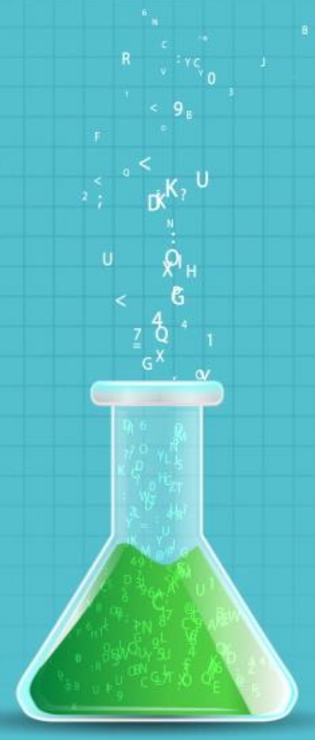
Data Science with Python

Lesson 11 — Web Scraping with BeautifulSoup

DATA SCIENCE

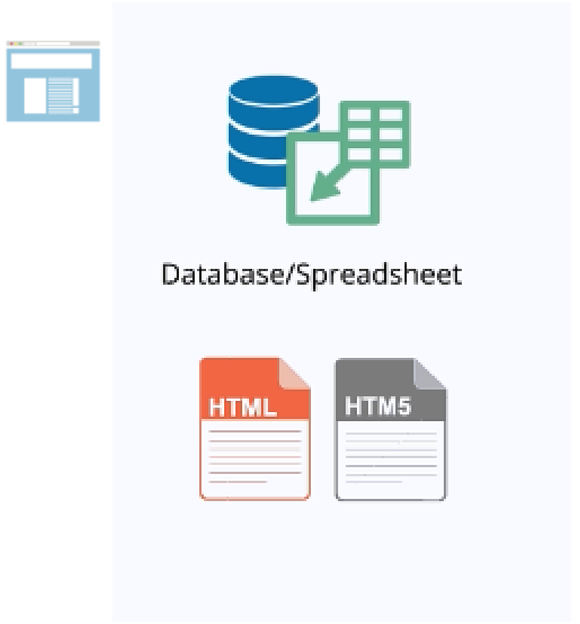
What’s In It For Me

- Define web scraping and explain the importance of web scraping
- Lists the steps involved in the web scraping process
- Describe basic terminologies such as parser, object, and tree associated with the BeautifulSoup
- Understand various operations such as searching, modifying, and navigating the tree to yield the required result



What is Web Scraping

Web scraping is a computer software technique of extracting information from websites in an automated fashion.

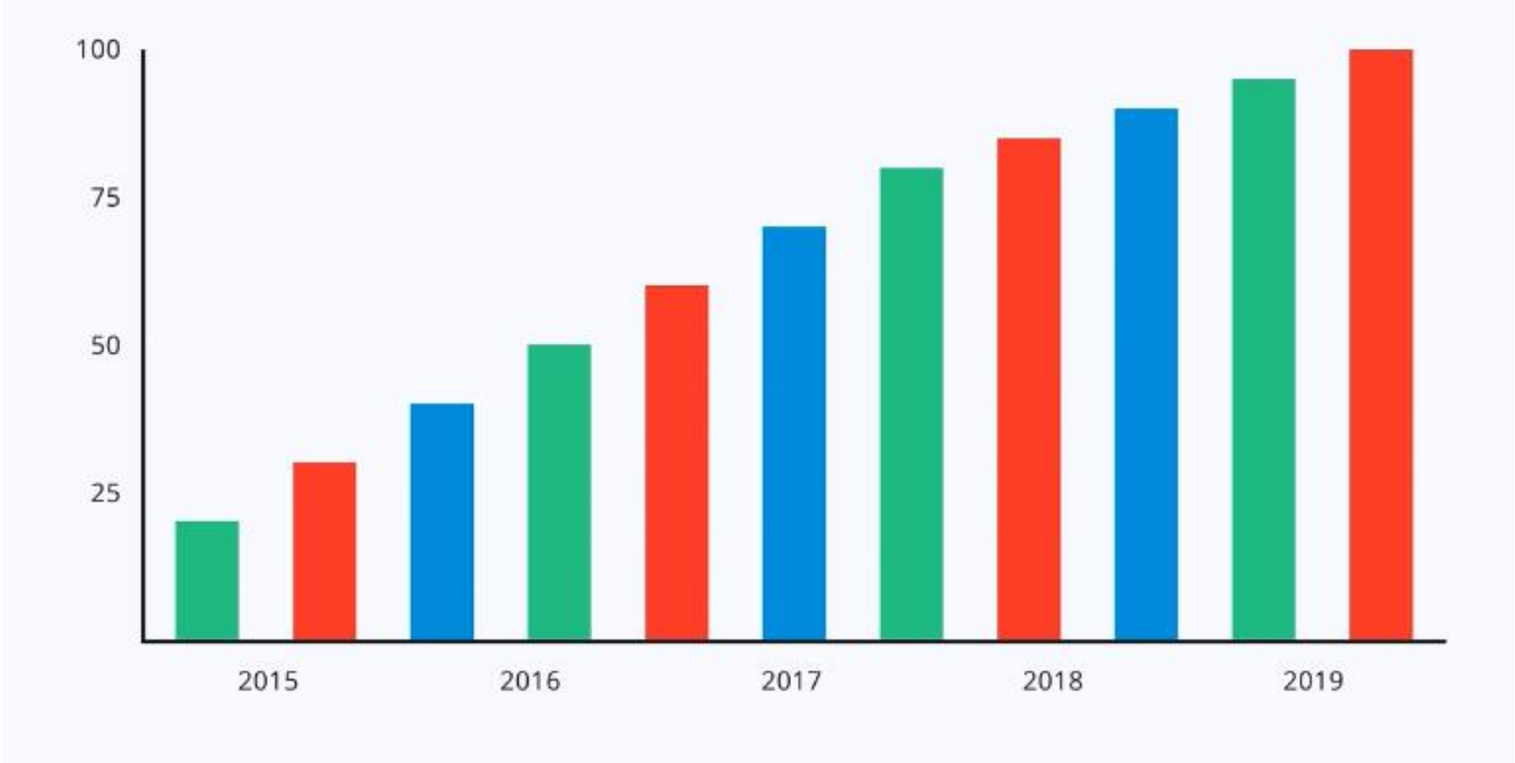


What Web Scraping is (contd.)

Web scraping is a computer software technique of extracting information from websites in an automated fashion.

Why Web Scraping

Every day, you find yourself in a situation where you need to extract data from the web.



Why Web Scraping (contd.)

Every day, you find yourself in a situation where you need to extract data from the web.

Web Scraping Process—Basic Preparation

There are two basic things to consider before setting up the web scraping process.

- Understand the target data on the Internet
- Finalize the list of websites

Web Scraping Process (contd.)

Once you have understood the target data and finalized the list of websites, you need to design the web scraping process.

The steps involved in a typical web scraping process are as follows:

Step 1: A web request is sent to the targeted website to collect the required data.

Web Scraping Process (contd.)

Once you have understood the target data and finalized the list of websites, you need to design the web scraping process.

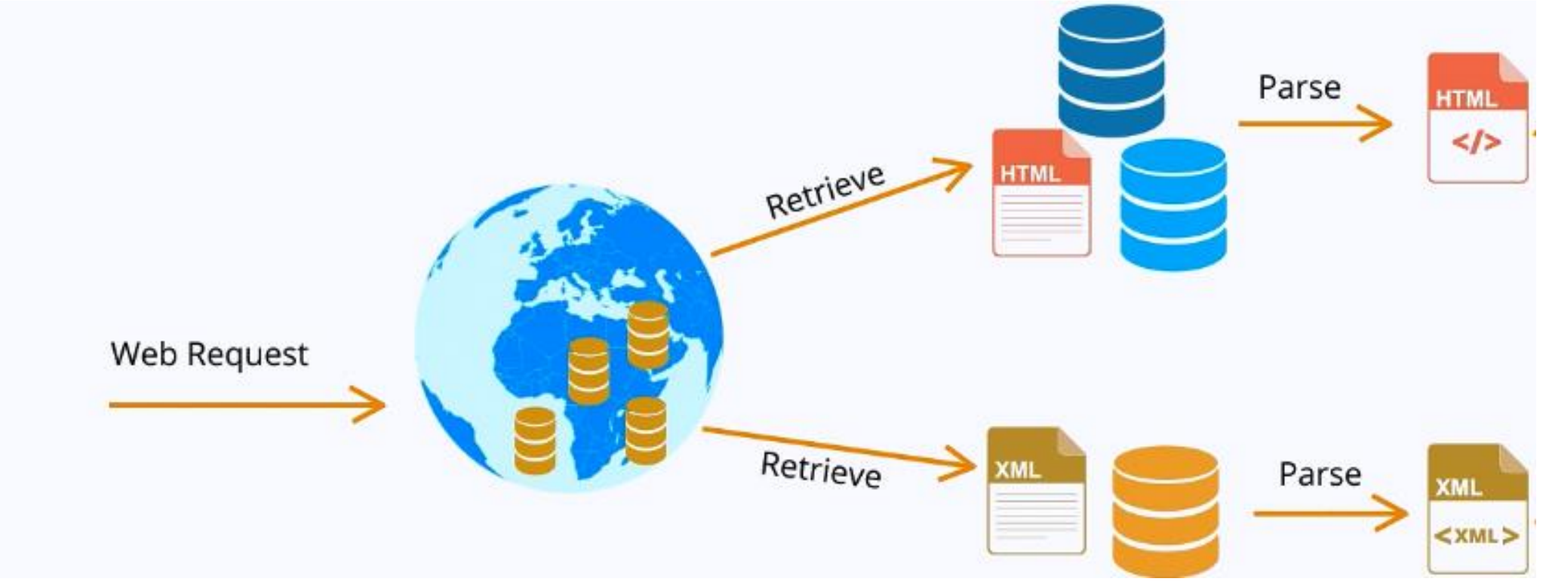
The steps involved in a typical web scraping process are as follows:

Step 2: The information is retrieved from the targeted website in HTML or XML format from web.

Web Scraping Process (contd.)

Once you have understood the target data and finalized the list of websites, you need to design the web scraping process.

The steps involved in a typical web scraping process are as follows:

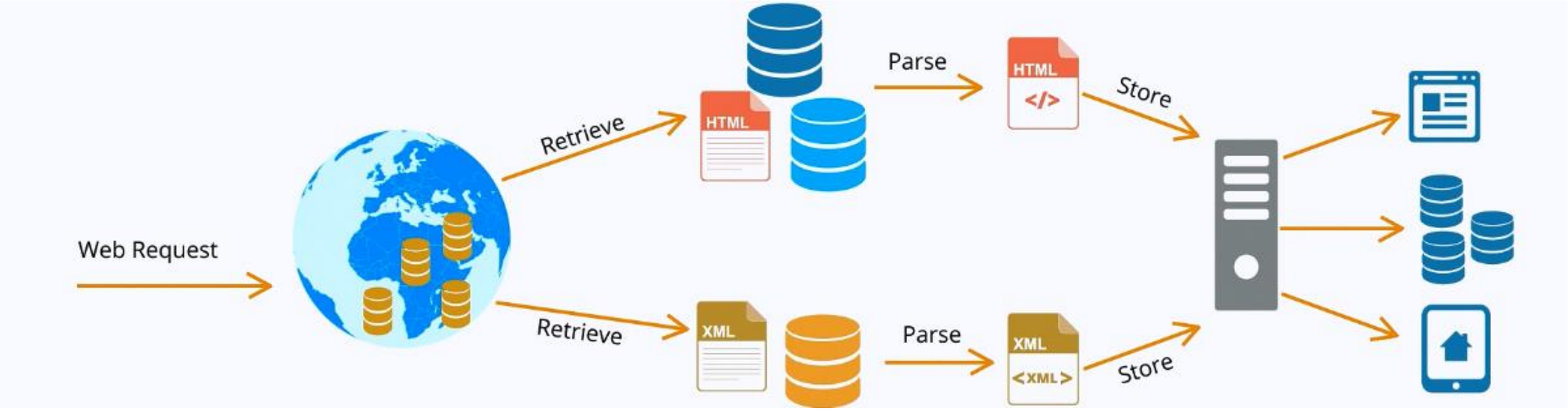


Step 3: The retrieved information is parsed to the several parsers based on the data format. Parsing is a technique to read data and extract information from the available document.

Web Scraping Process (contd.)

Once you have understood the target data and finalized the list of websites, you need to design the web scraping process.

The steps involved in a typical web scraping process are as follows:



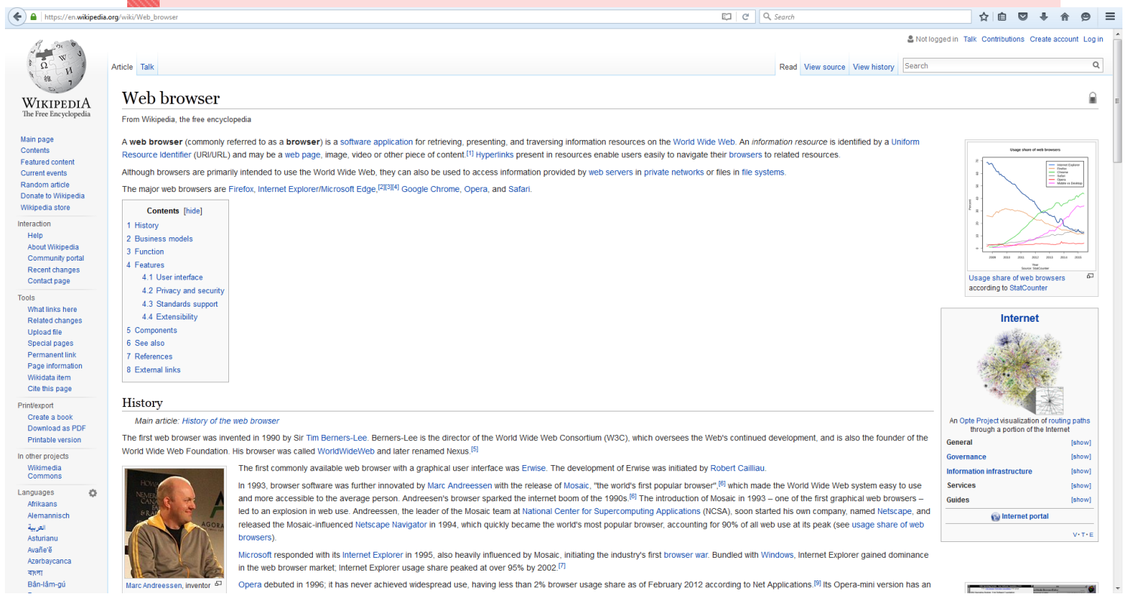
Step 4: The parsed data is stored in the desired format. You can follow the same process to scrap another targeted web.

