

Основы Искусственного интеллекта

Основы сбора и подготовки данных.

Этапы построения систем Машинного обучения

- Сбор данных
- Анализ/Чистка данных
- Обработка данных
- Построения модели
- Тестирование модели
- Развертывание модели



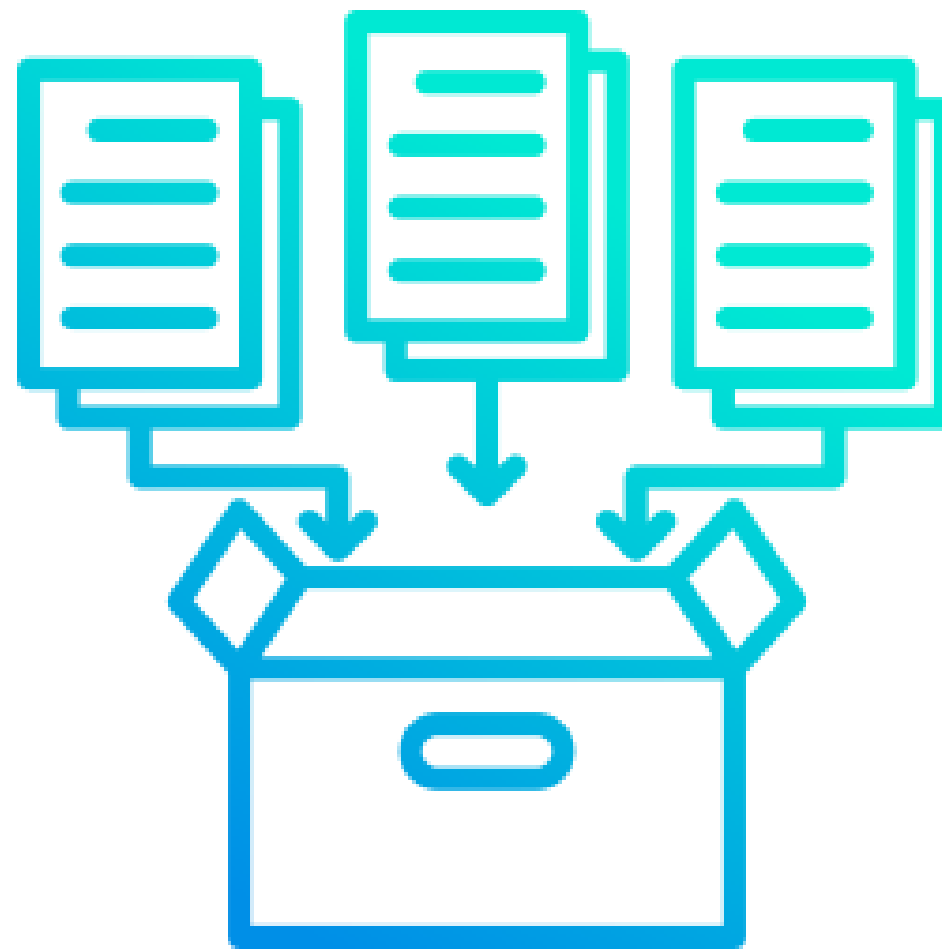
Ключевые компоненты для создания моделей ИИ



Сбор данных

Определение

- **Сбор данных в области** искусственного интеллекта (ИИ) и машинного обучения (МО) означает сбор информации из различных источников для обучения, проверки или тестирования моделей ИИ.



