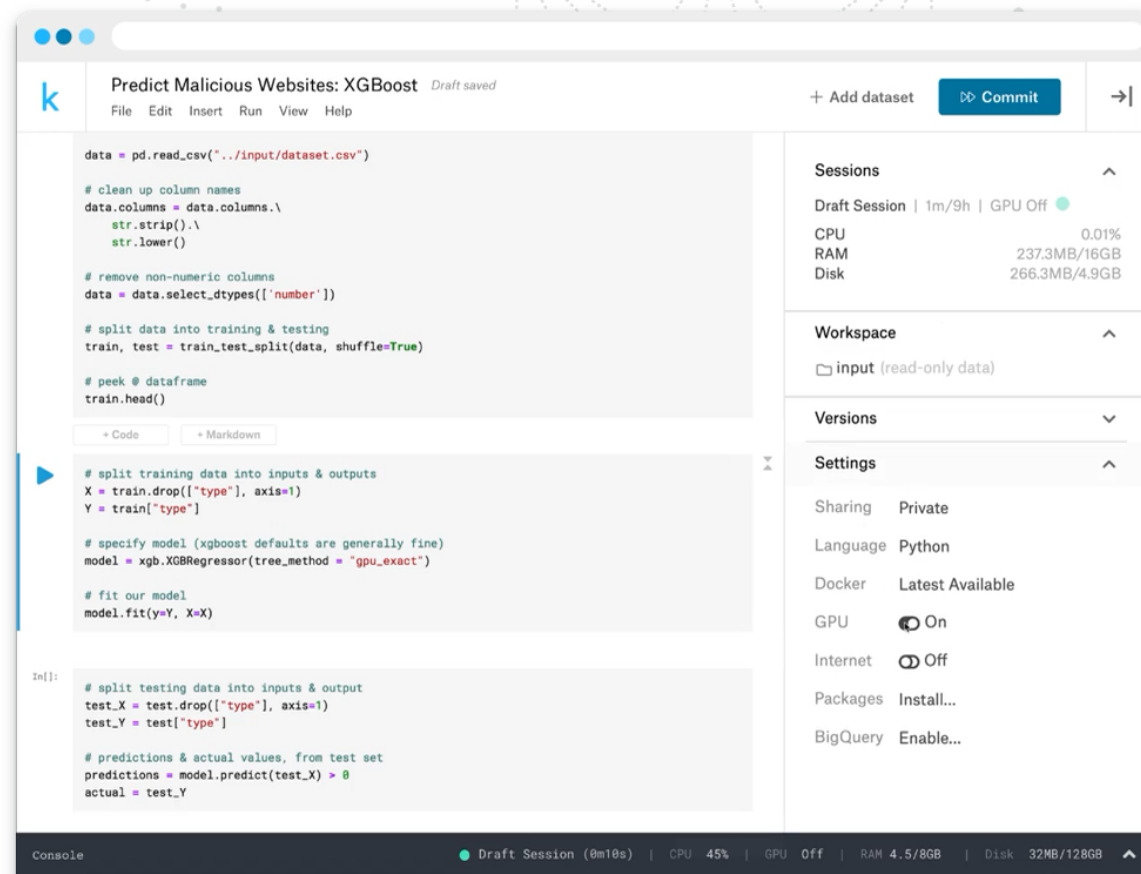
Sign In

Start with more than a blinking cursor

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access GPUs at no cost to you and a huge repository of community published data & code.

 REGISTER WITH GOOGLE

Register with Email



```
data = pd.read_csv("../input/dataset.csv")

# clean up column names
data.columns = data.columns.\
    str.strip().\
    str.lower()

# remove non-numeric columns
data = data.select_dtypes(['number'])

# split data into training & testing
train, test = train_test_split(data, shuffle=True)

# peek @ dataframe
train.head()

# split training data into inputs & outputs
X = train.drop(["type"], axis=1)
Y = train["type"]

# specify model (xgboost defaults are generally fine)
model = xgb.XGBRegressor(tree_method = "gpu_exact")

# fit our model
model.fit(y=Y, X=X)

# split testing data into inputs & output
test_X = test.drop(["type"], axis=1)
test_Y = test["type"]

# predictions & actual values, from test set
predictions = model.predict(test_X) > 0
actual = test_Y
```

Sessions

Draft Session | 1m/9h | GPU Off

CPU 0.01%

RAM 237.3MB/16GB

Disk 266.3MB/4.9GB

Workspace

input (read-only data)

Versions

Settings

Sharing Private

Language Python

Docker Latest Available

GPU On

Internet Off

Packages Install...

