

# Парсинг HTML. XPath

Урок 4



# План курса

1

Основы клиент-серверного взаимодействия. Парсинг API.

2

Парсинг HTML. BeautifulSoup.

3

СУБД MongoDB и ClickHouse в Python

4

Парсинг HTML. XPath.

5

Scrapy.

6

Scrapy. Парсинг фото и файлов.

7

Selenium в Python.

8






Работа с данными.

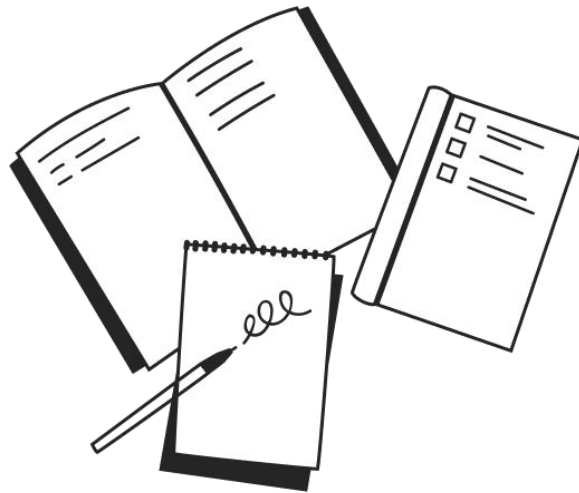
9

Инструменты разметки наборов данных.



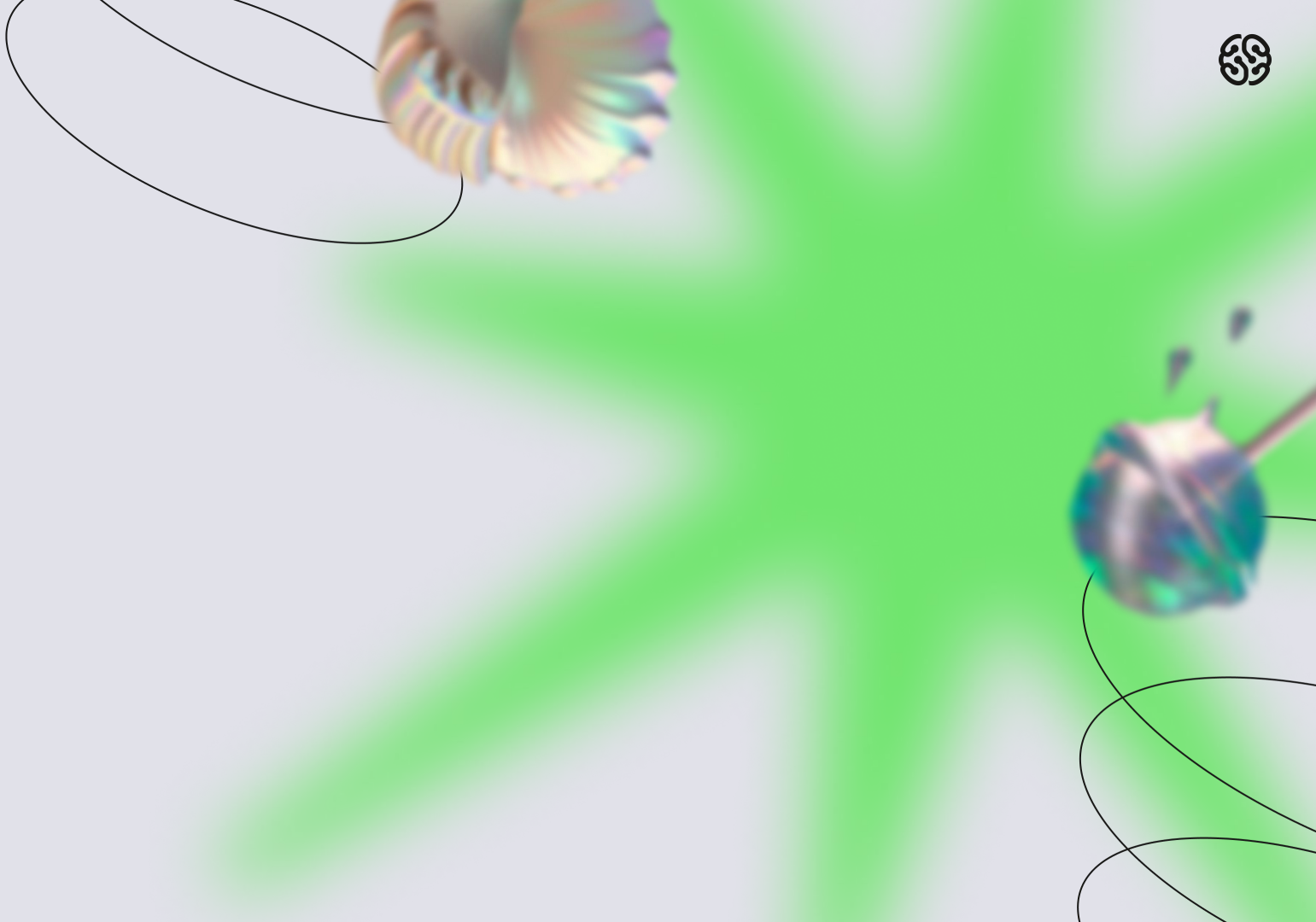
## Что будет на уроке сегодня

-  Основы LXML
-  XPath
-  XPath в Lxml
-  CSS селекторы
-  Скрейпинг веб-сайта с помощью XPath





LXML





Задание

Выведите текст абзаца `<p>Hello GeekBrains</p>`,  
используя метод `find`.





Задание

Почему в пути, который я указал в методе `find()`  
я начал не с тега `<html>`, а с тега `<head>`?





# Xpath

— это XML Path Language — язык запросов к элементам XML — документа.





# XPath Expression

— выражение, определяющее  
шаблон для выбора узлов.







```
//div

//elementName [@attribute = 'value']
//elementName [@id = 'value']
//elementName [@class = 'value']

//li[1]

//li [position() = 1 or position = 2]

//li [position() = 1 and contains(@text, "hello")]
```



axisName::elementName



## Оси для перехода вверх по дереву HTML



### **parent**

возвращает родителя определенного узла



### **ancestor**

возвращает всех предков определенного узла



### **preceding**

выбирает все узлы, которые появляются перед текущим узлом, за исключением предков, узлов атрибутов и пространства имен



### **preceding-sibling**

возвращает все элементы одного уровня до текущего узла



## Оси для перехода вниз по дереву HTML



### **child**

возвращает дочерние элементы (потомков) определенного узла



### **following**

возвращает все элементы, находящиеся после закрывающего тега определенного узла



### **following-sibling**

возвращает все элементы одного уровня после текущего узла



### **descendant**

возвращает всех потомков текущего узла



Задание

Попробуйте самостоятельно получить доступ к тексту тега `<p>` и вывести текст “Hello GeekBrains”





Задание

Используйте метод `cssselector` вместо  
XPath и выберите тег абзаца.





# Скрейпинг IMDb