Web Scraping Software vs. Web Browser

A web scraping software will interact with websites in the same way as your web browser.

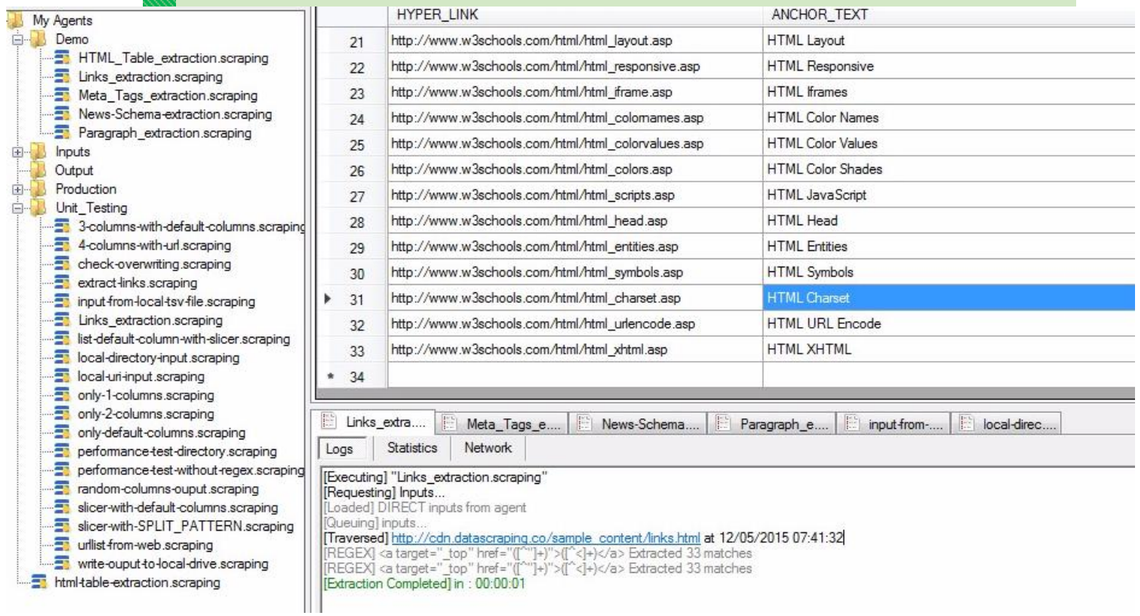
A Web scraper is used to extract the information from web in routine and automated manner.

Web Browser



Displays the data

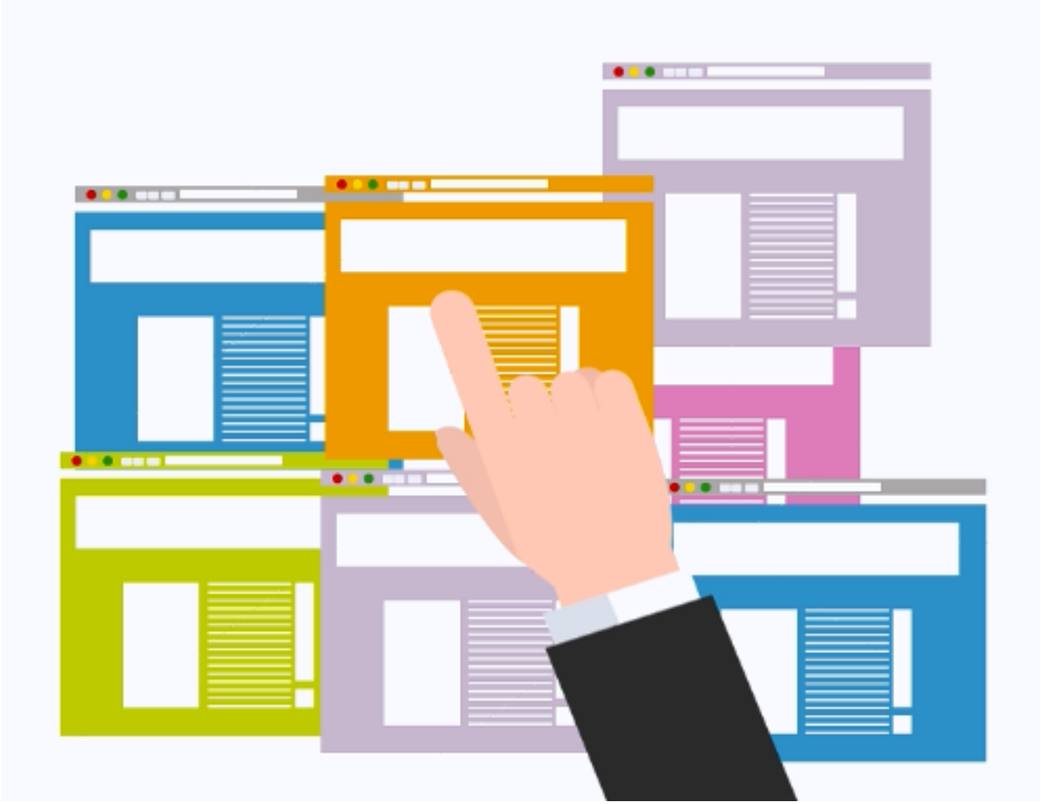
Web Scraping Software



Saves data from the web page to the local file or database

Web Scraping Considerations

It’s important to read and understand the legal information and terms and conditions mentioned in the website.



Web Scraping Considerations (contd.)

It's important to read and understand the legal information and terms and conditions mentioned in the website.

Legal Constraints

Notice

Copyright

Trademark Material

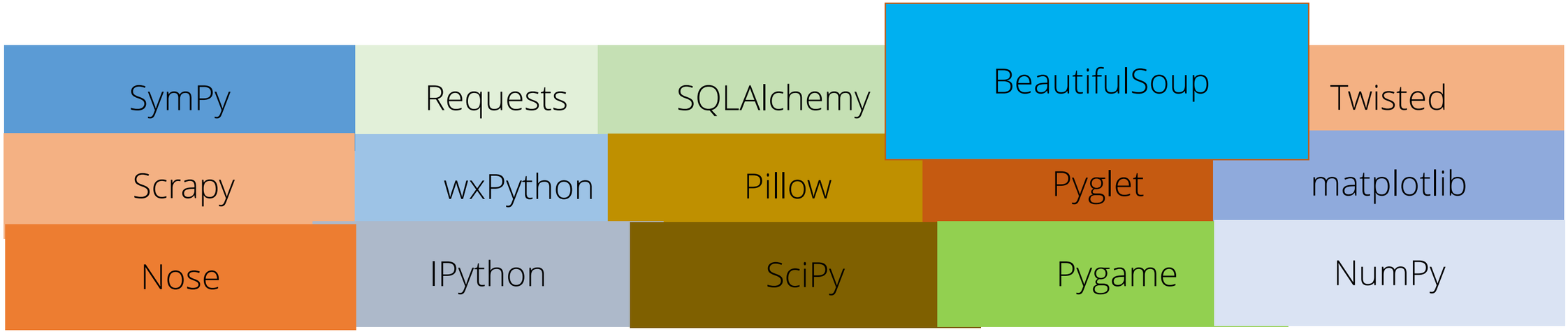
Patented Information

Web Scraping Tool—BeautifulSoup

SymPy	Requests	SQLAlchemy	BeautifulSoup	Twisted
Scrapy	wxPython	Pillow	Pyglet	matplotlib
Nose	IPython	SciPy	Pygame	NumPy

Web Scraping Tool: BeautifulSoup (contd.)

BeautifulSoup, is an easy, intuitive, and a robust Python library designed for web scraping.



Web Scraping Tool: BeautifulSoup (contd.)

BeautifulSoup, is an easy, intuitive, and a robust Python library designed for web scraping.

Some of the reasons to choose BeautifulSoup are as follows:

Efficient tool for dissecting documents and extracting information from the web pages

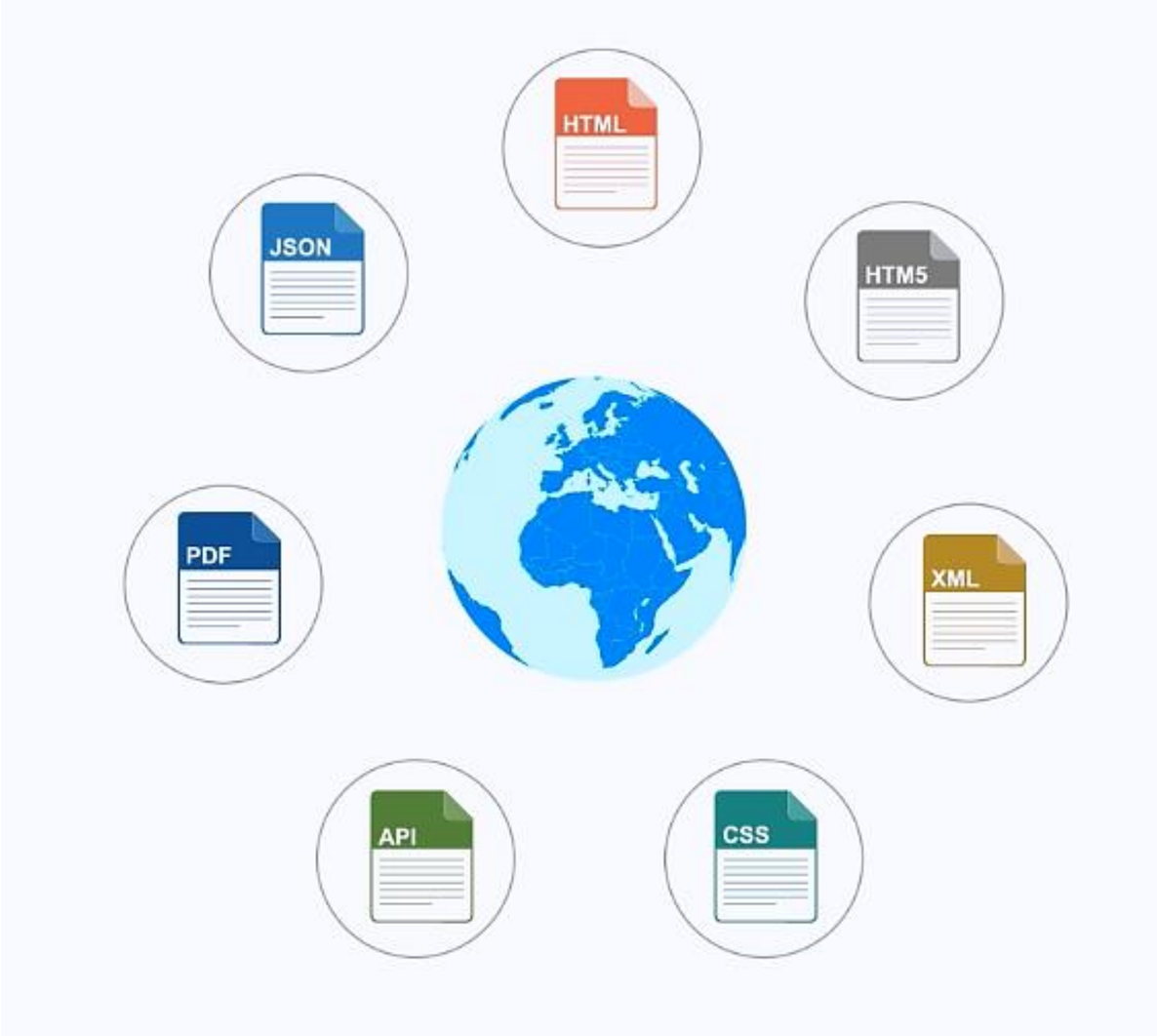
Powerful sets of built-in methods for navigating, searching, and modifying a parse tree