BigQuery Enable...

Console

Draft Session (0m10s) | CPU 45% | GPU Off | RAM 4.5/8GB | Disk 32MB/128GB

Kaggle

17.01.2023

Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

[+ New Dataset](#)

[All datasets](#)
[Computer Science](#)
[Education](#)
[Classification](#)
[Computer Vision](#)
[NLP](#)
[Data Visualization](#)
[Pre-Trained Model](#)

Trending Datasets

[See All](#)


(TOP 50)List of most expensive films 🎬🎬🏆...

Devrim Tuncer · Updated 5 days ago



Trending Music on Instagram and Snapchat

Ruchi Bhatia · Updated 18 hours ago



Employee-Attrition-Rate

Prachi Gopalani · Updated 16 days ago
Usability 9.4 · 19 kB



Carvana - Predict Car Prices

Ravi Shah · Updated 8 days ago

Search

Sign In

Register

Datasets

+ New Dataset

Search 5,497 datasets

Filters

Education X

5,497 Datasets

Hotness



QS World ranked Universities (2018-2022)

Aklima Akter Rimi · Updated 8 days ago

Usability 10.0 · 1 File (CSV) · 53 kB

32

Bronze



How Does Daily Yoga Impact Screen Time Habits

The Devastator · Updated 22 days ago

Usability 10.0 · 1 File (CSV) · 742 B

24

Bronze



College Majors and their Graduates

The Devastator · Updated a month ago

Usability 10.0 · 5 Files (CSV) · 40 kB

48

Bronze

17.01.2023

Udemy Courses

209K courses detailed information and comments



Data Card Code (4) Discussion (1)

About Dataset

This dataset contains detailed information on all available Udemy courses on Oct 10, 2022. This data was provided in the "Course_info.csv" file. Also, over 9 million comments were collected and provided in the "Comments.csv" file.

The information of over 209k courses was collected by web scraping the Udemy website. Udemy holds 209,734 courses and 73,514 instructors teaching courses in 79 languages in 13 different categories.

The related notebook was uploaded [here](#).

If you are interested in analytical data about online learning platforms, I recommend reading the below article to find attractive insight.

https://lnkd.in/gjCBhP_P

Usability ⓘ

10.00

License

CC BY-NC-SA 4.0

Expected update frequency

Annually

17.01.2023

Udemy Courses

Data Card Code (4) Discussion (1)

▲

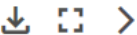
39

New Notebook

Download (643 MB)

Education Tabular Data Analytics

Comments.csv (1.61 GB)



Detail Compact Column

6 of 6 columns

id	course_id	# rate	date	display_name
comment ID	Course ID	Rate	Date	name
54127m	17694.91m	0.55	26Jun1123Oct22	
125535470	4913148	5.0	2022-10-07T11:17:41-07:00	Marlo
68767147	3178386	3.5	2020-10-19T06:35:37-07:00	Yamila Andrea

Data Explorer

Version 1 (1.69 GB)

- Comments.csv
- Course_info.csv

Датасеты библиотеки Scikit-Learn

Семь чистых датасетов

- Цены на жилье в Бостоне
- Диагностика рака молочной железы (Висконсин)
- Распознавание вина
- Растения Ирис
- Диабет
- Рукописные цифры
- Linnerud – набор данных регрессии

Dataset MNIST

- 70 тысяч картинок с цифрами от 0 до 9, приведенных к одинаковому виду.
- Формат CSV
- Размер 28×28 пикселей.
- Черный фон, на котором изображена белая цифра. Цифра помещена в середине, так, чтобы ее центр масс совпадал с центром изображения.
- Цифра меньше целой картинки — ее размер составляет 20×20 пикселей.
- Разделен на две части: train (60 тыс) и test (10 тыс)

