

Основы работы с данными: сбор, анализ, визуализация

Лекция 4.

Максим Карпов

 @buntar29
 mekarpov@hse.ru



HSE
University

Основы обработки данных

3. Для чего нужна статистика?

- **Изучать численное описание явлений, происходящих в обществе, науке и технике.**
- **Прорабатывать общие вопросы** сбора, измерения, мониторинга и анализа массовых статистических данных.
- **Находить специальную методологию исследования и обработки материалов.**
- **Познавать мир.**

Придет время, и статистическое мышление станет таким же необходимым качеством для истинного гражданина, как умение читать и писать.

Герберт Уэллс (1911)

3. Генеральная совокупность (Population)

Набор из всех объектов, которые интересны в условиях данной задачи.

Например: все люди, живущие в Москве; все автомобили, собранные на этом заводе.

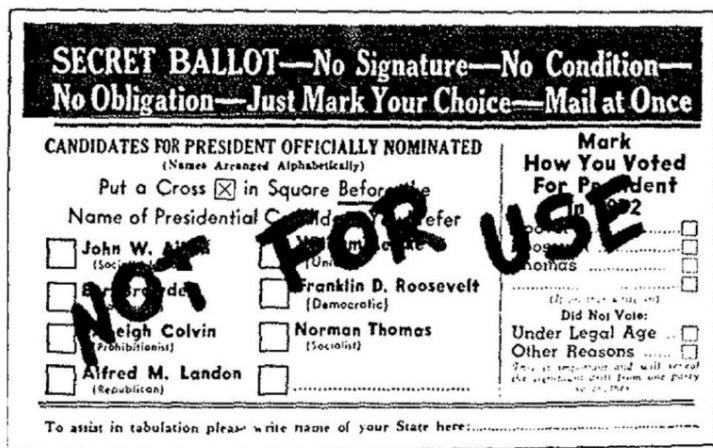
Как правило, вся генеральная совокупность недоступна исследователям.

3. Выборка (Sample)

Часть генеральной совокупности элементов, которая была отобрана для эксперимента. Выборка может быть представительной или непредставительной. Отбор выборки – важная часть любой исследовательской работы. Наиболее простой способ – случайная выборка.

3. Репрезентативность ваших данных

Репрезентативность - очень большая проблема для любого исследования. Вы всегда должны понимать, способна ли анализируемая вами выборка ответить на ваши вопросы.

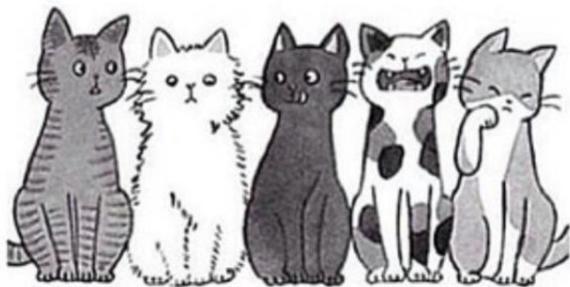


«Литературный дайджест» провел пост-опрос 1916 года, чтобы предсказать результаты выборов президента США. И это было успешно в 1916, 1920, 1924, 1928 и 1932 годах.

Генеральная
совокупность



Выборка



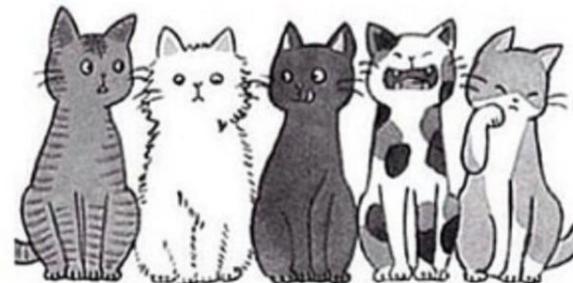
Наблюдение
(измерение)



3. Что делает статистика?

Статистика по случайной выборке судит о генеральной совокупности! Мы поймали 4 серых кошки и 1 черную. Какие выводы можно сделать о цвете кошек вообще?