Possess parser that supports both html and xml documents

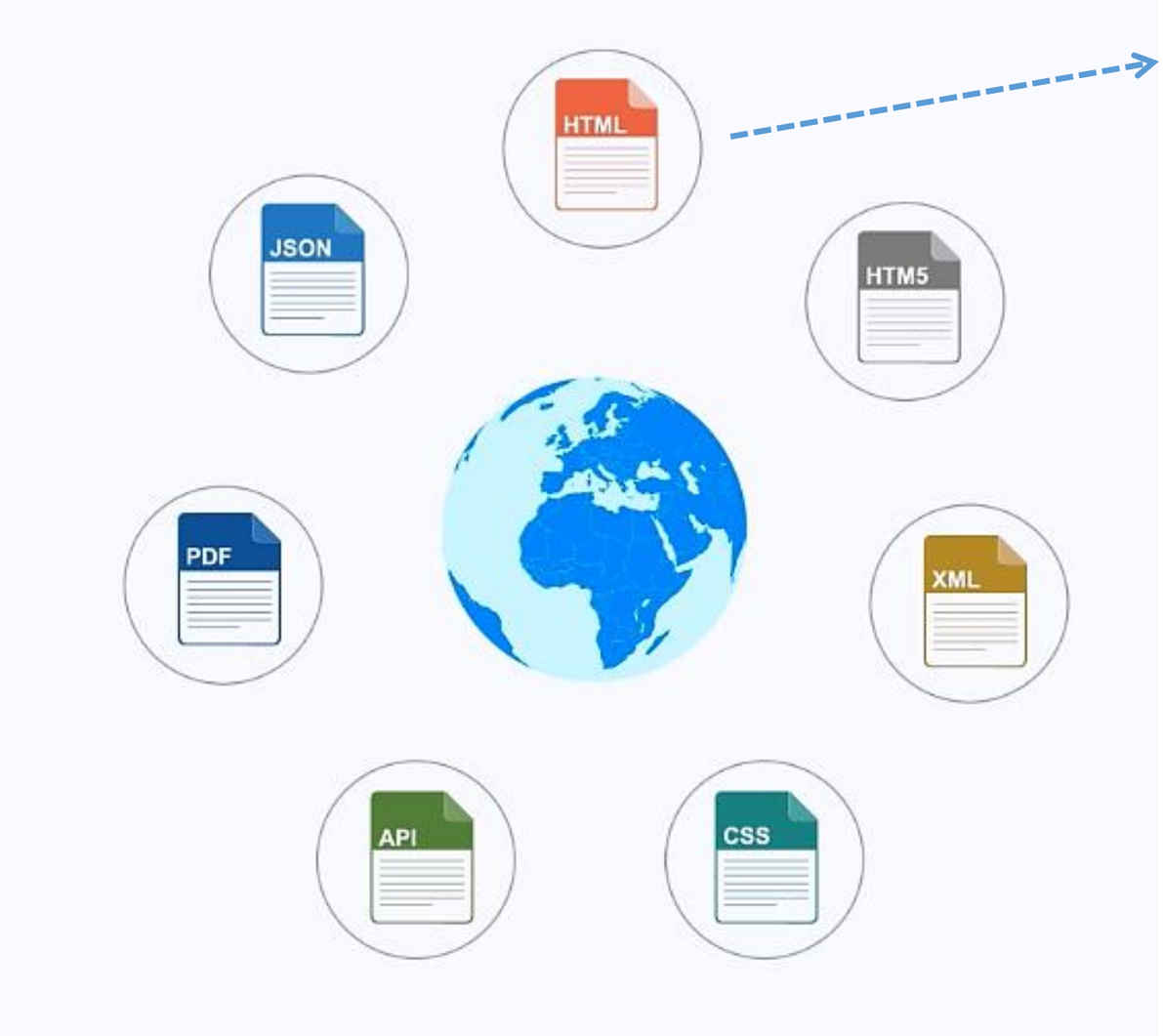
Converts all incoming documents to Unicode automatically

Converts all outgoing documents to UTF-8 automatically

Common Data/ Page Formats on The Web

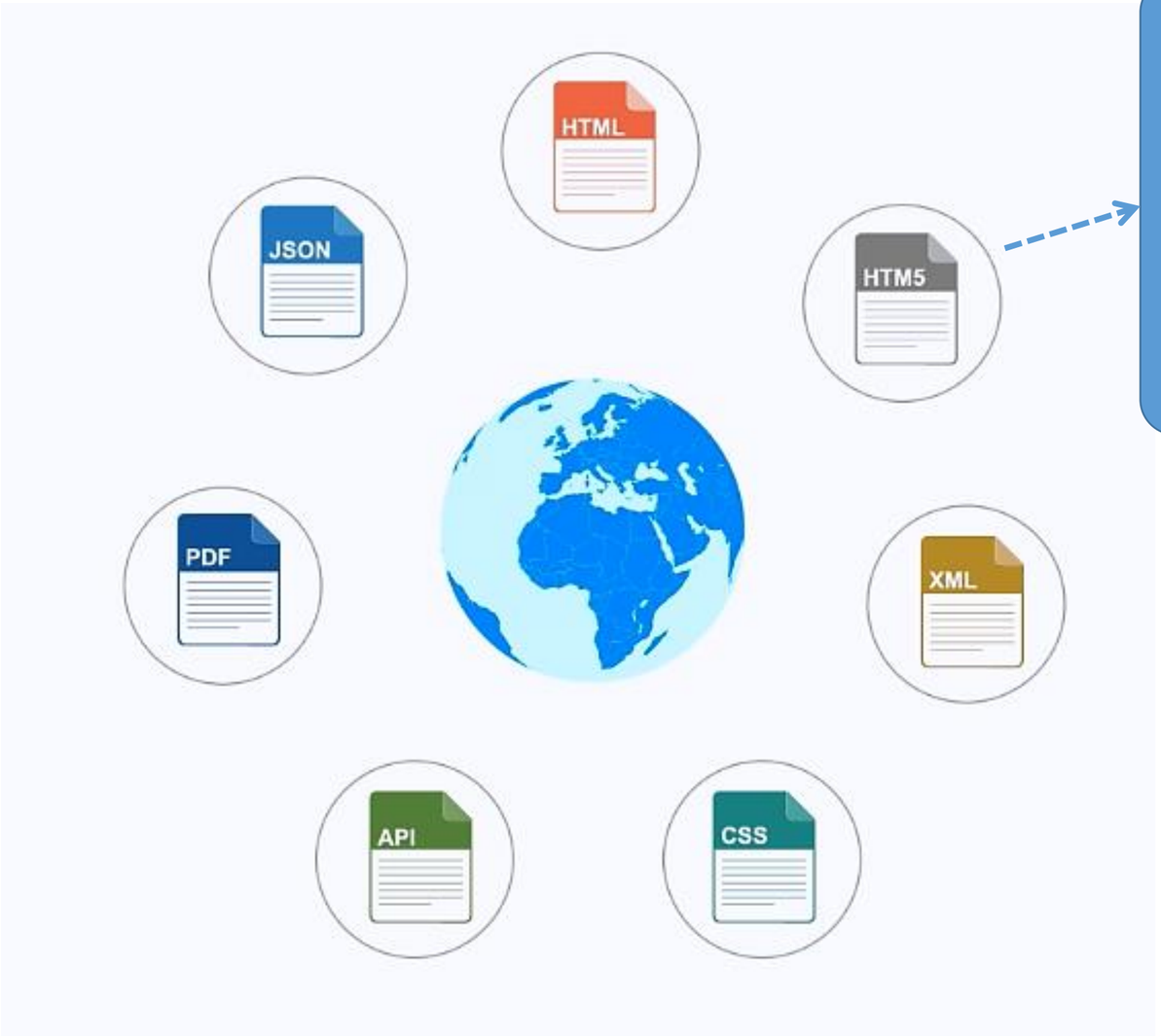


Common Data/ Page Formats on The Web (contd.)



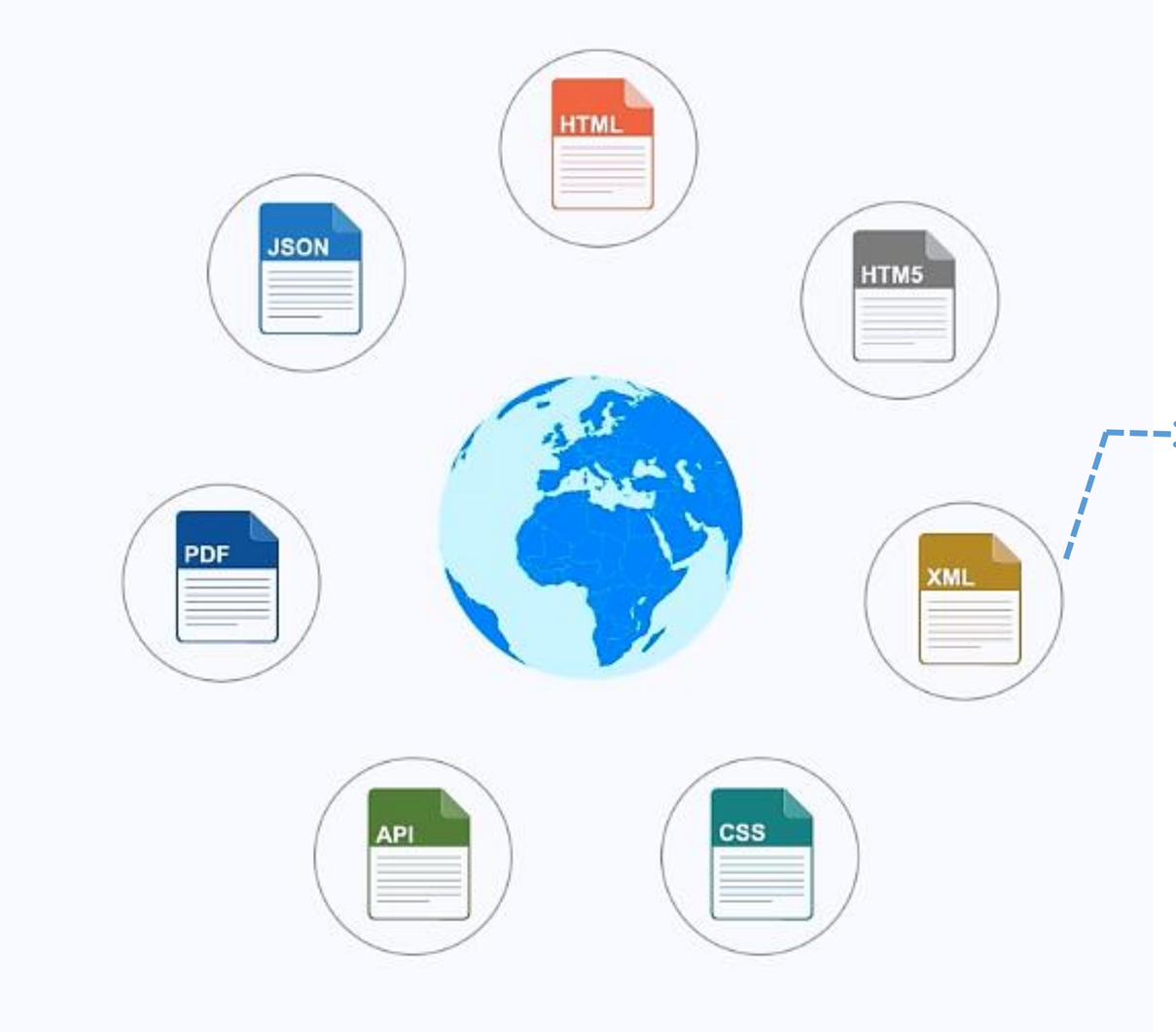
An HTML page is one of the oldest, easiest, and the most popular methods to upload information on the web.

Common Data/ Page Formats on The Web (contd.)



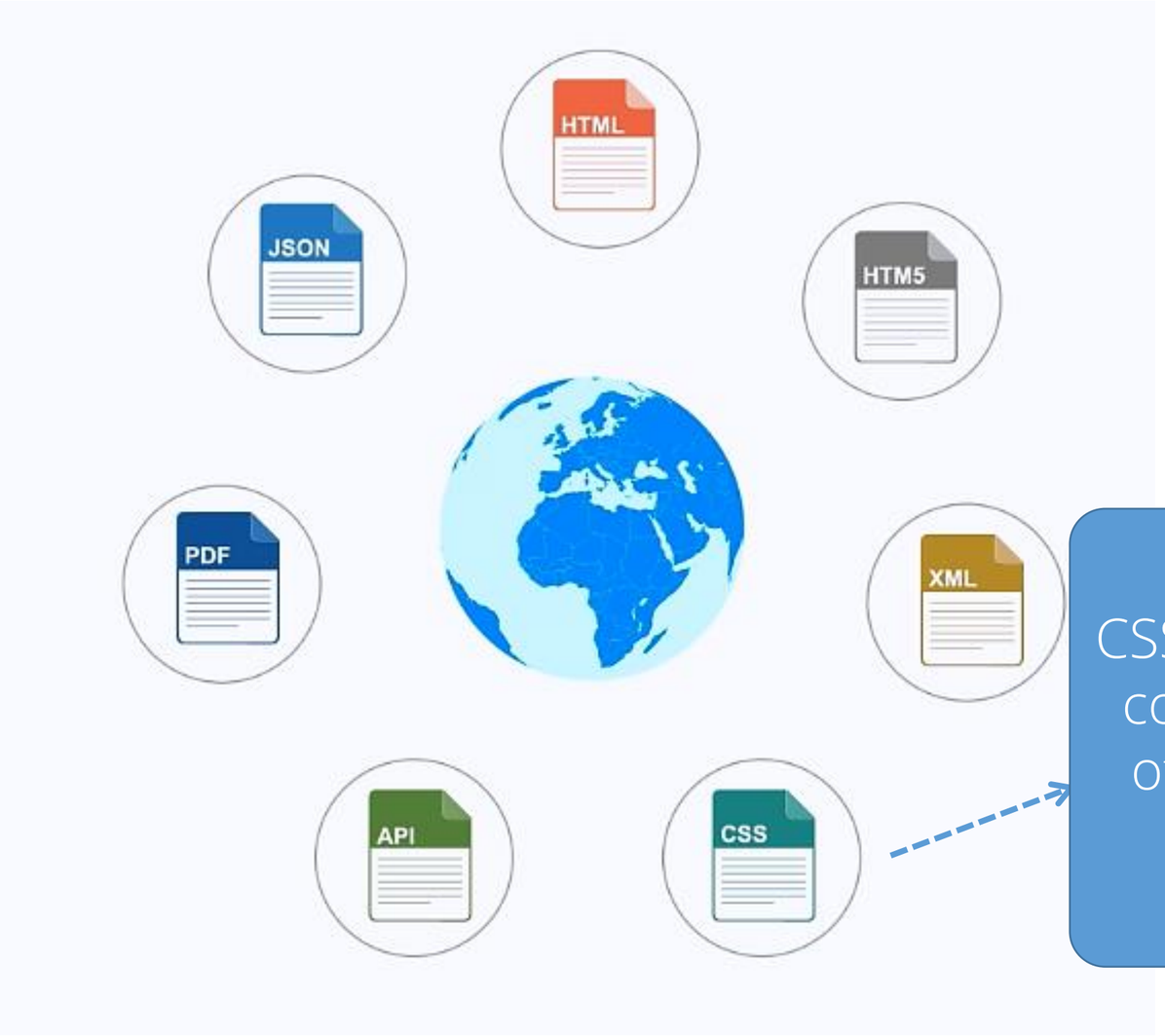
An HTML 5 is a new HTML standard which gained popularity with the mobile devices.

Common Data/ Page Formats on The Web (contd.)



XML is another popular way to upload your information on the web.

Common Data/ Page Formats on The Web (contd.)



CSS is mainly used for the consistent presentation of data using cascaded style sheets.

Common Data/ Page Formats on The Web (contd.)

Application Program Interface, or APIs, has now become a common practice to extract information from the web.

Common Data/ Page Formats on The Web (contd.)