Типы сбора данных для ИИ



Цели применения



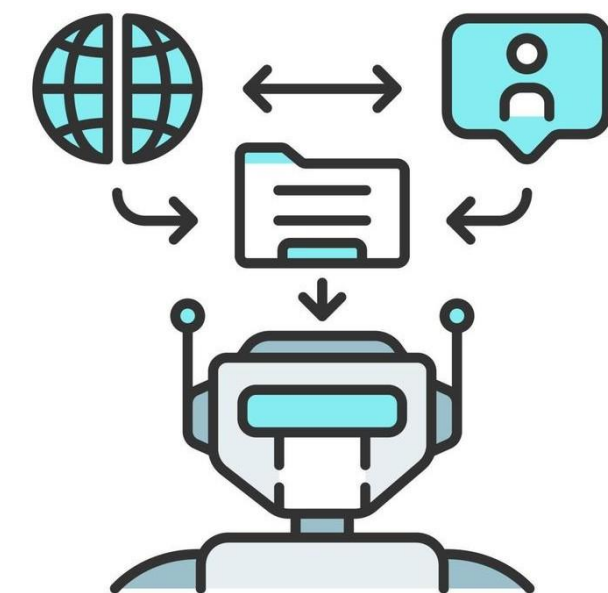
Дополнительные типы данных

- **Облако данных 3D точек.** Чтобы эффективно обучать беспилотные автомобили, системам искусственного интеллекта необходимы данные LiDAR для точного обнаружения с помощью лидарных датчиков с 3D-боксами. Производительность лидарных датчиков в автономных системах можно повысить только за счет 3D-данных высокого разрешения и высочайшего качества.
- **Dataways: новый вид сбора данных для AI.** Передовой поставщик услуг по сбору данных ИИ. Dataways собирает и анализирует контролируемые наборы данных для обучения и тестирования алгоритмов машинного обучения. Имея большой опыт сбора данных и специальную команду экспертов, которые могут собирать данные из всех регионов земного шара.

Сбор данных в ИИ и машинном обучении

Цель

- Основная цель сбора данных в области искусственного интеллекта и машинного обучения заключается в сборе всесторонних и репрезентативных наборов данных, охватывающих реальные сценарии.
- Эти наборы данных служат основополагающими блоками, на основе которых разработчики и исследователи могут обучать алгоритмы распознаванию закономерностей, делать прогнозы или выполнять другие когнитивные задачи.