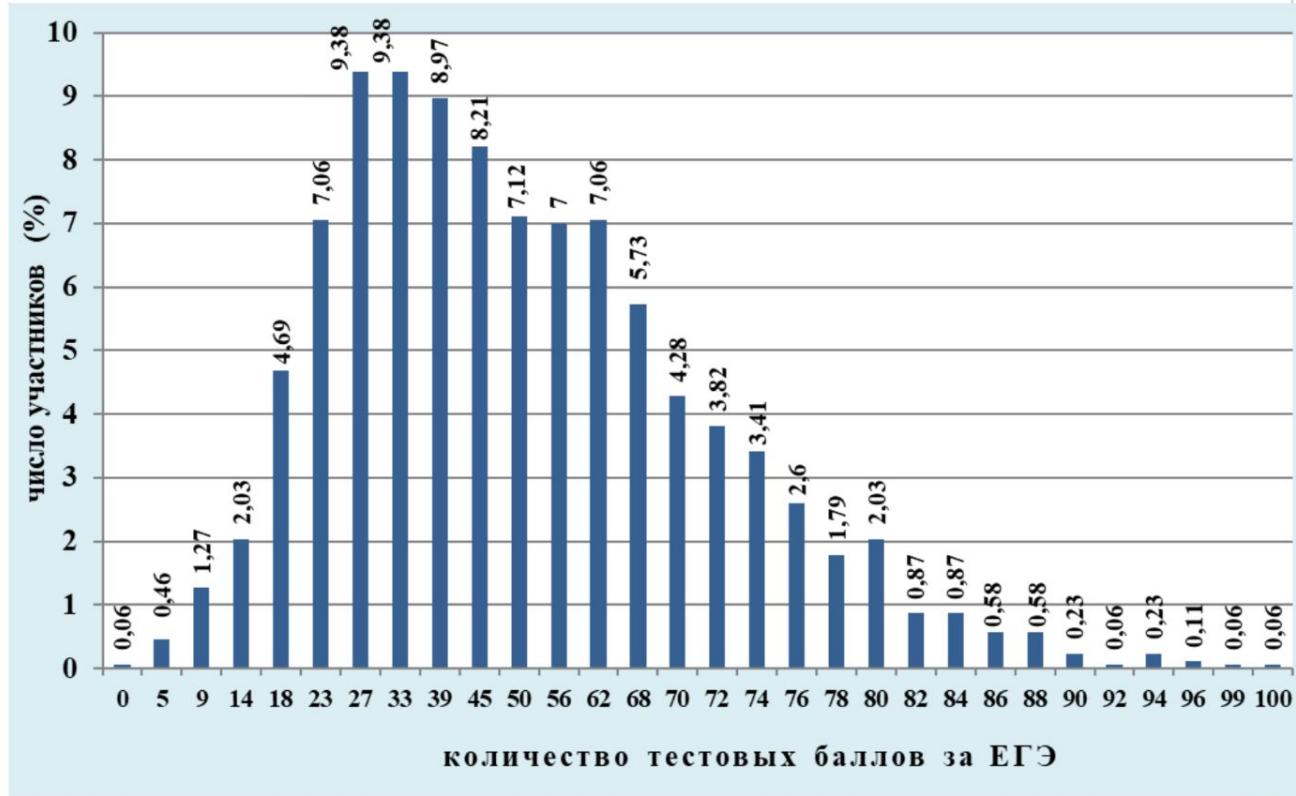
Получаем представления о значениях переменных в популяции, измеряя эти переменные (цвет кошек) в выборке



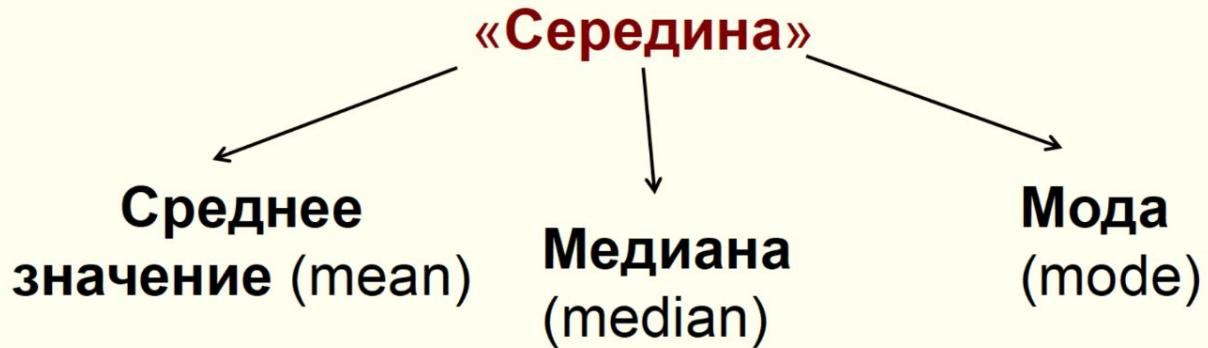
Гистограмма распределения

Чтобы получить полную информацию об исследуемых данных, необходимо получить закон распределения этих данных. Сейчас мы не будем давать определение понятия *распределение*. Но обсудим, что такое *гистограмма распределения*.