label = 5



label = 0



label = 4



label = 1



label = 9



label = 2



label = 1



label = 3



label = 1



label = 4



label = 3



label = 5



label = 3



label = 6



label = 1

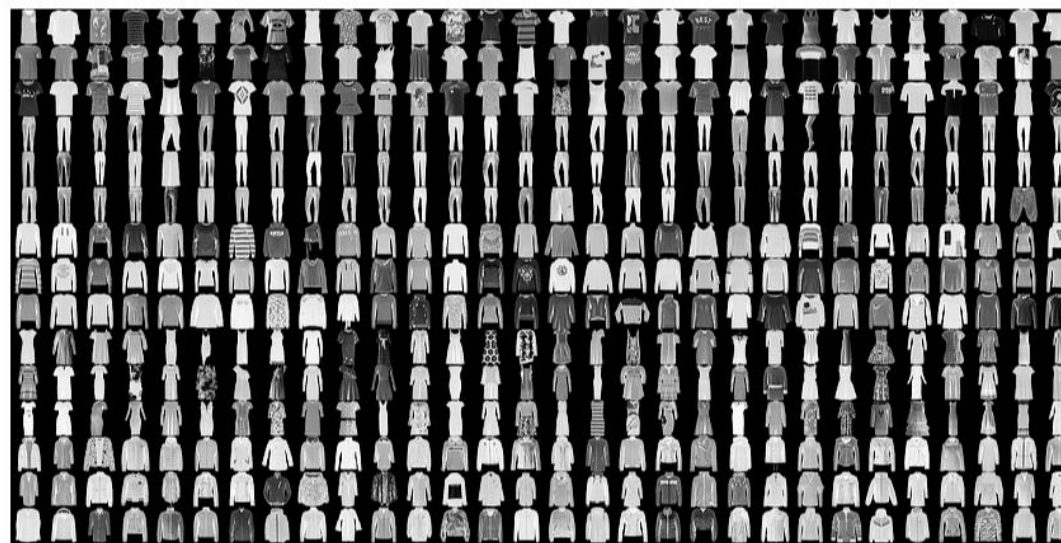


Датасет Fashion MNIST



- 70 тыс
- 28 на 28 пикселей

Метка	Класс
0	Футболка / топ
1	Шорты
2	Свитер
3	Платье
4	Плащ
5	Сандали
6	Рубашка
7	Кроссовки
8	Сумка
9	Ботинки