PDF is also widely used to upload information and reports.

Common Data/ Page Formats on The Web (contd.)

JavaScript Object Notation, or JSON, is a lightweight and popular format used for information exchange on the web.

The Parser

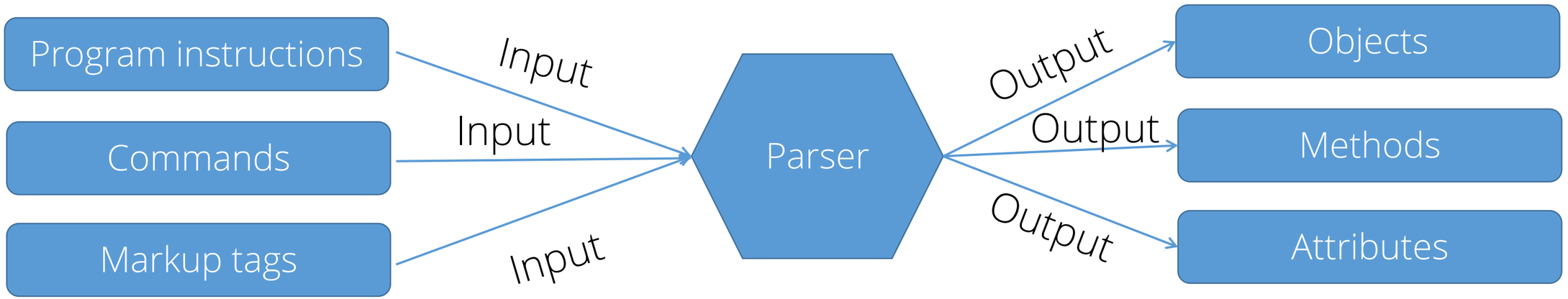
What is a parser?

How does it help Data Scientists in the web scraping process?

The Parser

A Parser is a basic tool to interpret or render information from a web document.

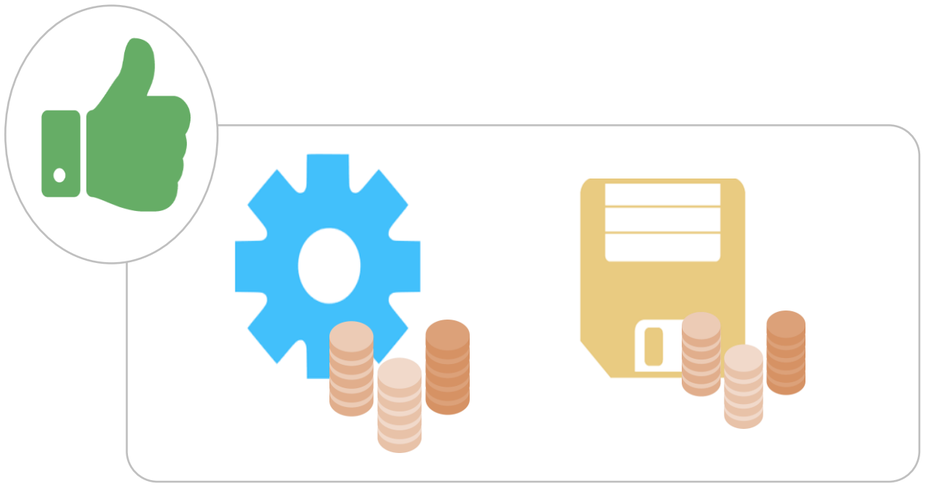
A Parser is also used to validate the input information before processing it.



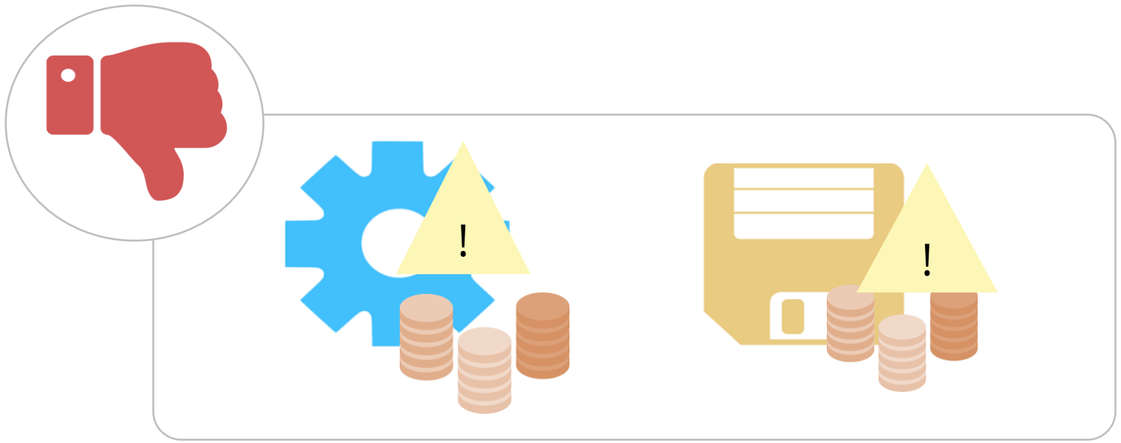
Importance of Parsing

Parsing data is one of the most important steps in the web scraping process.

Failing to parse the data would eventually lead to a failure of the entire process.



Parser



Parser

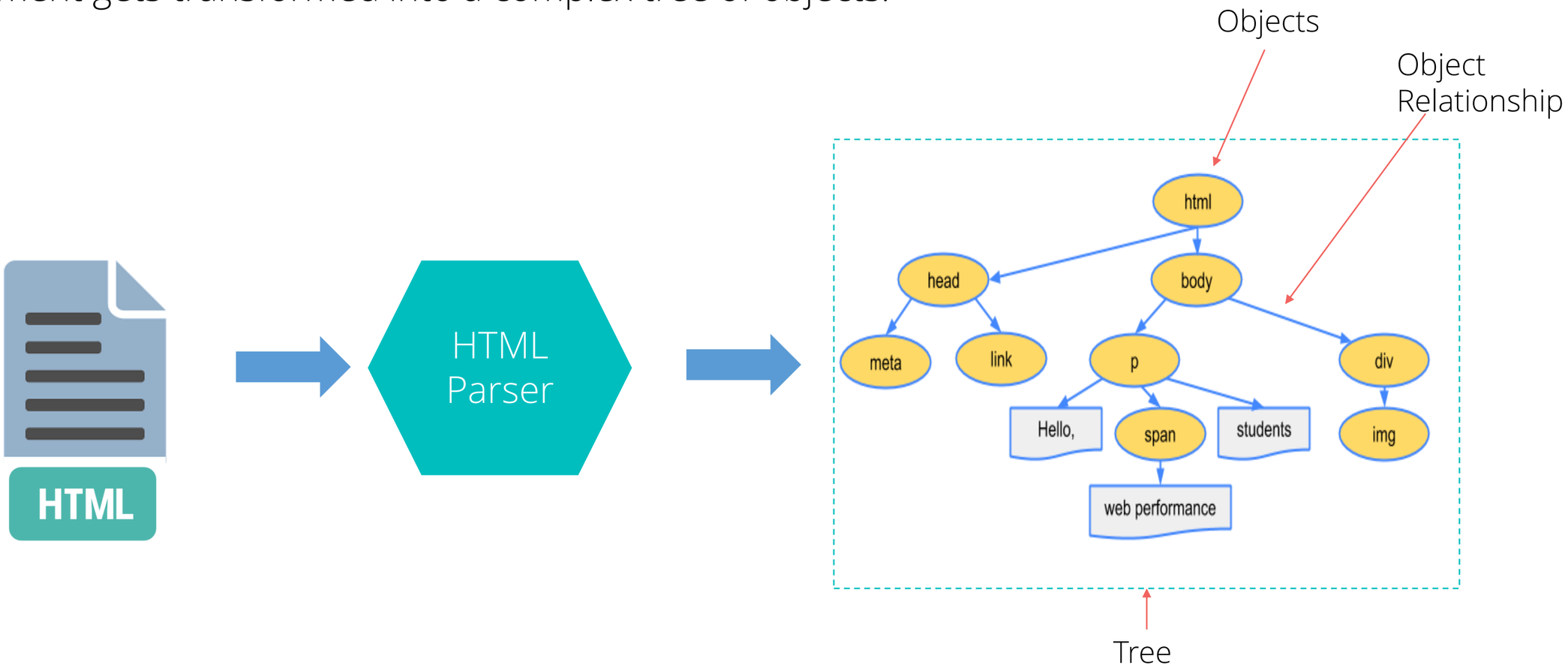
Various Parser

There are various parsers supported by BeautifulSoup:

html.parser	HTML parser is Python based, fast, and lenient.
lxml html	Lxml html is not built using Python and it depends on C. However, it is fast and lenient in nature.
lxml xml	Lxml xml is the only xml parser available and it also depends on C.
html5lib	HTML 5lib is another Python-based parser; however, it is slow and able to create valid HTML5.

Importance of Objects

A web document gets transformed into a complex tree of objects.

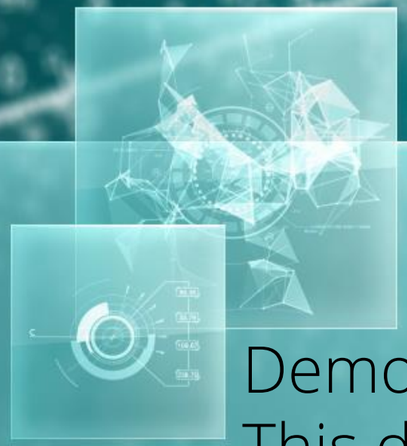


A tree is defined as a collection of simple and complex objects.

Types of Objects

BeautifulSoup transforms a complex HTML document into a complex tree of Python objects. There are four types of objects. They are:

Tag	A tag object is an XML or HTML tag in the web document. Tags have a lot of attributes and methods.
NavigableString	A NavigableString is a string or set of characters that corresponds to the text present within a tag.
BeautifulSoup	A BeautifulSoup represents the entire web document and supports navigating and searching the document tree.
Comment	A Comment represents the comment or information section of the document. It is a special type of NavigableString.



Demo: 01—Parsing web documents and extracting data using objects

This demo shows you how to scrape a web document, parse it, and use objects to extract information.

DATA SCIENCE

Knowledge Check

KNOWLEDGE CHECK

Which of the following object types represents a string or set of characters within a tag?

a. Tag

b. NavigableString

c. BeautifuSoup

d. Comment

KNOWLEDGE CHECK

Which of the following object types represents a string or set of characters within a tag?

a. Tag

b. NavigableString

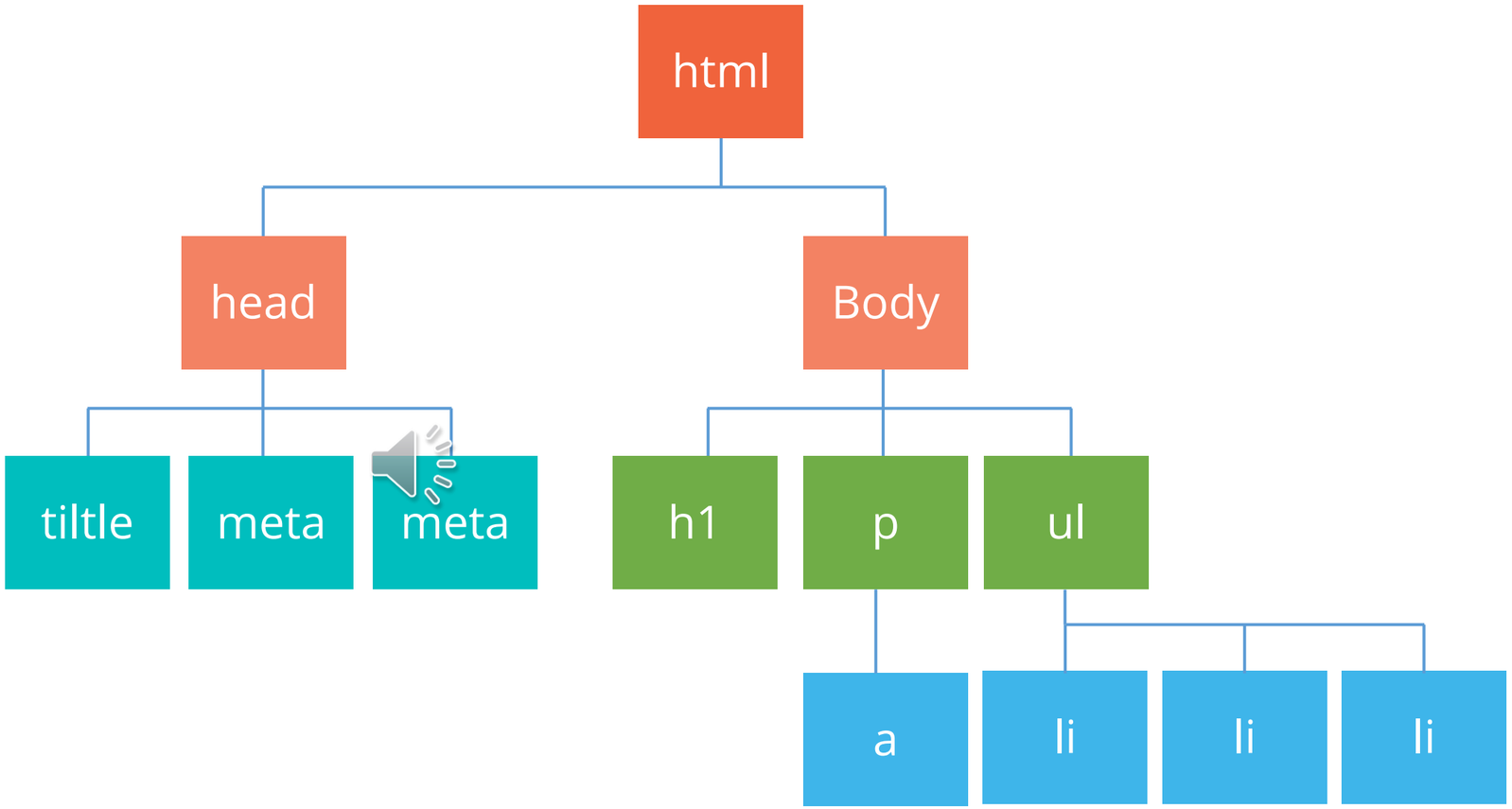
c. BeautifuSoup

d. Comment

The correct answer is b.