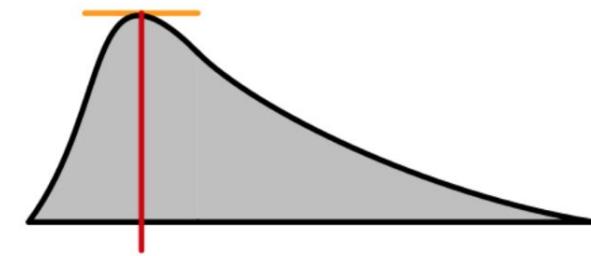
Гистограмма распределения (пример)



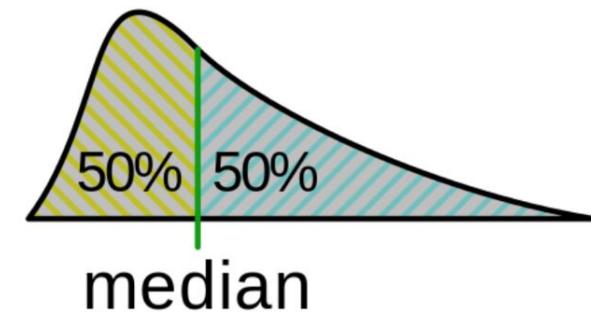
«Середина» распределения (central tendency)



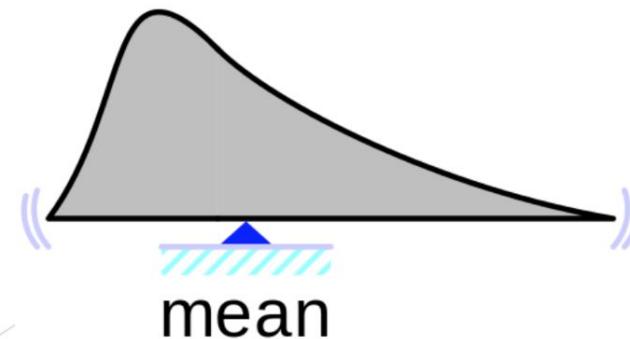
Мода (mode) -наиболее часто встречающееся значение, локальный максимум.



Медиана(median)-значение, которое делит распределение пополам(его площадь в т.ч.): половина значений больше медианы, половина -не меньше.

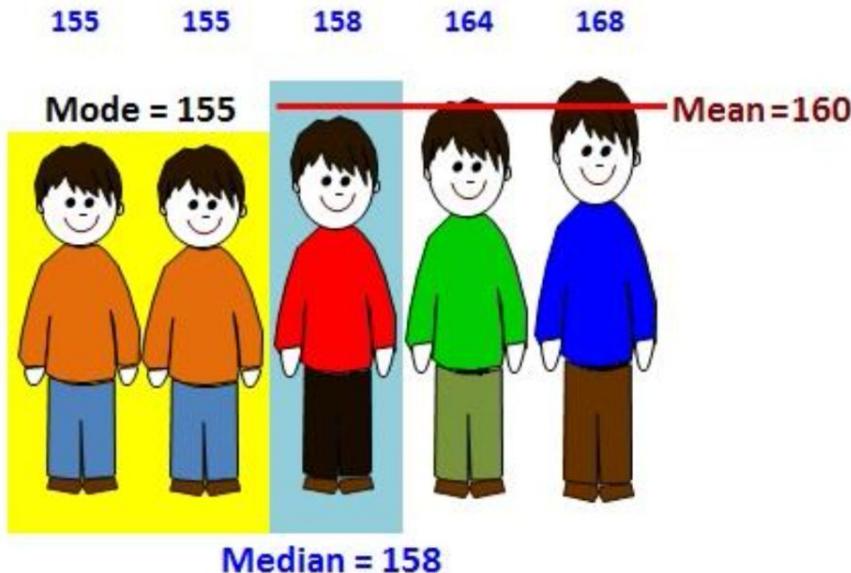


Среднее значение (mean) -сумма всех значений переменной, делённая на количество значений.