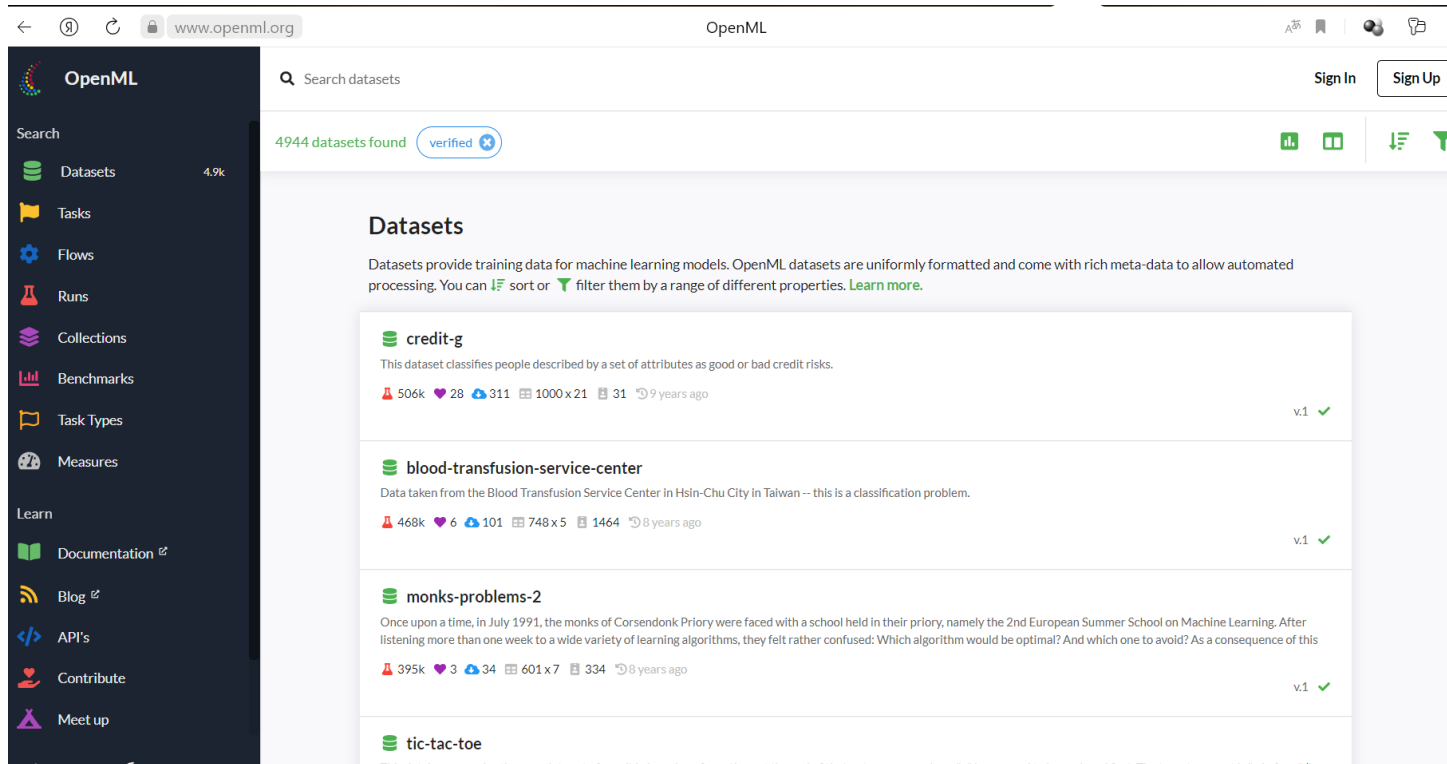
(с) Андрей Шмиг

СИНТАКСИС ИСПОЛЬЗОВАНИЯ ДАТАСЕТОВ

```
from sklearn import datasets  
data = datasets.load_breast_cancer()
```

```
# импорт pandas  
import pandas as pd  
# Считайте DataFrame, используя данные функции  
df = pd.DataFrame(data.data, columns=data.feature_names)  
# Добавьте столбец "target" и заполните его данными.  
df['target'] = data.target  
# Посмотрим первые пять строк  
df.head()
```


Использование публичных датасетов



The screenshot shows the OpenML website interface. The browser address bar displays 'www.openml.org'. The left sidebar contains navigation links: OpenML, Search, Datasets (4.9k), Tasks, Flows, Runs, Collections, Benchmarks, Task Types, Measures, Learn, Documentation, Blog, API's, Contribute, and Meet up. The main content area has a search bar with 'Search datasets' and buttons for 'Sign In' and 'Sign Up'. Below the search bar, it indicates '4944 datasets found' with a 'verified' badge. The 'Datasets' section explains that datasets provide training data for machine learning models and lists several datasets: 'credit-g', 'blood-transfusion-service-center', 'monks-problems-2', and 'tic-tac-toe'. Each dataset entry includes a description, statistics (e.g., 506k instances, 28 features), and a version indicator (v.1).

```
from sklearn.datasets import fetch_openml
```

```
X,y = fetch_openml("wine", version=1, as_frame=True, return_X_y=True)
```



ФОРМИРОВАНИЕ ДАТАСЕТА

Этапы формирования датасета



Этап планирования

- Формулировка задачи;
- Составление перечня признаков и/или характеристик исходных данных;
- Определение методики верификации значений выбранных признаков и/или характеристик у элементов формируемого набора данных;
- Определение источников данных;

Этап отбора исходных (сырых) данных

- выбор как можно большего спектра исследований интересующей модальности и процедуры;
- сохранение необходимого для решения задачи количества сопроводительной информации
- по возможности, обезличивание исследований «на месте», не покидая информационного контура учреждения, в котором происходит отбор данных.

Этап разметки и верификации

- **Разметка** — процесс определения значения признаков или характеристик у элемента данных в датасете.

Этап документирования

- Публикация датасета и сопроводительной документации.