Explanation: NavigableString is a string or set of characters that corresponds to the text present within a tag.

Understanding the tree



Understanding The Tree

```
<!DOCTYPE html>
<html>
        <body>
                <div class="oraganizationlist">
                        <ul id="HR">
                                <li class="HRmanager">
                                        <div class="name">Jack</div>
                                        <div class="ID">101</div>
                                </li>
                                <li class="HRmanager">
                                        <div class="name">Daren</div>
                                        <div class="ID">65</div>
                                </li>
                        </ul>
                        <ul id="IT">
                                <li class="ITmanager">
                                        <div class="name">Morris</div>
                                        <div class="ID">39</div>
                                </li>
                                <li class="ITmanager">
                                        <div class="name">Jane</div>
                                        <div class="ID">11</div>
                                </li>
                        </ul>
                        <ul id="Finance">
                                <li class="accountmanager">
                                        <div class="name">Tom</div>
                                        <div class="ID">22</div>
                                </li>
                                <li class="accountmanager">
                                        <div class="name">Kelly</div>
                                        <div class="ID">95</div>
                                </li>
                        </ul>
        </body>
</html
```

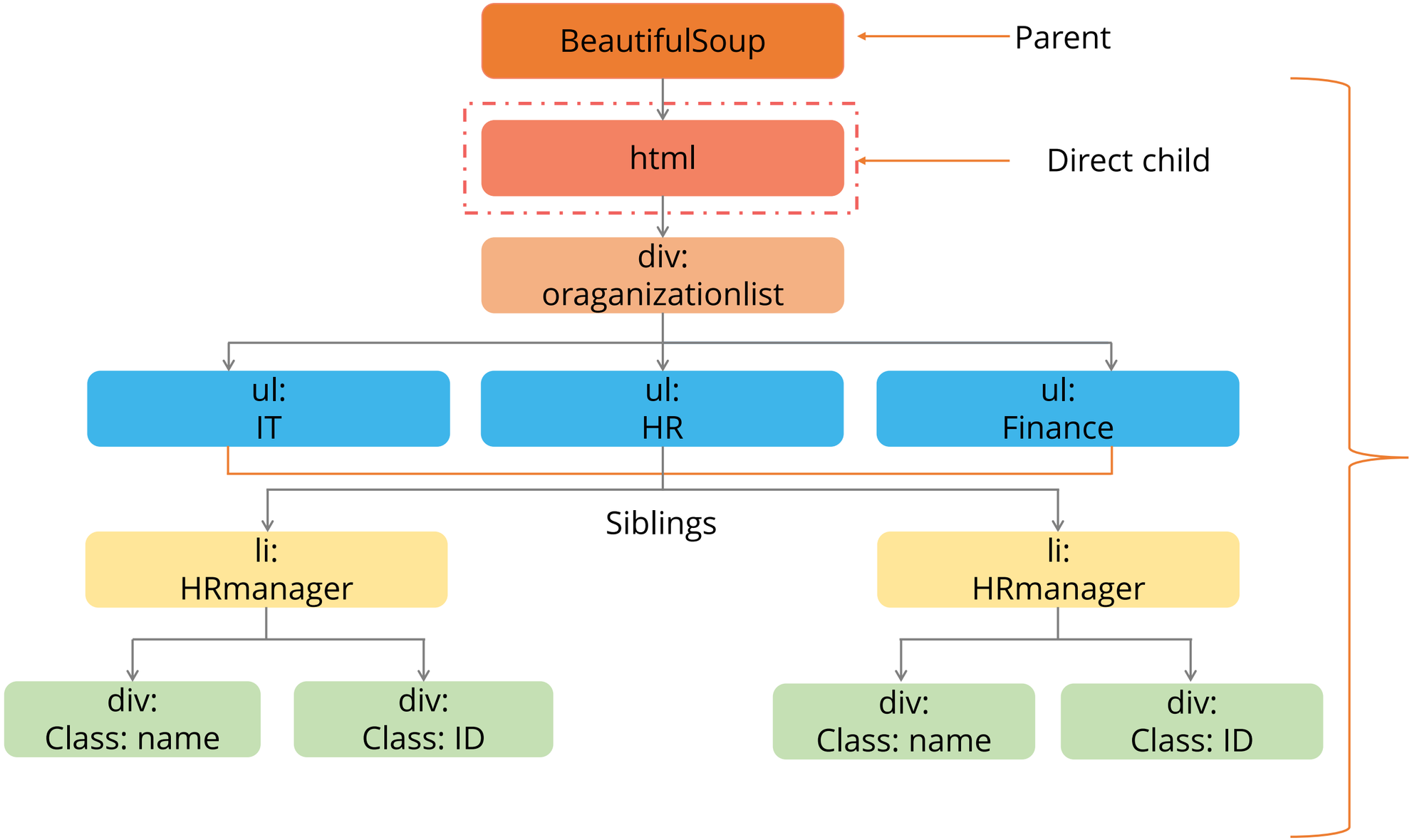
html tag

Body tag

Division or a Section

Cascaded style sheets

Understanding The Tree (contd.)



Searching The Tree – Filters

With the help of the search filters technique, you can extract specific information from the parsed document.

The filters can be treated as search criteria for extracting the information based on the elements present in the document.

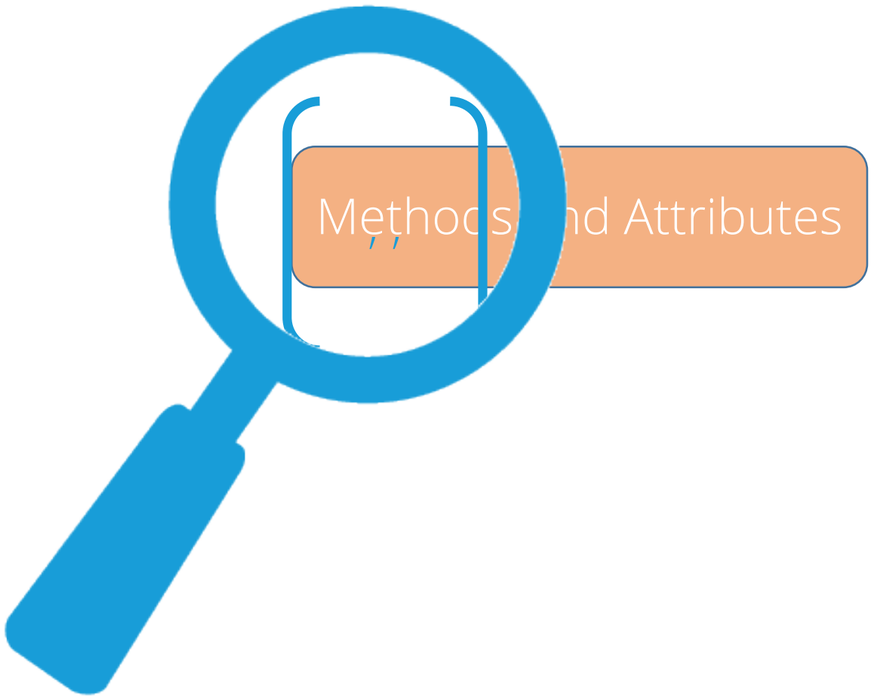
Searching The Tree – Filters (contd.)

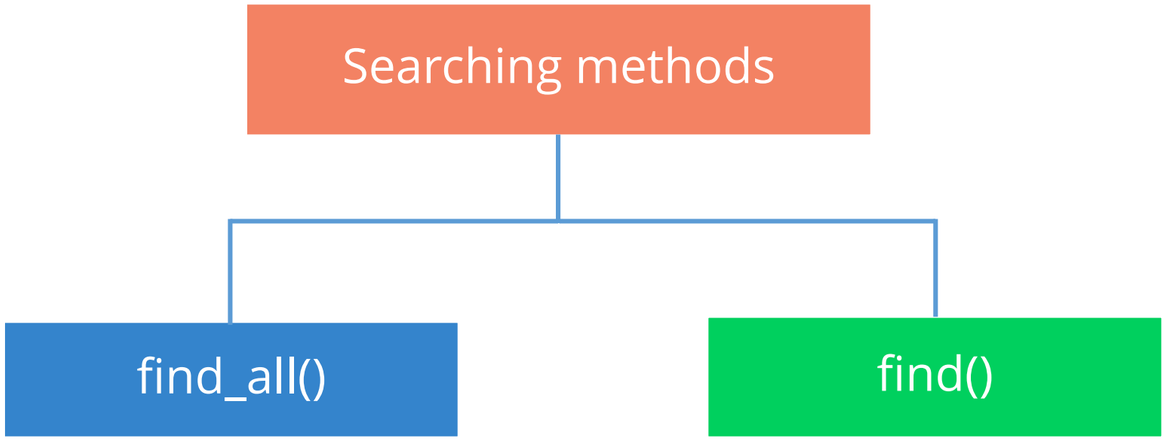
There are various kinds of filters used for searching an information from a tree:

String	A string is the simplest filter. BeautifulSoup will perform a match against the search string.
Regular Expressions	A regular expression filters the match against the search criteria.
List	A list filters the string that matches against the search item in the list.
Function	A function filters the elements that matches against its only argument.

Searching the Tree—find_all()

BeautifulSoup defines a lot of methods for searching the parsed tree.

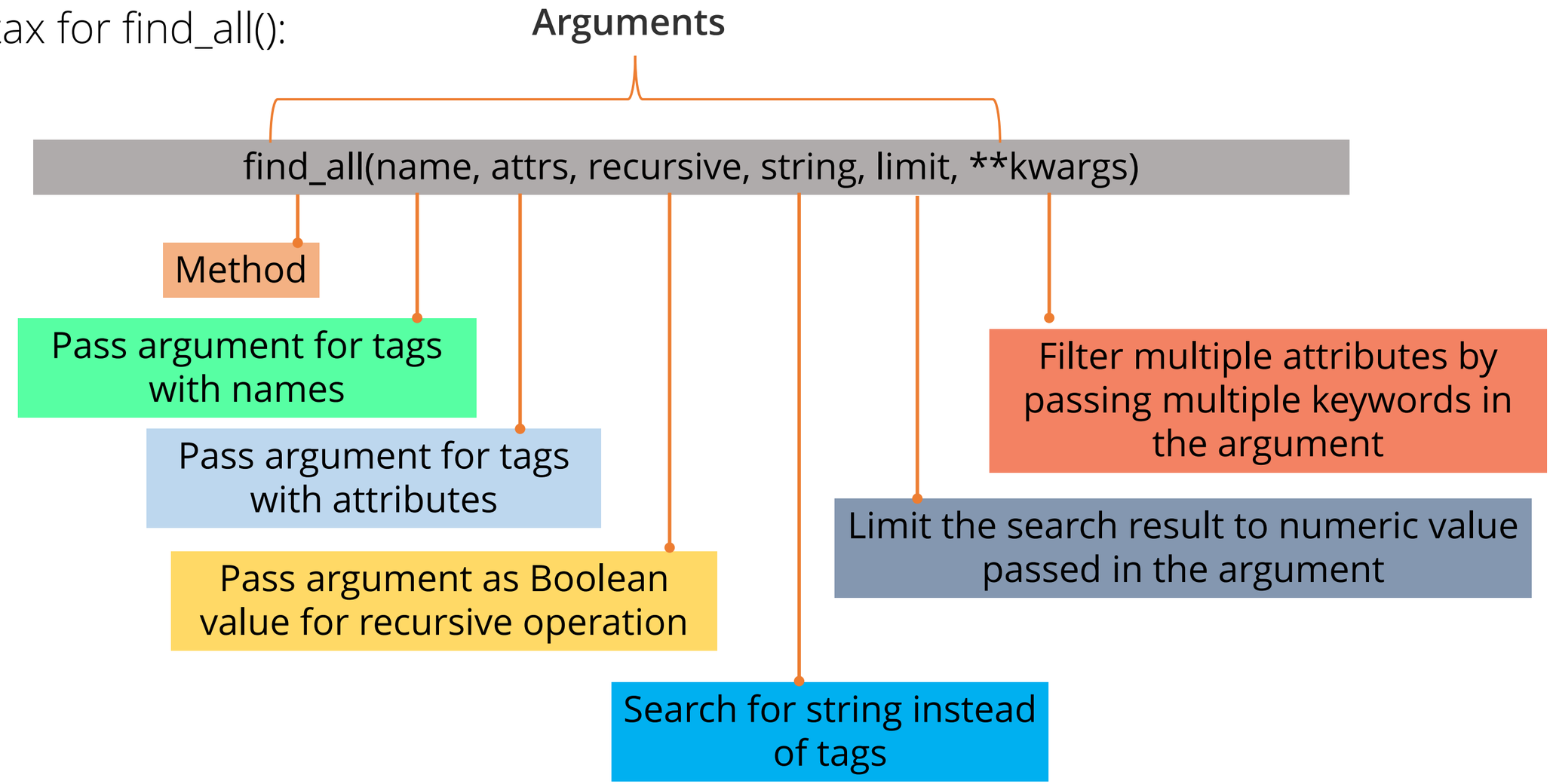




Searching the tree with find_all()

The find_all() searches and retrieves all tags' descendants that matches your filters.

The syntax for find_all():



Searching the tree with find ()

The find_all() finds the entire document looking for results.