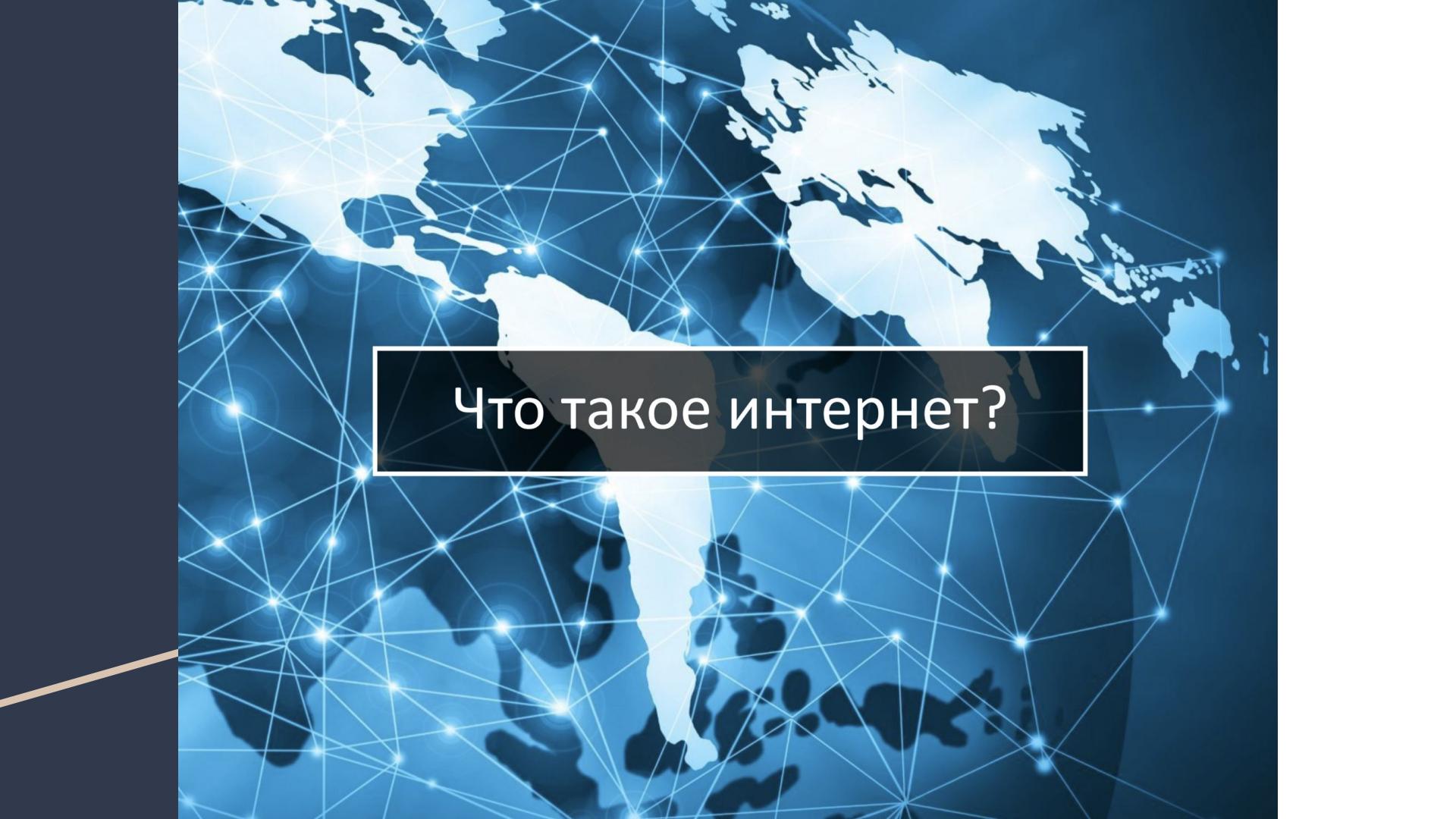
4. Среднее значение, медиана и мода

Среднее значение, медиана и мода - это различные способы измерить “центральное” число в наборе данных. Каждый из этих способов пытается выразить информацию о данных с помощью одного “усредненного” числа.



Типы данных

- Типы переменных/признаков
 - Качественные/числовые
 - . Непрерывные (масса человека, площадь дома)
 - . Дискретные (возраст в годах)
 - Качественные/категориальные
 - . Бинарные (пол: 1 – женский, 0 – мужской)
 - . Номинальные, неупорядоченные (категория товара)
 - . Порядковые, упорядоченные (степень согласия с утверждением)



Что такое интернет?

The image shows a bundle of various colored electrical wires and cables. Some are sheathed in thick, smooth plastic, while others are exposed, showing their individual copper or aluminum strands. The colors include red, blue, yellow, green, white, and black. A large, semi-transparent rectangular box with a black border is positioned in the center-left area of the image, containing the text.