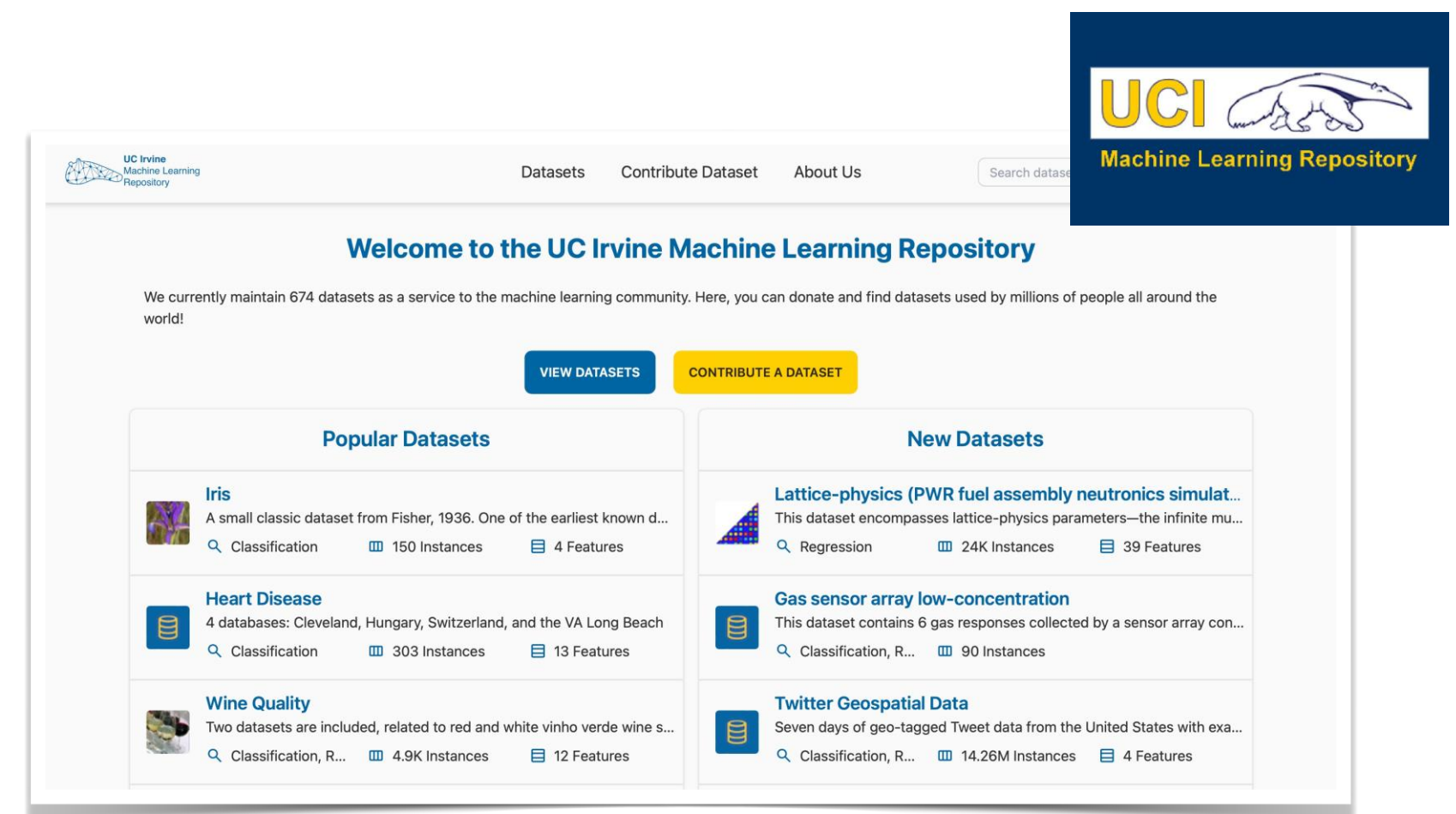