Создание графического датасета

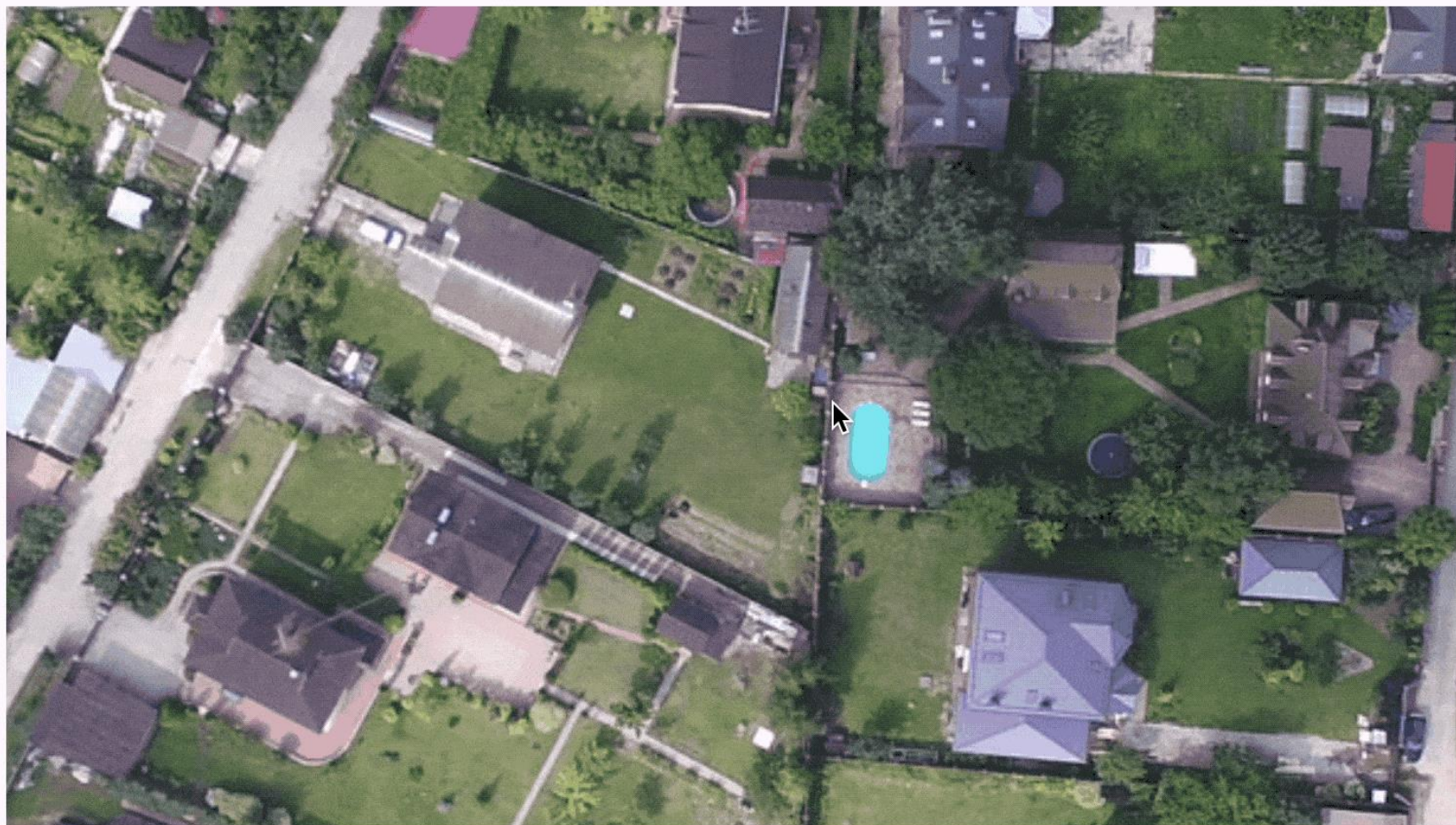
Сбор фотографий

- Качество, размер

Разметка

- Label Studio, ImageTagger, LabelImg, Lionbridge AI, TrainingData.io, Hive, Scale, Supervise.ly

Экспорт датасета



⊕

⊕

⊖

📄

🔄

Home 1

Pool 2

Fence 3

Driveway 4

Other 5


```
project-ID-at-YEAR-MONTH-DAY-HOUR-MINUTE-STRING.zip
```

```
notes.json
```

```
classes.txt
```

```
labels
```

```
    image_filename1.txt
```

```
    image_filename2.txt
```

```
    image_filename3.txt
```

```
    ...
```

```
images
```

```
    image_filename1.jpg
```

```
    image_filename2.jpg
```

```
    image_filename3.jpg
```

```
    ...
```