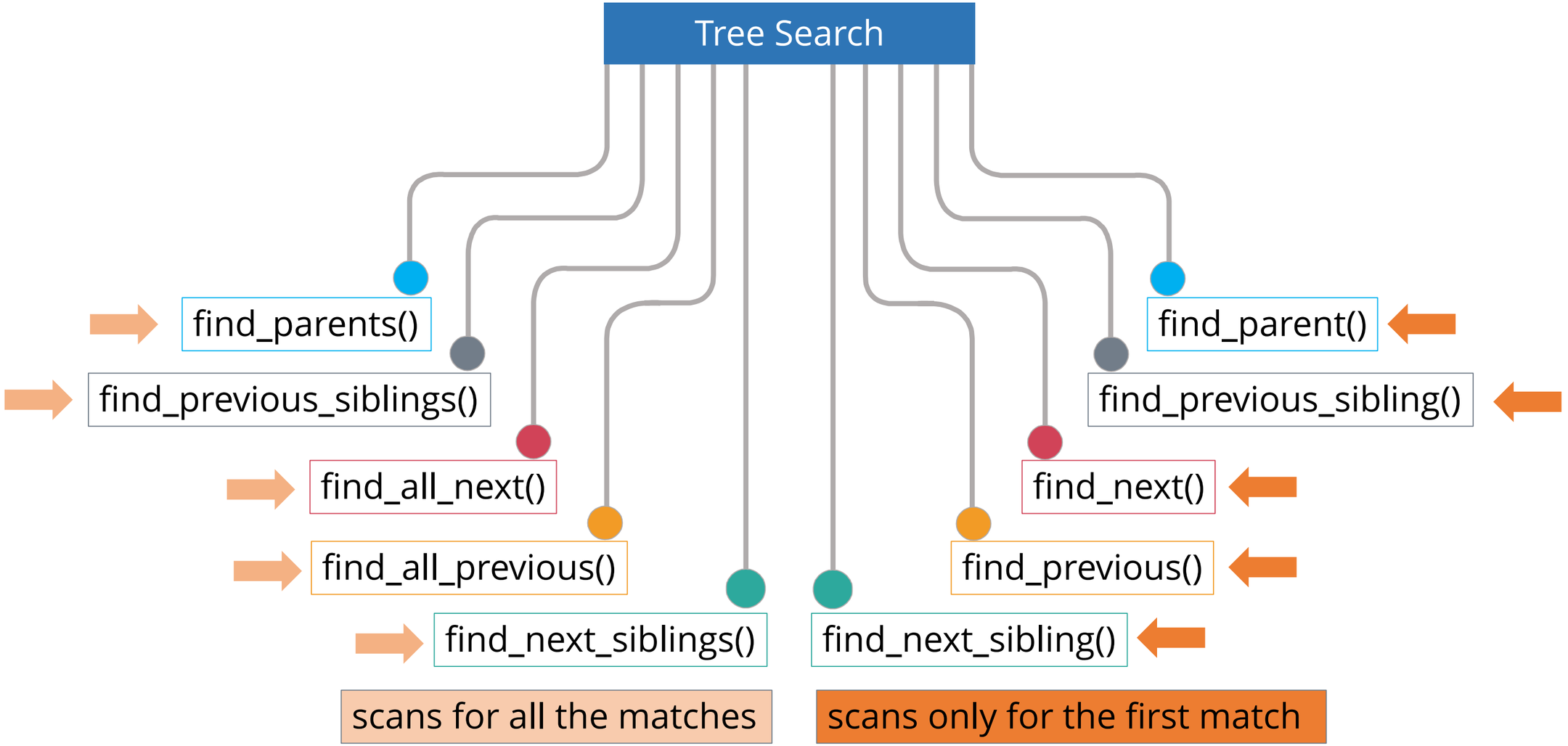
To find one result, use find().

The find() method has a syntax similar to that of the find_all() method; however, there are some key differences.

Method name	Search Scope	Match Found	Match Not Found
Find_all()	Scans entire document	Returns list with values	Returns empty list
Find()	Searches only for passed argument	Returns only the first match value	Returns None

Searching the tree with other methods

Searching the parse tree can also be performed by various other methods such as the following:



Demo: 02—Demo: 02—Searching in a Tree with Filters

This demo shows the ways to search in a tree using filters.

DATA SCIENCE

simplilearn

Knowledge Check

KNOWLEDGE CHECK

The method get_text() is used to _____.

a. parse the entire document

b. parse only part of the document

c. search the tree

d. navigate the tree

KNOWLEDGE CHECK

The method get_text() is used to _____.

a. parse the entire document

b. parse only part of the document

c. search the tree

d. navigate the tree

The correct answer is. b.

Explanation The method get_text() is used to parse only part of the document.

Navigating options

With the help of BeautifulSoup, it is easy to navigate the parse tree based on the need.

There are four options to navigate the tree:

Click each tab to know more.

- Navigating Down
- Navigating Up
- Navigating Sideways
- Navigating Back and Forth

Navigating options

With the help of BeautifulSoup, it is easy to navigate the parse tree based on the need.

There are four options to navigate the tree. They are:

Click each tab to know more.

Navigating Down

Navigating Up

Navigating Sideways

Navigating Back and Forth

This technique shows you how to extract information from children tags. Following are the attributes used to navigate down:

- .contents and .children
- .descendants
- .string
- .strings and stripped_strings

Navigating options

With the help of BeautifulSoup, it is easy to navigate the parse tree based on the need.

There are four options to navigate the tree:

Click each tab to know more.

- Navigating Down
- Navigating Up
- Navigating Sideways
- Navigating Back and Forth

Navigating Up:

Every tag has a parent and two attributes, .parents and .parent,to help navigate up the family tree.

Navigating options

With the help of BeautifulSoup, it is easy to navigate the parse tree based on the need.

There are four options to navigate the tree:

Click each tab to know more.

- Navigating Down
- Navigating Up
- Navigating Sideways
- Navigating Back and Forth

Navigating Sideways:

This technique shows you how to extract information from the same level in the tree.
The attributes used to navigate sideways are .next_sibling and .previous_sibling.

Navigating options

With the help of BeautifulSoup, it is easy to navigate the parse tree based on the need.

There are four options to navigate the tree:

Click each tab to know more.

- Navigating Down
- Navigating Up
- Navigating Sideways
- Navigating Back and Forth

Navigating Back and Forth:

This technique shows you how to parse the tree back and forth.
Following are the attributes used to navigate back and forth are:
.next_element and .previous_element
.next_elements and .previous_elements

Demo: 03—Navigating a Tree

This demo shows how to navigate the web tree using various techniques.

DATA SCIENCE

simplilearn

Knowledge Check

KNOWLEDGE CHECK

Which of the following attributes is used to navigate up?

a. .next_element

b. .parent

c. .previous_elements

d. .next_sibling

KNOWLEDGE CHECK

Which of the following attributes is used to navigate up?

a. .next_element

b. .parent

c. .previous_elements

d. .next_sibling

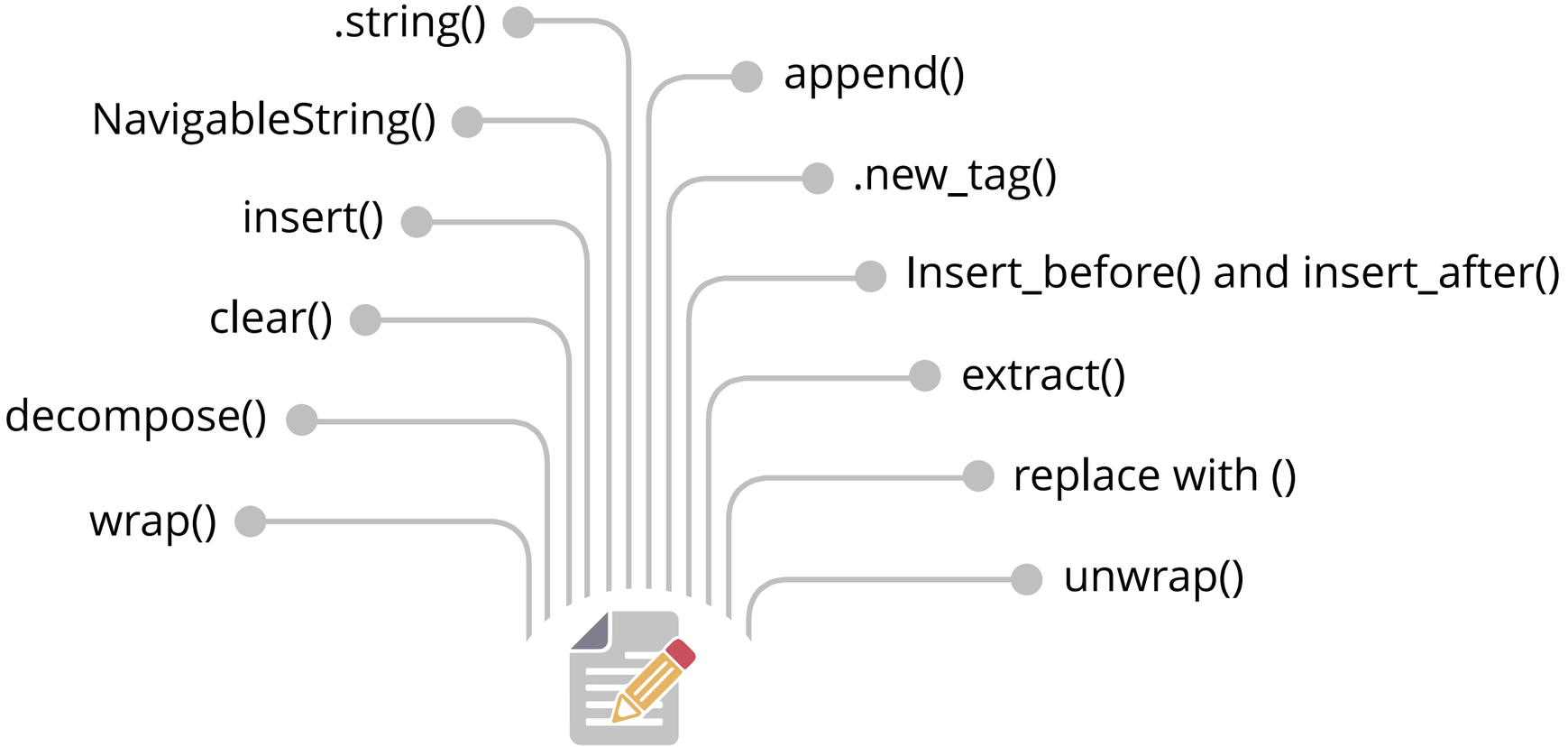
The correct answer is. b.

Explanation: The .parent attribute is used to navigate up.

Modifying The Tree

With BeautifulSoup, you can also modify the tree and write your changes as a new HTML or XML document.

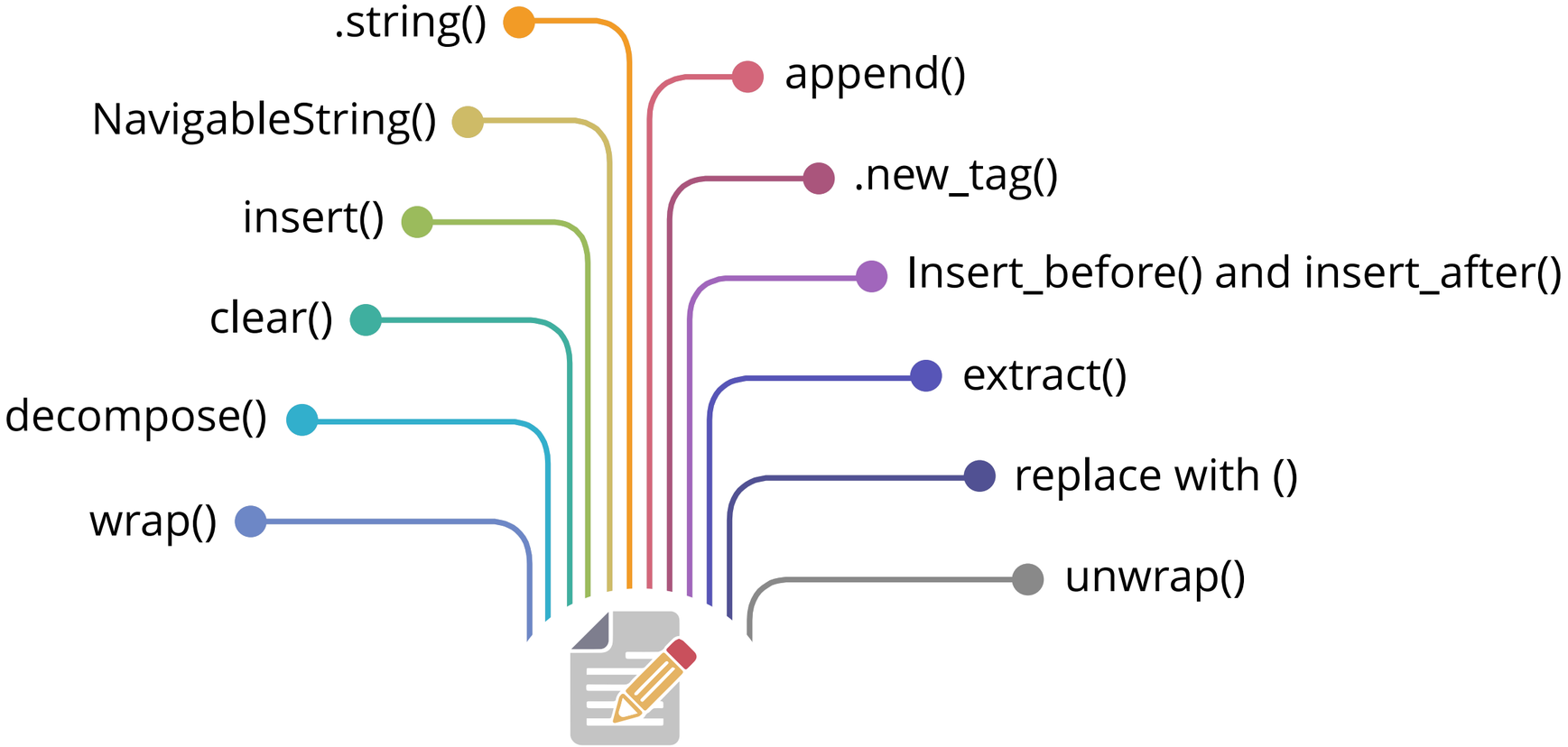
There are several methods to modify the tree:

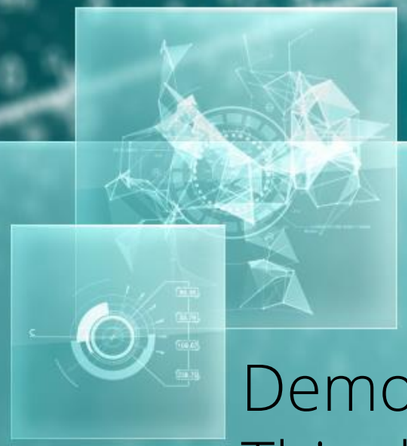


Modifying The Tree

With BeautifulSoup, you can also modify the tree and write your changes as a new HTML or XML document.