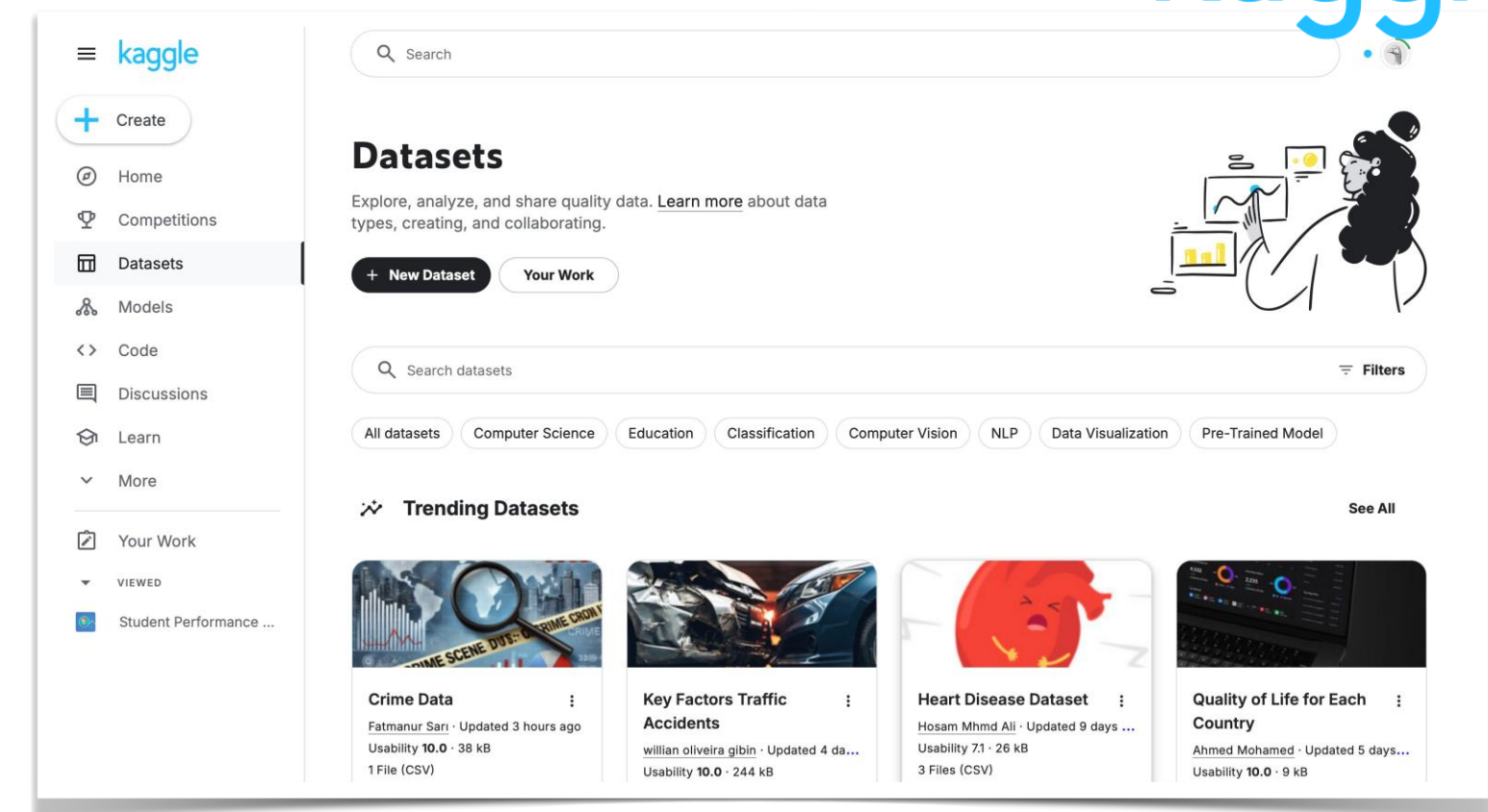
Методы сбора данных