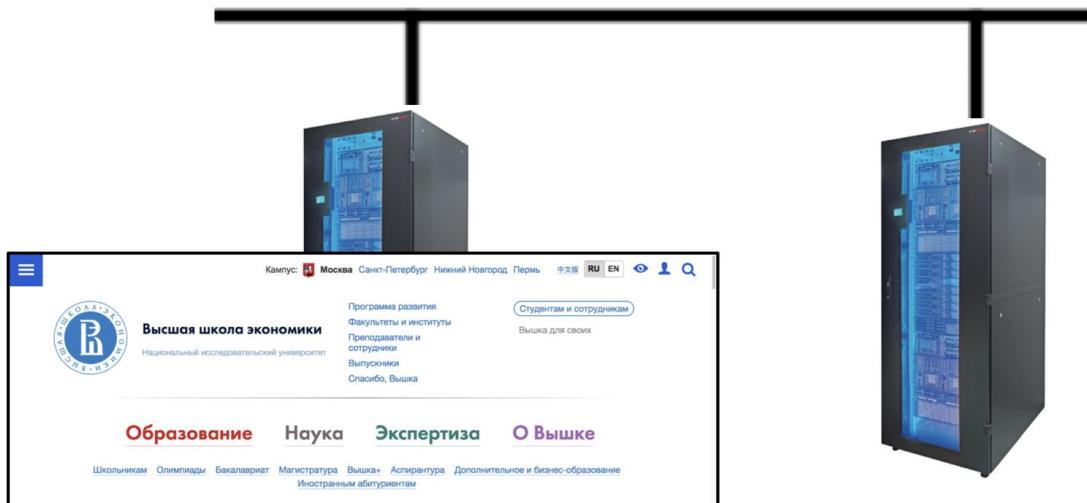
Это провода

Компьютеры, соединенные
проводами, могут общаться



Сервер

- Компьютер, подключенный к интернету напрямую
- Все вебсайты живут на серверах



IP адрес

- Есть у каждого узла интернета (у каждого сервера)
- По ним компьютеры находят друг друга
 - 172.217.18.14
 - 212.8.235.7

Откуда взялся интернет?

1972

1970

197

- 1957 — СССР запустил Спутник
- 1958 — создана ARPA (Advanced Research Projects Agency)
- 1960 — Joseph Licklider (1960). "Man-Computer Symbiosis". IRE Transactions on Human Factors in Electronics. HFE-1: 4–11.
- 1962 — Licklider стал руководителем в ARPA
- 1969 — запуск ARPANET; первая передача информации между удаленными компьютерами (слово "LO")
- 1971 — первая программа для электронной почты



Откуда взялся интернет?

1972

- 1973 — по дну Атлантики прикладывают интернет-кабель; к сети подключаются Великобритания и Норвегия
- 1983 — разработана система доменных имен (DNS)
- 1984 — сеть NSFNet (National Science Foundation Network); станет каркасом современного интернета
- 1989 — Тимоти Бернерс-Ли высказывает идею «Всемирной паутины»;
- 1992-93 — он же создает протокол HTTP и язык HTML
- 1993 — первый браузер Mosaic

SATELLITE CIRCUIT
□ IMP
○ TIP
△ PLURIBUS IMP
NOTE: THIS MAP DOES NOT SHOW AREA'S EXPERIMENTAL
SATELLITE CONNECTIONS!
NAMES SHOWN ARE IMP NAMES, NOT NECESSARILY HOST NAMES