There are several methods to modify the tree:





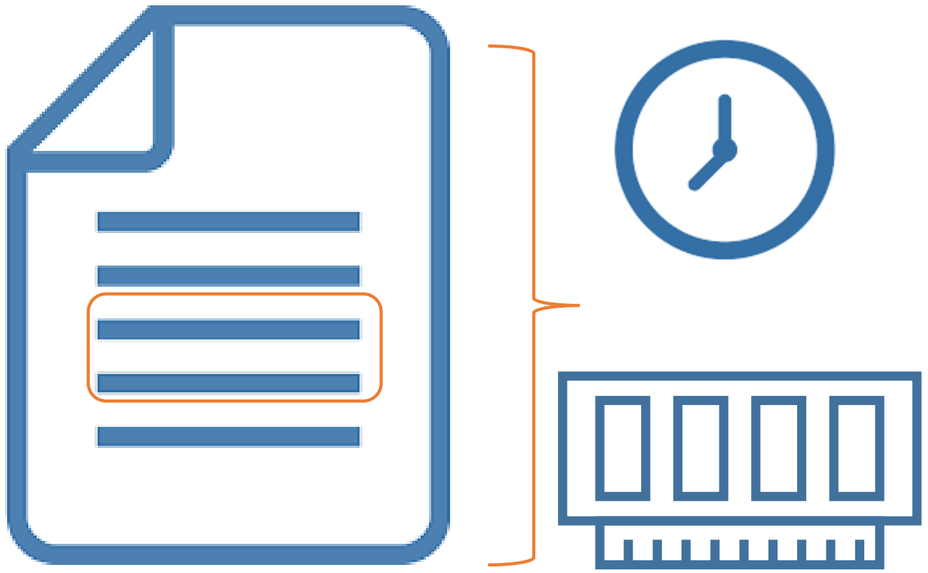
Demo: 04—Modifying the Tree

This demo shows you ways to modify a web tree to get the desired result with the help of an example.

DATA SCIENCE

simplilearn

Parsing Only Part of the Document



But how can you overcome this problem?

Use SoupStrainer class

Allows you to choose the part of the document to be parsed

This feature of parsing a part of the document will not work with the html5lib parser.

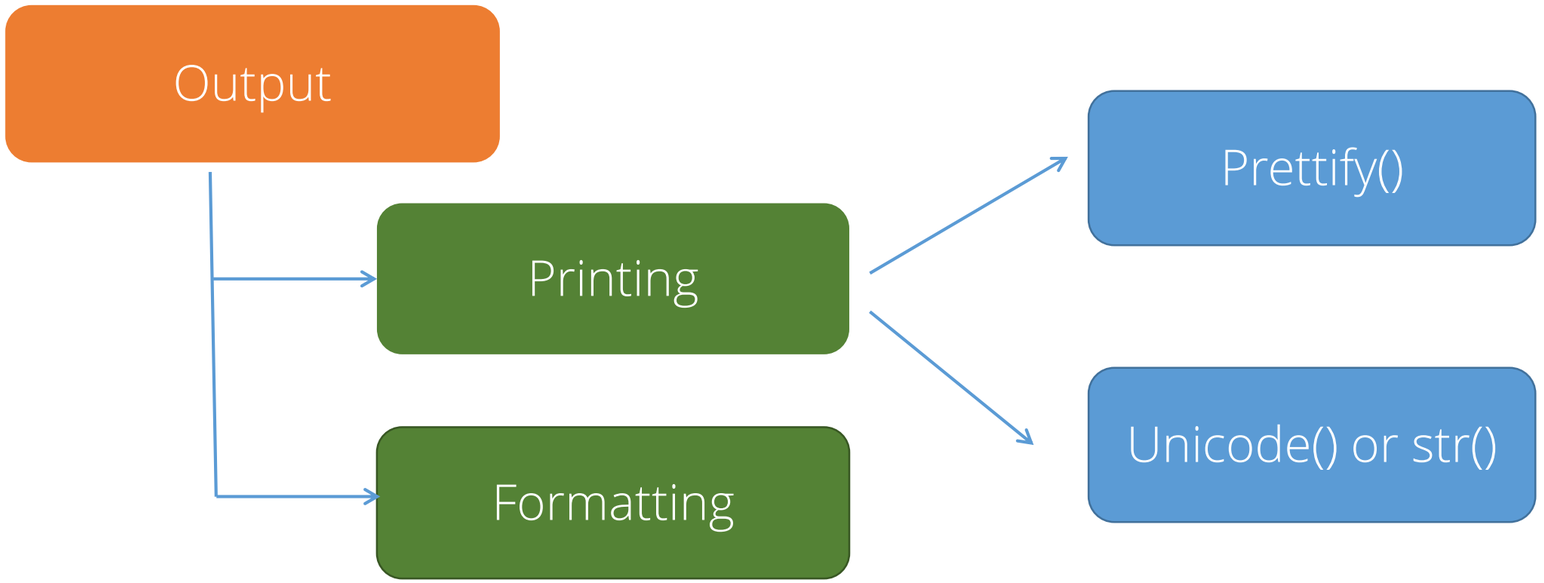
Demo: 05—Parsing part of the document

This demo shows you how to parse only a part of document with the help of an example.

DATA SCIENCE

simplilearn

Output : Printing and Formatting



Output : Printing and Formatting

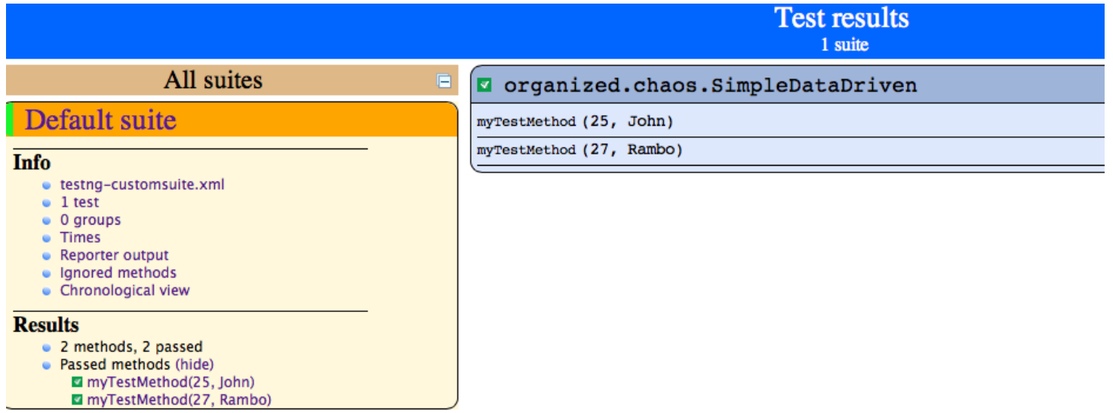
Output

Printing

Formatting

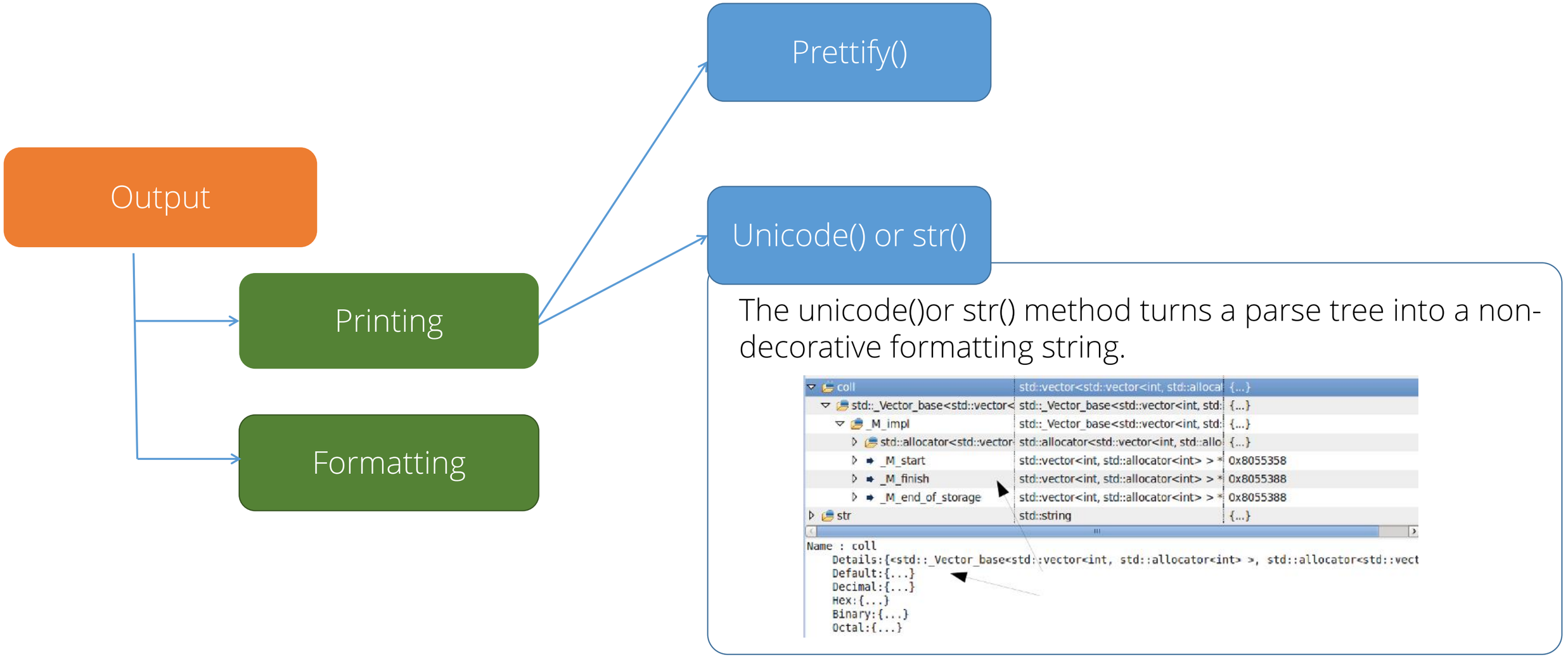
Prettify()

The prettify or pretty printing method turns a parse tree into a decorative formatted Unicode string.



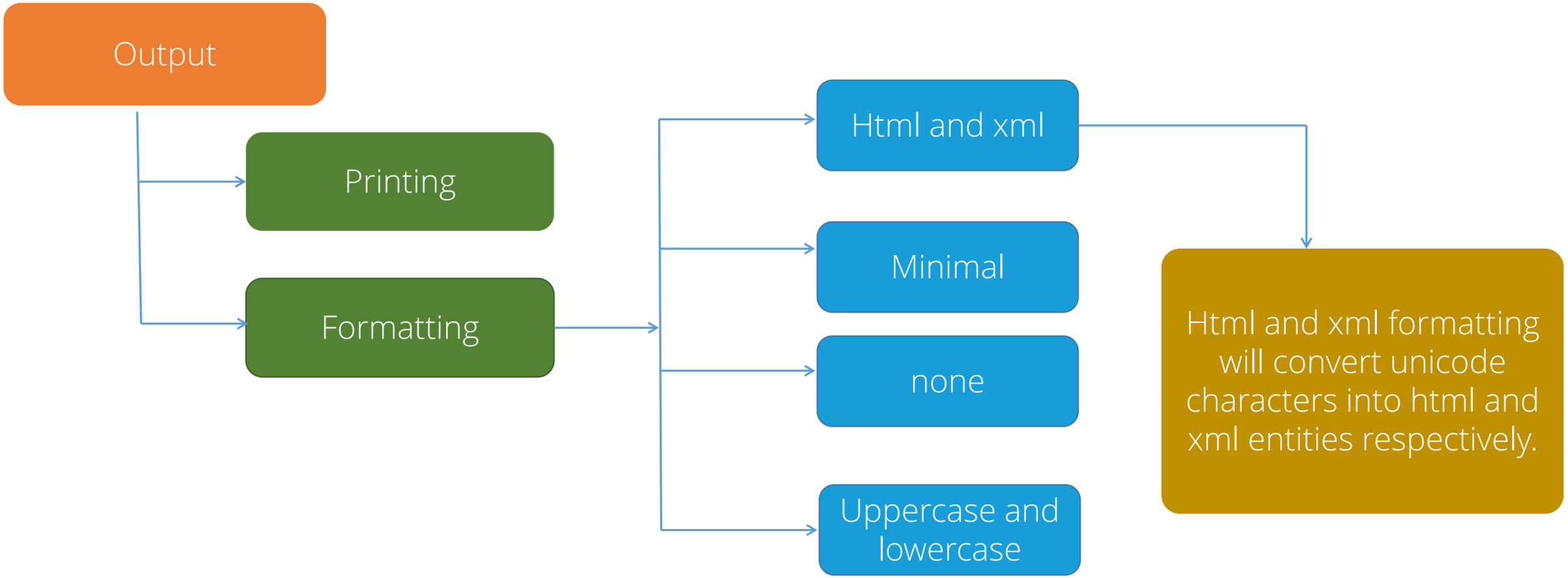
Unicode() or str()

Output : Printing and Formatting (contd.)