- **Выполнение веб-сканирования (web-scraping):** веб-сканирование включает автоматическое извлечение данных с веб-сайтов. Этот метод использует ботов или сканеров для навигации по веб-страницам и извлечения релевантной информации для анализа.
- **Использование опросов и обратной связи пользователей:** Опросы и тесты пользователей собирают целевую информацию, задавая конкретные запросы отдельным лицам или группам. Этот метод помогает собирать качественные данные, которые являются субъективными и не состоящими из цифр.
- **Сбор данных с датчиков:** В приложениях Интернета вещей (IoT) датчики собирают данные в реальном времени с различных устройств или систем. Эти датчики захватывают экологические, поведенческие или операционные данные.
- **Использование общедоступных наборов данных:** Многие организации предоставляют широкий спектр общедоступных наборов данных по областям для исследовательских целей.

Общедоступные наборы данных

kaggle

- kaggle.com
- Необходимо пройти регистрацию (“Continue with Google”)
- Репозиторий машинного обучения Калифорнийского университета в Ирвине
- <https://archive.ics.uci.edu/>

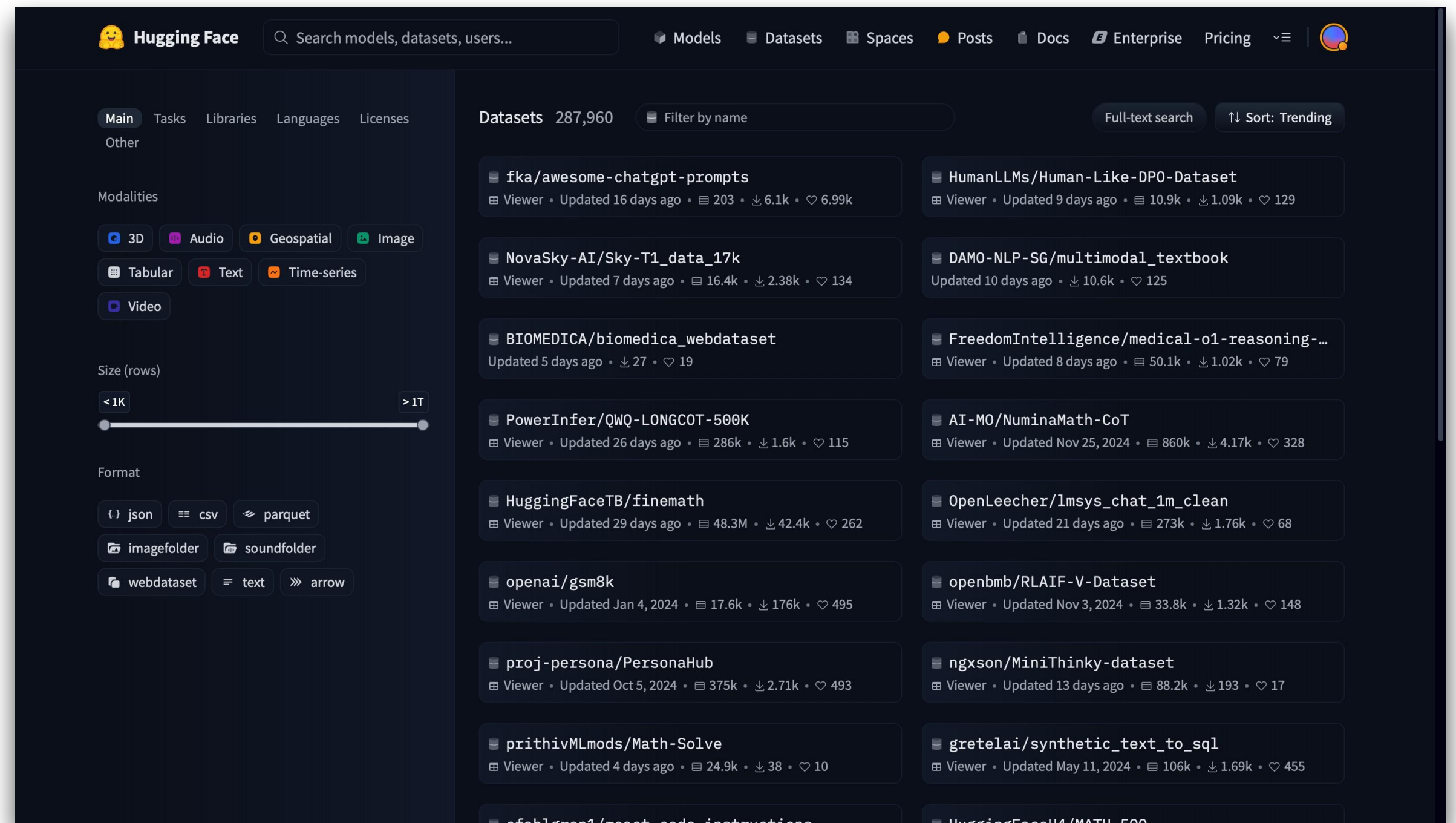


Общедоступные наборы данных



Hugging Face

- huggingface.co - пройти регистрацию, вход на платформу
- <https://huggingface.co/datasets> - локация наборов данных



Чистка и подготовка данных

- Обработка пропущенных значений
- Стандартизация форматов данных
- Нормализация числовых значений
- Кодирование категориальных переменных

Инструменты для анализа данных

- Google Collab: <https://colab.research.google.com/>



- Google Sheets: <https://docs.google.com/spreadsheets>



Google Sheets

Набор данных для использования на предстоящие дни

- Набор данных - Титаник



- <https://www.kaggle.com/c/titanic/data>
- Необходимо загрузить набор в Google Collab