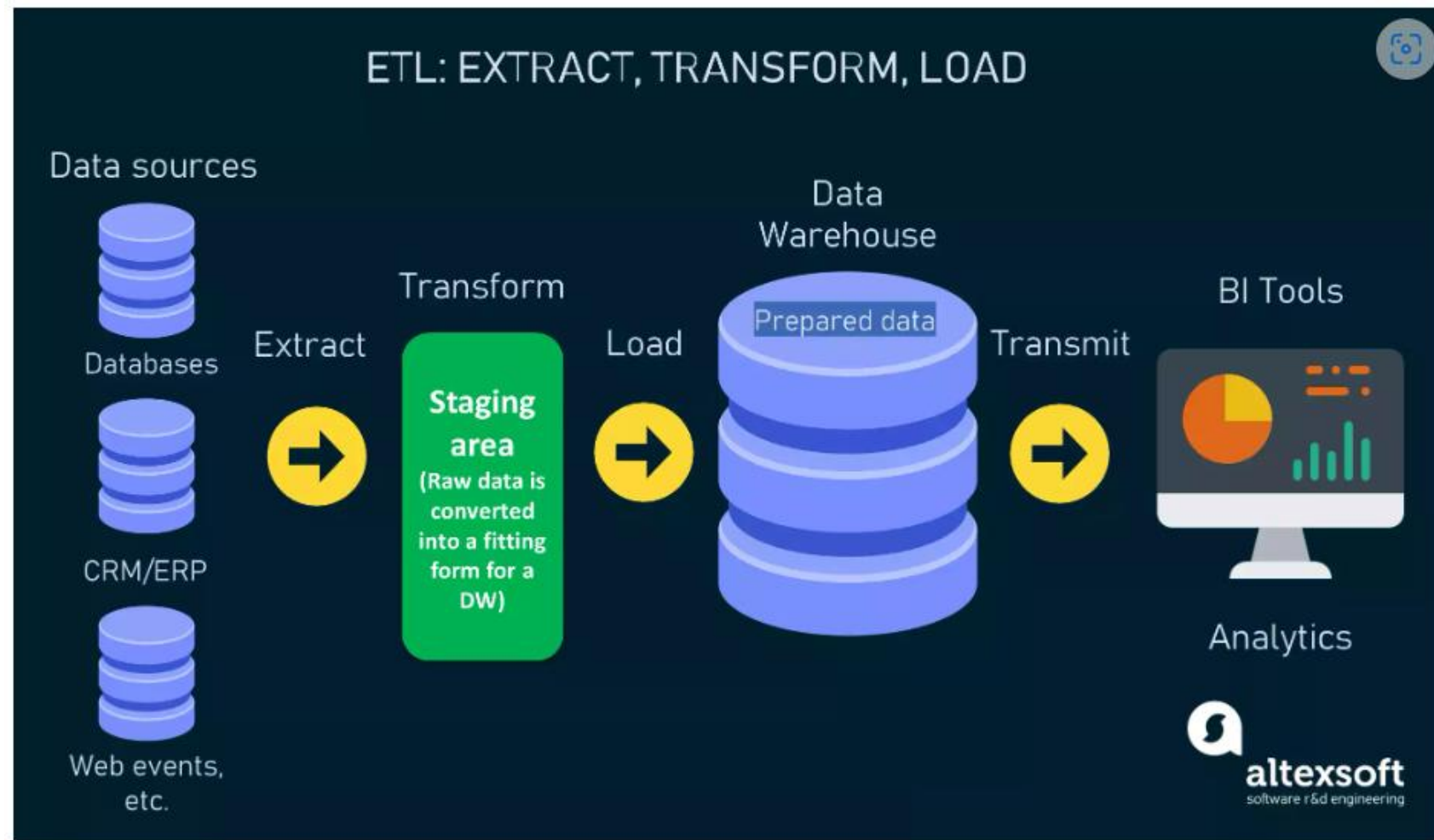
17.01.2023

МЕХАНИЗМЫ СБОРА ДАННЫХ

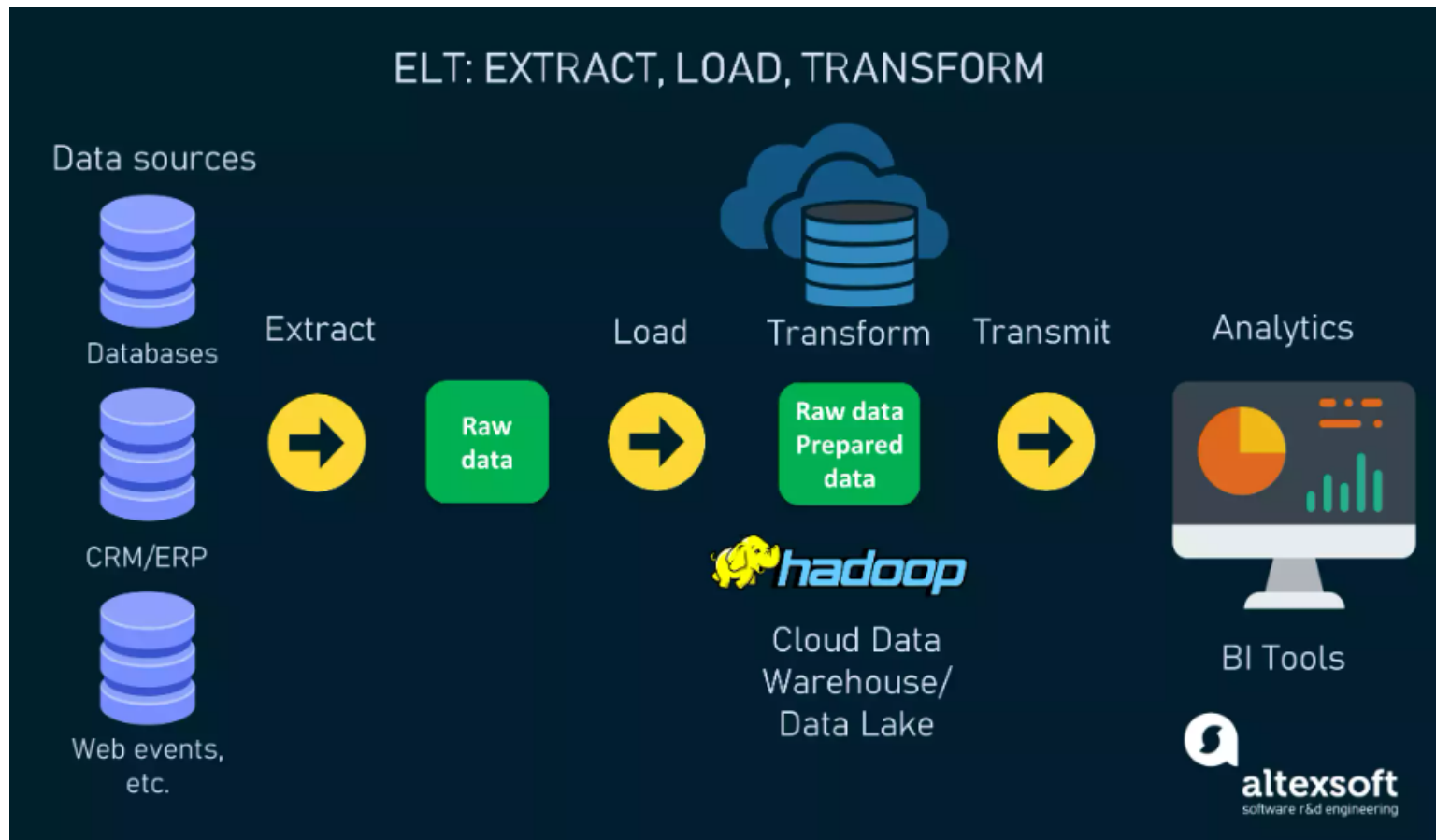
Хранилище Warehouse

- Для структурированных данных (SQL);
- Подход **ETL**

Extract,
Transform
Load



Data Lakes



- способны хранить и структурированные, и неструктурированные данные, в том числе изображения, видео, аудиозаписи, файлы PDF и т.д.
- Подход ELT (Extract, Load and Transform)

Data Fabric

- **Data Fabric** — это архитектура управления **данными**, позволяющая оптимизировать доступ к разрозненным **данным** и интеллектуально организовать и согласовать их для доставки потребителям в режиме самообслуживания.
- Data Lake - концепция централизации данных, Data Fabric - концепция распределенной работы с данными.