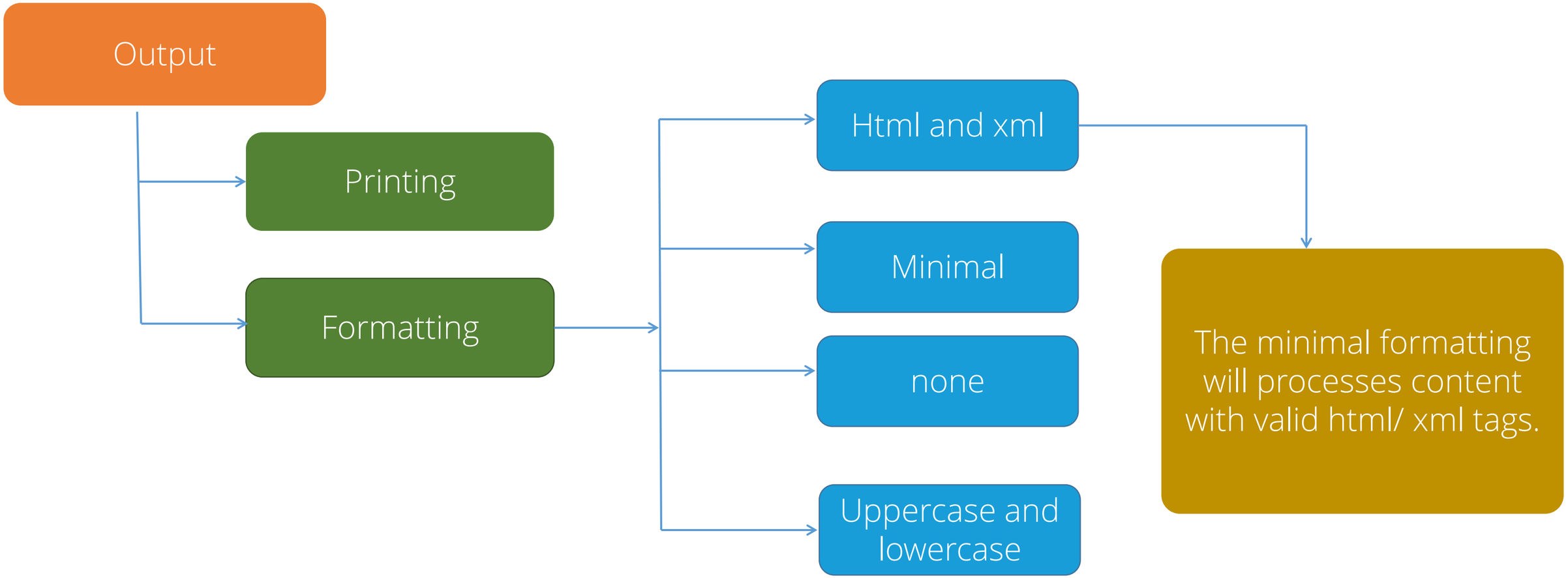
Output : Printing and Formatting (contd.)

The formatters are used to generate different types of output with the desired formatting.



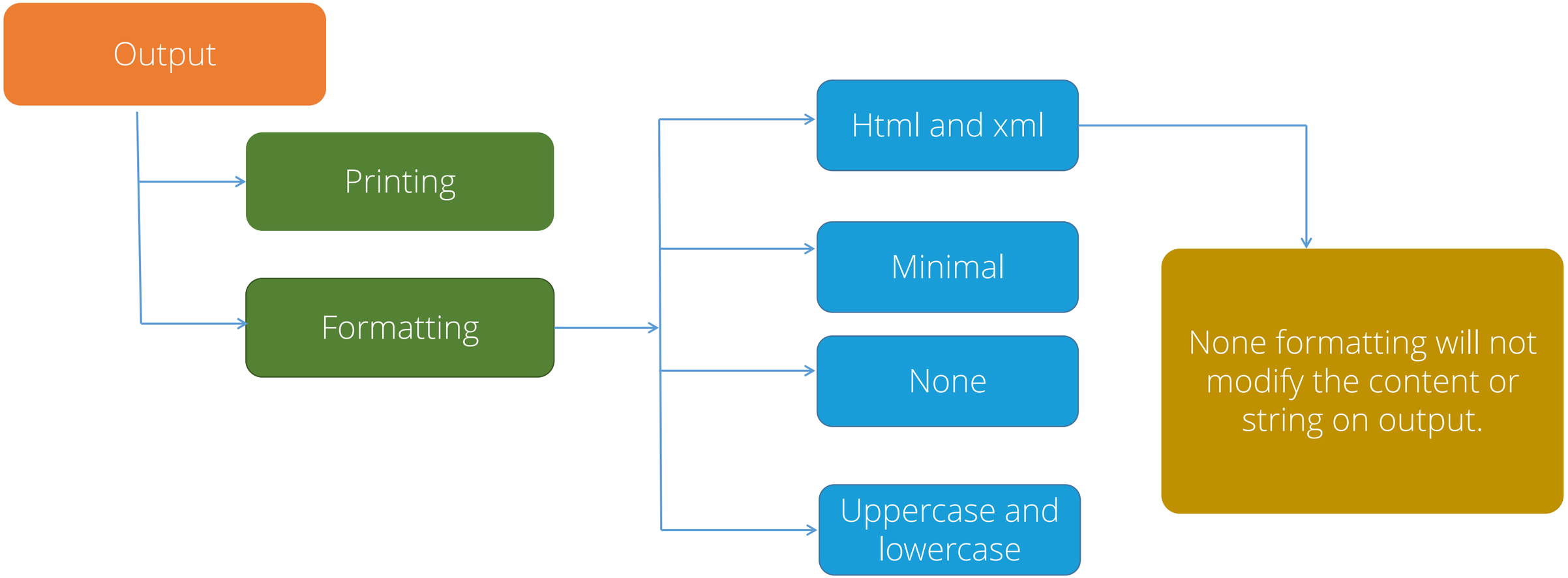
Output : Printing and Formatting (contd.)

The formatters are used to generate different types of output with the desired formatting.



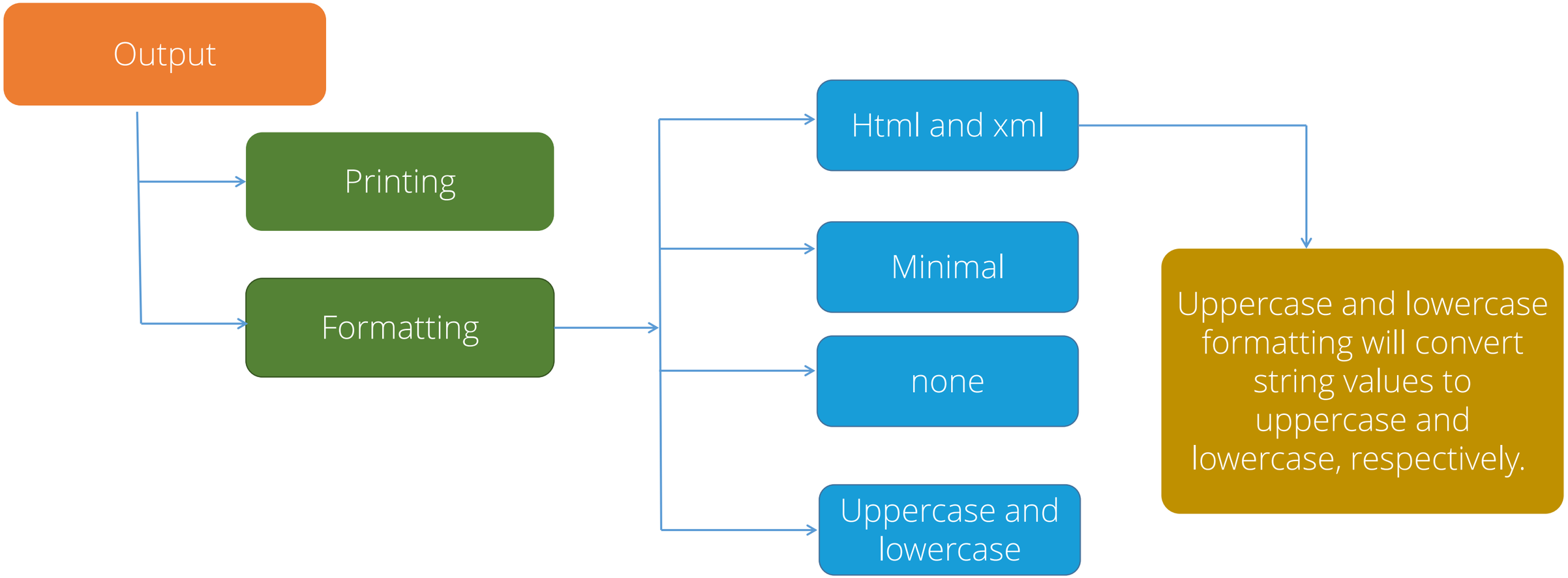
Output : Printing and Formatting (contd.)

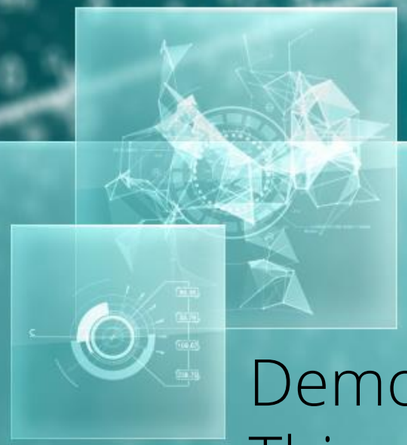
The formatters are used to generate different types of output with the desired formatting.



Output : Printing and Formatting (contd.)

The formatters are used to generate different types of output with the desired formatting.





Demo: 06—Formatting and Printing

This demo shows the ways to format, print, and encode the web document.

DATA SCIENCE

Encoding

Document Encoding

- HTML or XML documents are written in specific encodings such as, ASCII or UTF-8.
- When we load the document into BeautifulSoup, it gets converted into Unicode.
- The original encoding can be extracted from attribute .original encoding of the Beautiful Soup object.

Output Encoding

- When you write a document from Beautiful Soup, you get a UTF-8 document irrespective of the original encoding.
- If some other encoding is required, we can pass it to prettify.

Assignment

Assignment 01

Problem | Instructions

Scrape the Simplilearn website page and perform the following tasks:

- View and print the Simplilearn web page content in a proper format
- View the head and title
- Print all the href links present in the Simplilearn web page

Simplilearn website URL: http://www.simplilearn.com/

Assignment 01

Problem Instructions

Instructions to perform the assignment:

- Use Simplilearn's resource page URL in the Jupyter notebook to view and evaluate it.

Common instructions:

- If you are new to Python, download the "Anaconda Installation Instructions" document from the "Resources" tab to view the steps for installing Anaconda and the Jupyter notebook.
- Download the "Assignment 01" notebook and upload it on the Jupyter notebook to access it.
- Follow the provided cues to complete the assignment.

Assignment

Assignment 02

Problem | Instructions

Scrape the Simplilearn website resource page and perform the following tasks:

- View and print the Simplilearn web page content in a proper format
- View the head and title
- Print all the href links present in the Simplilearn web page
- Search and print the resource headers of the Simplilearn web page
- Search resource topics
- View the article names and navigate through them

Simplilearn website URL: http://www.simplilearn.com/resources

Assignment 02

Problem Instructions

Instructions to perform the assignment:

- Download the web scraping dataset from the “Resource” tab. Upload the dataset to your Jupyter notebook to view and evaluate it.

Common instructions:

- If you are new to Python, download the “Anaconda Installation Instructions” document from the “Resources” tab to view the steps for installing Anaconda and the Jupyter notebook.
- Download the “Assignment 02” notebook and upload it on the Jupyter notebook to access it.
- Follow the provided cues to complete the assignment.

Quiz

QUIZ

1

Which of the following is the only xml parser?

a. html.parser

b. lxml

c. lxml.xml

d. html5lib

QUIZ 1

Which of the following is the only xml parser?

a. html.parser

b. lxml

c. lxml.xml

d. html5lib

The correct answer is c.

Explanation: lxml.xml is the only xml parser available for BeautifulSoup object.

QUIZ 2

In which of the following formats is the BeautifulSoup output encoded?

a. ASCII

b. Unicode

c. latin-1

d. UTF-8

QUIZ 2

In which of the following formats id the BeautifulSoup output encoded?

a. ASCII

b. Unicode

c. latin-1

d. UTF-8

The correct answer is **d** .

Explanation: The output of the BeautifulSoup is always UTF-8 encoded.

QUIZ

3

Which of the following libraries is used to extract a web page?

a. Beautiful Soup

b. Pandas

c. Requests

d. NumPy

QUIZ 3

Which of the following libraries is used to extract a web page?

a. Beautiful Soup

b. Pandas

c. Requests

d. NumPy

The correct answer is **c** .

Explanation: Requests is the right API to extract the web page.

QUIZ

4

Which of the following is NOT an object in BeautifulSoup?

a. Tag

b. NextSibling

c. NavigableString

d. Comment

QUIZ

4

Which of the following is NOT an object in BeautifulSoup?

a. Tag

b. Next sibling

c. NavigableString

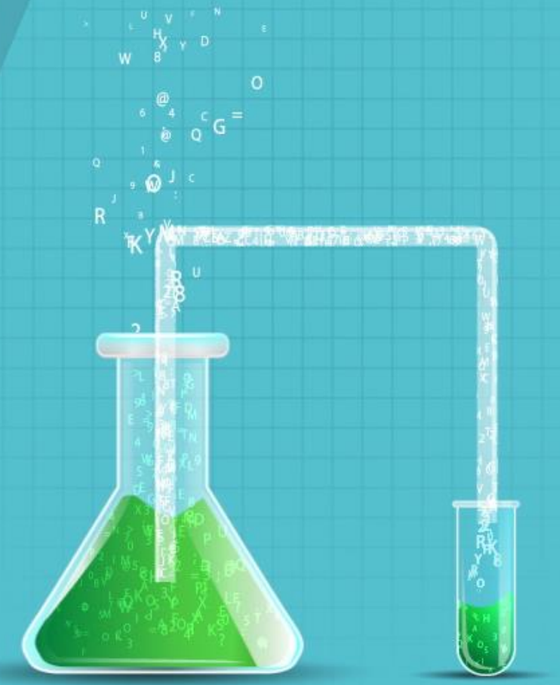
d. Comment

The correct answer is **b** .

Explanation: NextSibling is a navigation method.

Key Takeaways

- Web scraping is a computer software technique of extracting information from websites in an automated fashion.
- A Parser is a basic tool to interpret or render the information from a web document.
- Objects are used to extract the required information from a tree structure by searching or navigating through the parsed document.
- A tree can be defined as a collection of simple and complex objects.
- BeautifulSoup transforms a complex HTML document into a complex tree of Python objects.



This concludes “Web Scraping with BeautifulSoup”

The next lesson is “Python integration with Hadoop, MapReduce, and Spark”