Чистка и подготовка данных

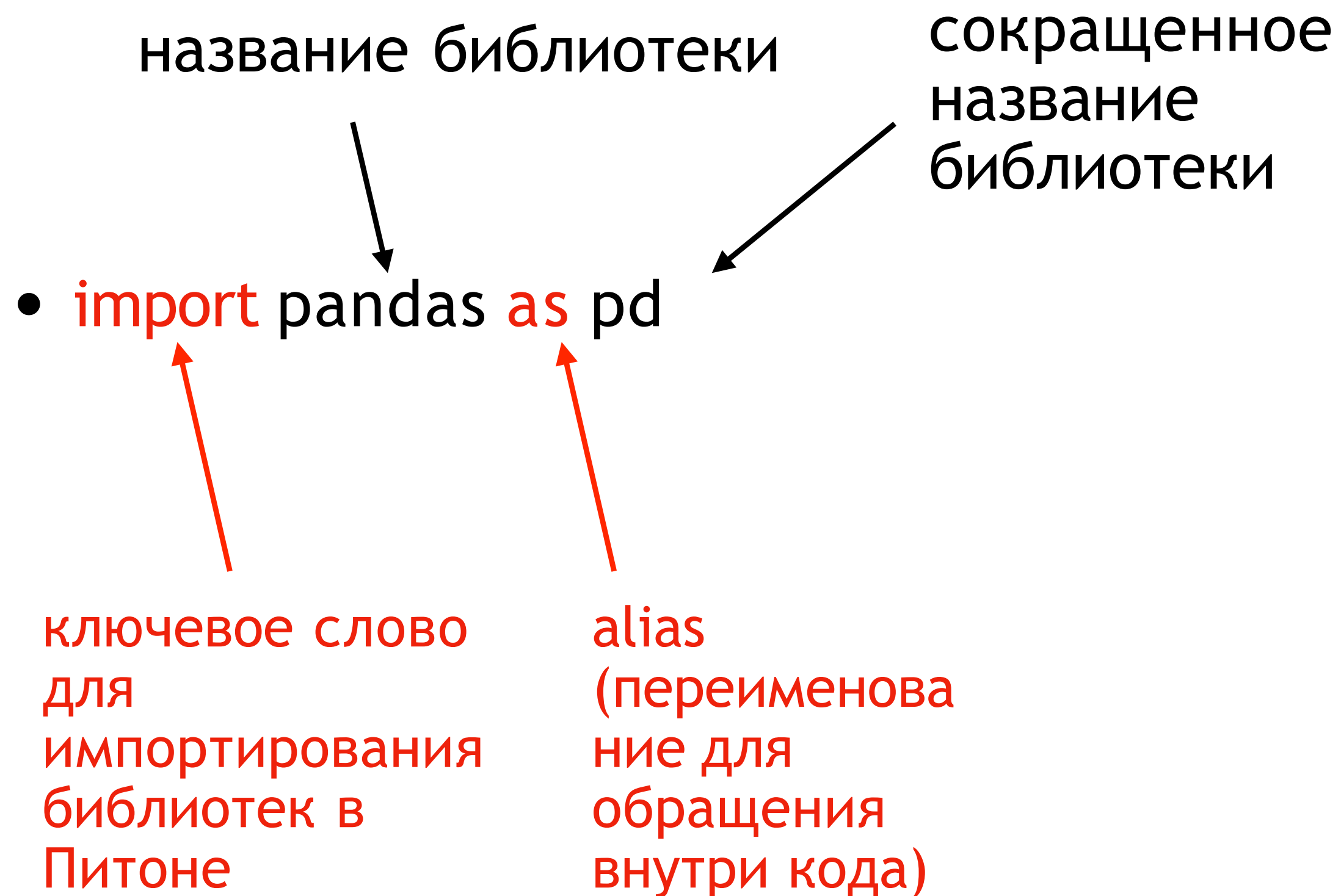
- Обработка пропущенных значений
- Стандартизация форматов данных
- Нормализация числовых значений
- Кодирование категориальных переменных