197

Веб-скрейпинг



Устройство интернет-сайтов

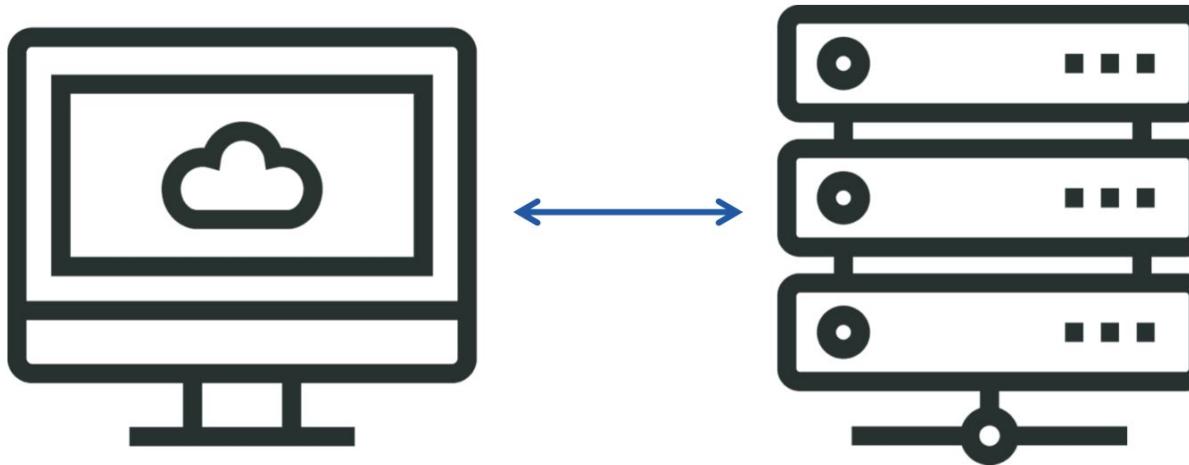
Введение в HTML

Структура кода страницы

Основные теги



Устройство интернет-сайтов

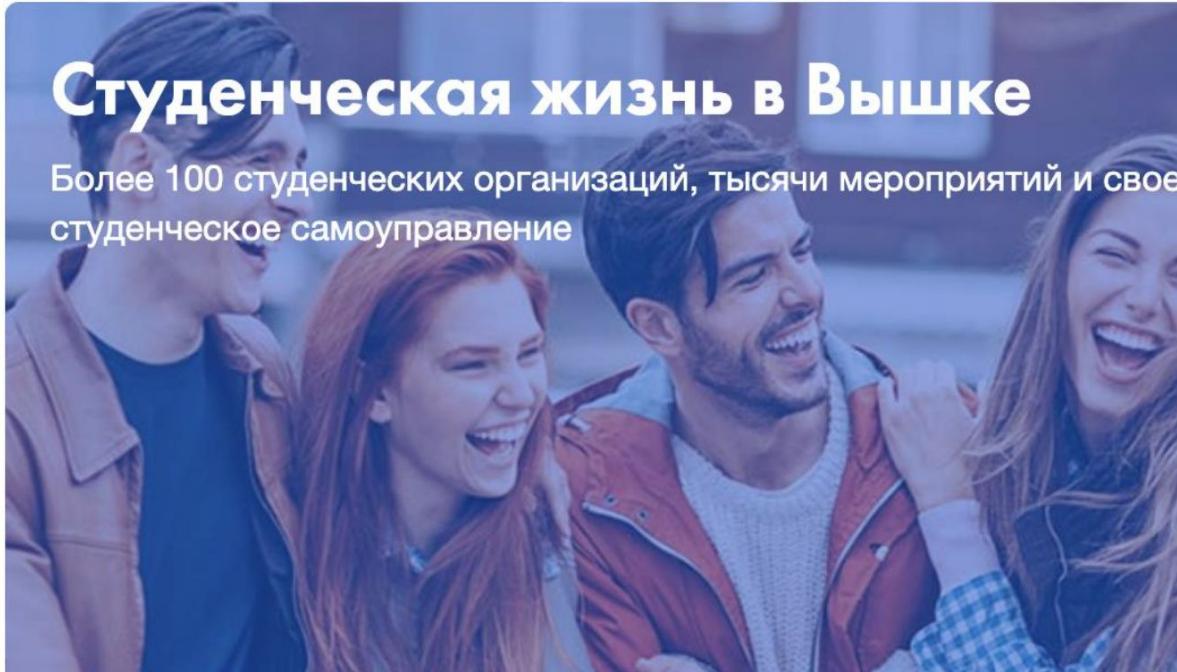


Устройство интернет-сайтов

Вышка 360°

Студенческая жизнь в Вышке

Более 100 студенческих организаций, тысячи мероприятий и свое студенческое самоуправление



Устройство интернет-сайтов



```
class="section">

## Вышка 360°



Студенческая жизнь в ВышкеБолее 100 студенческих организаций, тысячи мероприятий и свое студенческое самоуправлениеСпорт и активный отдых


```



Устройство интернет-сайтов

Введение в HTML

Структура кода страницы

Основные теги



Введение в HTML

Теги

<имя_тега></имя_тега>

Например:

Этот текст написан жирным

Этот текст написан **жирным**



Введение в HTML

Атрибуты

<тег атрибут="значение"></тег>

Например:

При нажатии откроется сайт
Вышки

При нажатии откроется сайт Вышки



Введение в HTML

Одиночные теги

<тег>

Например:

Первая строка
Вторая строка

Первая строка
Вторая строка



Введение в HTML

Вложенные теги

```
<тег1><тег2></тег2></тег1>
```

Например:

Написано ***жирным курсивом***!

Написано ***жирным курсивом!***



Устройство интернет-сайтов

Введение в HTML

Структура кода страницы

Основные теги



Структура страницы

```
<!DOCTYPE html>
<html>

    <head>
        <title>Название страницы</title>
    </head>

    <body>
        <h1>Заголовок</h1>
        <p>Параграф текста.</p>
    </body>

</html>
```



Структура страницы

```
<!DOCTYPE html>
<html>
  <head>
    <title>Название страницы</title>
  </head>

  <body>
    <h1>Заголовок</h1>
    <p>Параграф текста.</p>
  </body>

</html>
```

Заголовок

Параграф текста.