Библиотеки для работы с данными

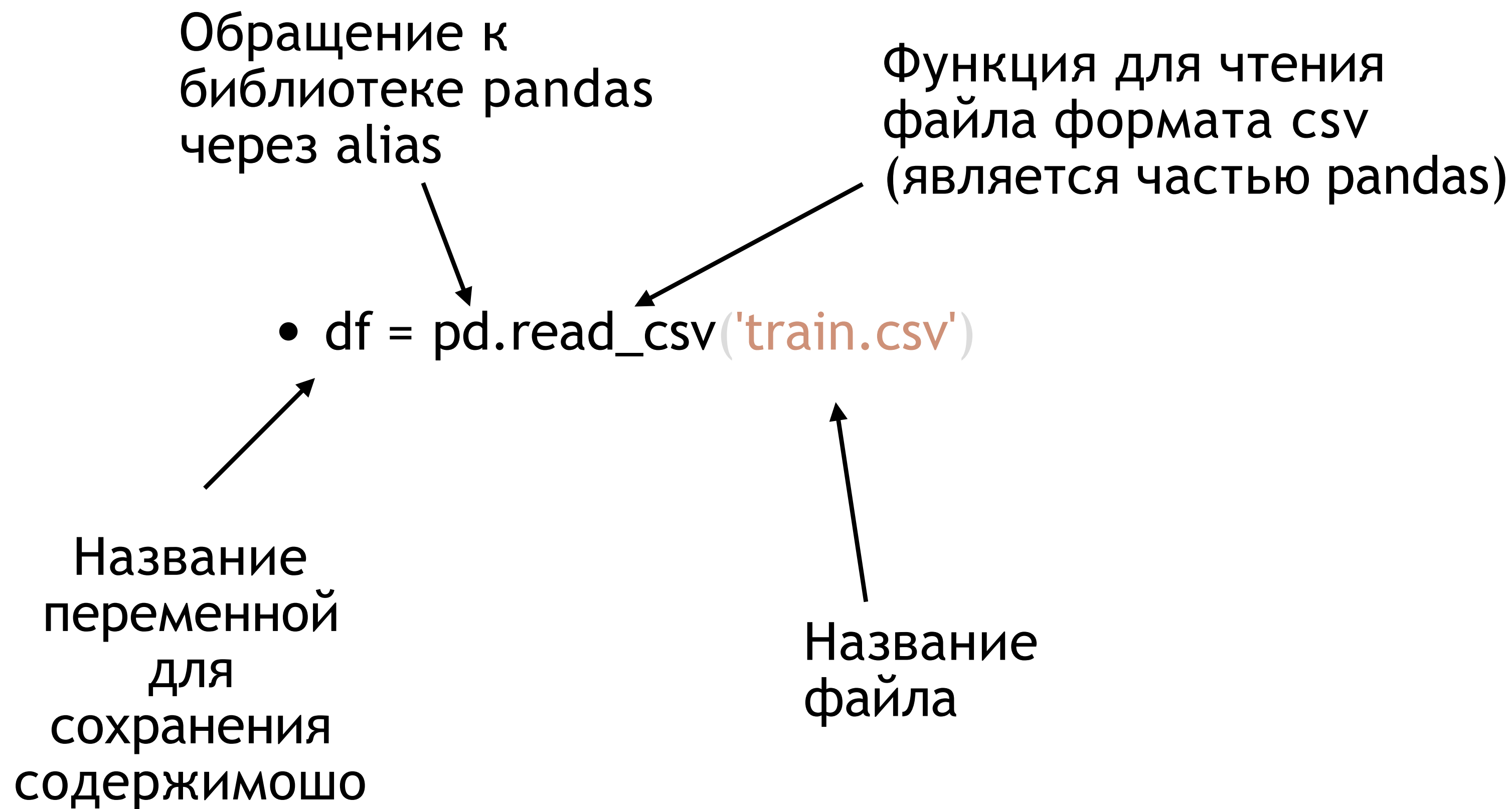
- **Pandas** — это программная библиотека, написанная для языка программирования Python для манипулирования и анализа данных.
- В частности, он предлагает структуры данных и операции для управления числовыми таблицами и временными рядами.



Импортирование библиотеки



Считывание csv файла



Описание набора данных Титаник

Переменная	Определения	Значения
Выжил	Признак выживания	0 = Нет, 1 = Да
Класс	Класс билета	1 = Первый, 2 = Второй, 3 = Третий
Гендер	Гендерная принадлежность	
Возраст	Возраст в годах	
# братьев и сестер / супругов на борту	# братьев и сестер / супругов на борту	
# родителей/детей на борту	# родителей/детей на борту	
Номер билета	Номер билета	
Стоимость проезда пассажира	Плата за билет	
Номер каюты	Номер каюты	
Порт отправления	Порт отправления	C = Cherbourg, Q = Queenstown, S = Southampton

Описание набора данных Титаник

Примечания по переменным

- **Класс:** показатель социально-экономического статуса (SES).
 - 1-й = Верхний
 - 2-й = Средний
 - 3-й = Нижний
- **Возраст:** возраст является дробным, если меньше 1. Если возраст предварительный, он имеет форму xx.5.
- **# братьев и сестер / супругов на борту:** Набор данных определяет семейные отношения таким образом:
 - Братья и Сестра = брат, сестра, сводный брат, сводная сестра
 - Супруга = муж, жена
- **# родителей/детей на борту:** Набор данных определяет семейные отношения таким образом:
 - Родитель = мать, отец
 - Ребенок = дочь, сын, падчерица, пасынок
 - Некоторые дети путешествовали только с няней, поэтому для них значение=0.

Чистка набора данных Титаник с помощью Pandas

- **Цель:** понять, как очистить и предварительно обработать набор данных «Титаник» с помощью pandas.
- **Важность.** Чистка данных — важнейший этап любого анализа данных или в подготовке данных для машинного обучения. Набор данных «Титаник» представляет собой реальный пример обработки данных.

Просмотр данных

По шапке

- `df.head()`
- `df.head(5)`
- `df.head(10)`

Просмотр данных

По концу

- `df.tail()`
- `df.tail(5)`
- `df.tail(10)`

Описание данных

- `df.describe()`
- `df.info()`

Загрузка набора данных Титаник без скачивания файла

```
import pandas as pd

# Загрузка данных

url = 'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv'

df = pd.read_csv(url)

# Показать первые строки

print(df.head())
```

Обработка недостающих данных

- **Шаг 1.** Определить недостающие значения.

```
df.isnull().sum()
```

- **Шаг 2.** Определить стратегию обработки пропущенных значений:
 - Удалить строки/столбцы.
 - Заполните недостающие значения средним значением, медианой, модой или заполнителем.

Обработка недостающих данных

- # Заполните недостающее значение Возраст (Age) медианным значением.

```
df['Age'] = df['Age'].fillna(df['Age'].median())
```

- # Заполните отсутствующее слово Порт отправления (Embarked) с помощью режима

```
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])
```

- # Удалить Номер каюты (Cabin) из-за слишком большого количества пропущенных значений.

```
df = df.drop(columns=['Cabin'])
```


Преобразование типов данных

- При необходимости преобразуйте категориальные столбцы в числовые форматы.
- Пример: закодируйте Гендер (`Sex`) и Порт отправления (`Embarked`) для целей машинного обучения.

Преобразование типов данных

- # Преобразовать Гендер (Sex) в числовое значение (0 для мужчин (male), 1 для женщин (female))

```
df['Sex'] = df['Sex'].map({'male': 0, 'female': 1})
```

- # Горячее кодирование Порт отправления (Embarked)

```
df = pd.get_dummies(df, columns=['Embarked'])
```

Embarked_C	Embarked_Q	Embarked_S
False	False	True
True	False	False
False	False	True
False	False	True
False	False	True

Sex	
0	0
1	1
2	1
3	1
4	0
...	...
886	0
887	1
888	1
889	0
890	0

Горячее кодирование

Пример

					
Car_model			Car_model_BMW	Car_model_Lexus	Car_model_Toyota
0	Toyota		0.0	0.0	1.0
1	Lexus		0.0	1.0	0.0
2	BMW		1.0	0.0	0.0
3	Toyota		0.0	0.0	1.0

Спасибо за внимание!