Устройство интернет-сайтов

Введение в HTML

Структура кода страницы

Основные теги



Основные теги

`<h1></h1>, <h2></h2>, ... <h6></h6>...` — заголовки

`` — создание гиперссылки

`` — вставка картинки

`<p></p>` — выделяет абзац текста

`` — жирное начертание шрифта

`<i></i>` — курсивное начертание шрифта

`
` — перевод строки на новую строчку, аналог `\n`



Запрос страницы из Сети

**Атрибуты страницы:
код ответа, кодировка**

**Извлечение информации
из страницы**



Запрос страницы из Сети

Чтобы работать с информацией из интернета, нам потребуются два модуля:

- 1 Модуль для запроса страниц из Сети;
- 2 Модуль для извлечения информации из страниц.

```
import requests  
from bs4 import BeautifulSoup
```



Запрос страницы из Сети

Запросим файл страницы <https://online.hse.ru/python-as-foreign/1/> из Сети:



Запрос страницы из Сети

Запросим файл страницы <https://online.hse.ru/python-as-foreign/1/> из Сети:

- 1 Подключим модуль requests;
- 2 Запросим страницу;
- 3 Проверим, что она загрузилась, посмотрев на статус загрузки;
- 4 Установим правильную кодировку;
- 5 Посмотрим на html-код страницы.



Запрос страницы из Сети

Запросим файл страницы <https://online.hse.ru/python-as-foreign/1/> из Сети:

```
import requests  
url = 'https://online.hse.ru/python-as-  
foreign/1/'  
page = requests.get(url)
```



Запрос страницы из Сети

**Атрибуты страницы:
код ответа, кодировка**

**Извлечение информации
из страницы**



Атрибуты страницы: код ответа, кодировка

Посмотрим на содержимое файла страницы

<https://online.hse.ru/python-as-foreign/1/> из Сети:

```
print(f'Страница загрузилась с кодом  
{page.status_code}')  
# Страница загрузилась с кодом 200
```

Если бы мы запросили несуществующую страницу, код ответа был бы другой:

```
page2 = requests.get(  
'https://online.hse.ru/python-as-  
foreign/8/')  
print(page2.status_code)  
# 404
```



Атрибуты страницы: код ответа, кодировка

Посмотрим на содержимое файла страницы

<https://online.hse.ru/python-as-foreign/1/> из Сети:

```
print(f'Страница загрузилась с кодом  
{page.status_code}')  
# Страница загрузилась с кодом 200  
  
print(page.text)
```

Компьютер выведет на экран:

```
<html>  
<head>  
    <meta charset="UTF-8">  
    <title>ДД³Д»Д°Д²Д»ДµД½Д, Дµ</title>
```



Атрибуты страницы: код ответа, кодировка

Посмотрим на содержимое файла страницы

<https://online.hse.ru/python-as-foreign/1/> из Сети, установив правильную кодировку:

```
page.encoding = 'utf-8'  
print(page.text)
```

Компьютер выведет на экран:

```
<html>  
<head>  
    <meta charset="UTF-8">  
    <title>Оглавление</title>
```



Атрибуты страницы: код ответа, кодировка

Чаще всего предыдущий способ сработает, но иногда из-за особенностей страниц кодировка все равно не изменится. Тогда, скорее всего поможет:

```
page3 = requests.get(url)
page3_text = page3.content.decode('utf-8')
print(page3_text)
```

Компьютер выведет на экран:

```
<html>
<head>
    <meta charset="UTF-8">
    <title>Оглавление</title>
```



Запрос страницы из Сети

**Атрибуты страницы:
код ответа, кодировка**

**Извлечение информации
из страницы**



Извлечение информации из страницы

Посмотрим на нашу страницу:

```
<html>
<head>
    <meta charset="UTF-8">
    <title>Оглавление</title>
</head>
<body>
    Статьи о персонажах: <a href="1.html">Гарри Поттер</a>,
    <a href="2.html">Джинни Уизли</a>,
    <a href="3.html">Лили Поттер</a>,
    <a href="4.html">Гермиона Грейндженер</a>,
    <a href="5.html">Сириус Блэк</a>
```



Извлечение информации из страницы

Распознаем код страницы с помощью BeautifulSoup:

```
from bs4 import BeautifulSoup  
soup = BeautifulSoup(page.text)
```

Посмотрим на нашу страницу ещё раз:

```
<html>  
<head>  
    <meta charset="UTF-8">  
    <title>Оглавление</title>  
</head>  
<body>  
    Статьи о персонажах: <a  
    href="1.html">Гарри Поттер</a>,  
    <a href="2.html">Джинни Уизли</a>
```



Извлечение информации из страницы

Найдём какую-нибудь ссылку:

```
link = soup.find('a')
print(link)
# <a href="1.html">Гарри Поттер</a>

print(link.get('href'))
# 1.html

print(link.text)
# Гарри Поттер
```



Извлечение информации из страницы

Найдём все ссылки:

```
print(soup.find_all('a'))  
# [<a href="1.html">Гарри Поттер</a>, <a  
href="2.html">Джинни Уизли</a>, <a  
href="3.html">Лили Поттер</a>, <a  
href="4.html">Гермиона Грейнджен</a>, <a  
href="5.html">Сириус Блэк</a>, <a  
href="6.html">Рубеус Хагрид</a>,  
<ещё 5 ссылок>  
<a  
href="https://harrypotter.fandom.com">Гарри  
Поттер вики</a>]
```



Извлечение информации из страницы

Найдём все ссылки и просмотрим их список с помощью цикла **for**:

```
for link in soup.find_all('a'):
    print(link)
```

Компьютер выведет на экран:

```
<a href="1.html">Гарри Поттер</a>
<a href="2.html">Джинни Уизли</a>
<a href="3.html">Лили Поттер</a>
<a href="4.html">Гермиона Грейндженер</a>
<a href="5.html">Сириус Блэк</a>
<a href="6.html">Рубеус Хагрид</a>
```

...



Статистические данные в журналистике

1. Использовать результаты исследований и опросов как инфоповод.
Нужно знать хотя бы основы статистики и научного метода, чтобы верно интерпретировать результаты исследования!

Пример: важна правильная интерпретация результатов исследования

ВОЗ не нашла доказательств вреда от микропластика в питьевой воде

<https://nplus1.ru/news/2019/08/22/no-harm>

ВОЗ: Микропластик в питьевой воде безопасен для здоровья



КОММЕНТАРИИ

<https://ru.euronews.com/2019/08/22/no-health-risk-from-microplastics-in-drinking-water-ru>

Пример: важен контекст (особенно для подозрительных данных)

Reality Check: Is Malmo the 'rape capital' of Europe?

<https://www.bbc.com/news/uk-politics-39056786>

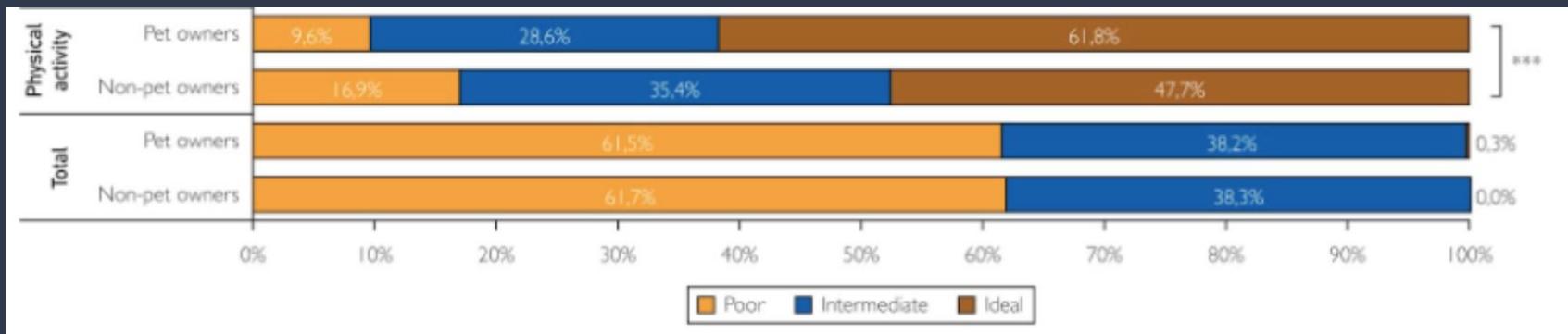
Sweden approves new law recognising sex without consent as rape

<https://www.bbc.com/news/world-europe-44230786>

Пример: нужно подвергать сомнению выводы авторов

Собачники отличились здоровьем сердца и сосудов

<https://nplus1.ru/news/2019/08/23/cvd-health-in-dog-owners>



Пример: нужно подвергать сомнению источники информации

E-cigarettes: Can They Help People Quit?

<https://www.bbc.co.uk/programmes/p03hnfb6>

... и валидность опросов общественного мнения

The 1936 *Literary Digest* Poll (нерепрезентативная выборка)

https://en.wikipedia.org/wiki/The_Literary_Digest

The 1992 UK general election (shy Tories, social acceptability)

<https://www.itv.com/news/2017-04-18/times-the-election-polls-got-it-wrong/>

Цифровое литературоведение!

- Распределение разных тематических слов в 50 000 романов:

