Working with the Parse Tree in Beautiful Soup

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

The HTML tree structure

Getting started with Beautiful Soup

Understanding tags, attributes, NavigableStrings and comments

Using filters with tags, attribute values, regular expressions and functions

Extracting links from documents

Using Soup Strainer to parse just parts of a document

The HTML Parse Tree

HTML

*H*yper*t*ext *M*arkup *L*anguage is the standard markup language for documents designed to be displayed in a web browser.

```
<html>

  <head>
    <title>Page Title</title>
  </head>

  <body>
    <h1>Page Header</h1>
    <a href=“https:/www.pluralsight.com”>Some Link</a>
  </body>

</html>
```

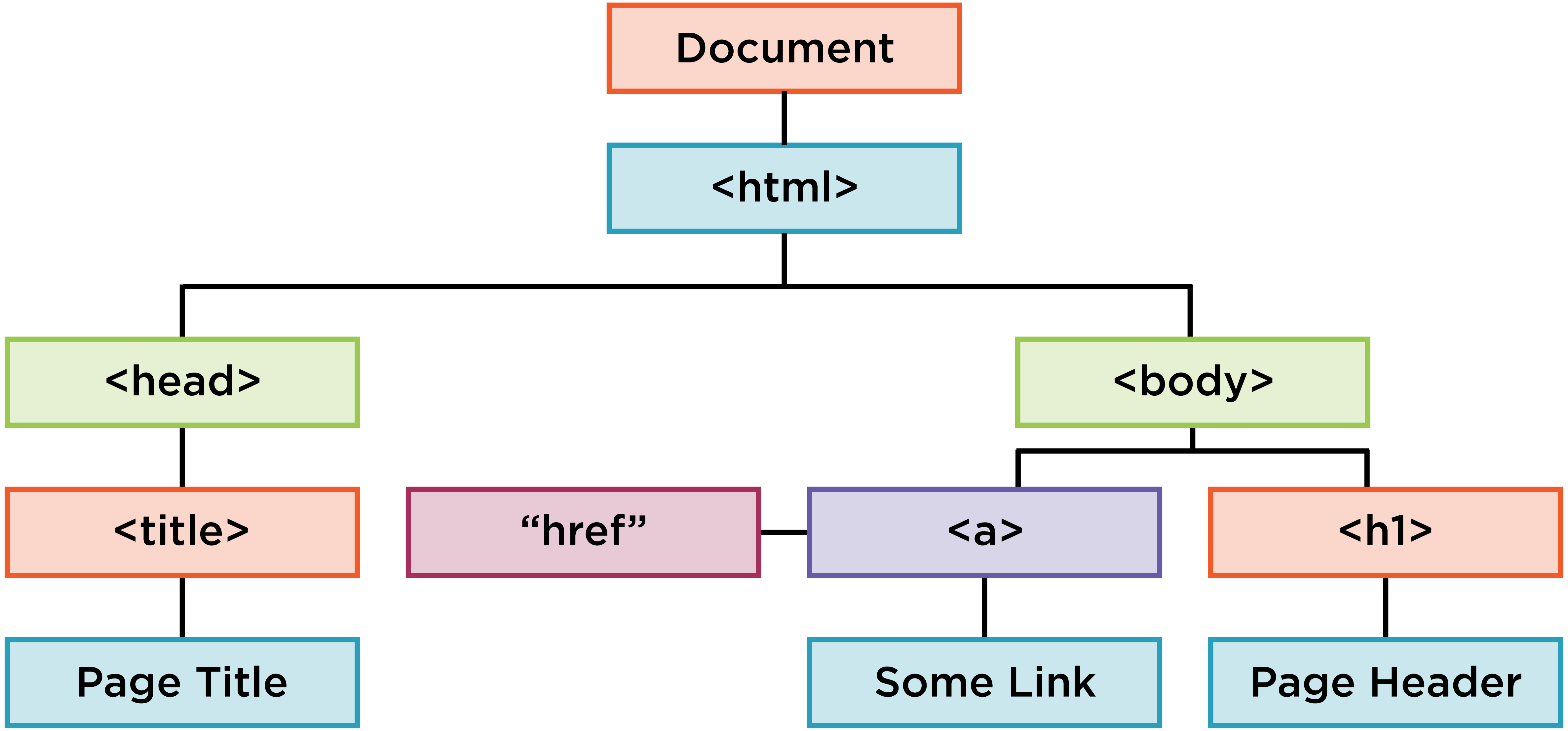
```
<html>

  <head>
    <title>Page Title</title>
  </head>

  <body>
    <h1>Page Header</h1>
    <a href=“https:/www.pluralsight.com”>Some Link</a>
  </body>

</html>
```

```
<html>

  <head>
    <title>Page Title</title>
  </head>

  <body>
    <h1>Page Header</h1>
    <a href=“https:/www.pluralsight.com”>Some Link</a>
  </body>

</html>
```

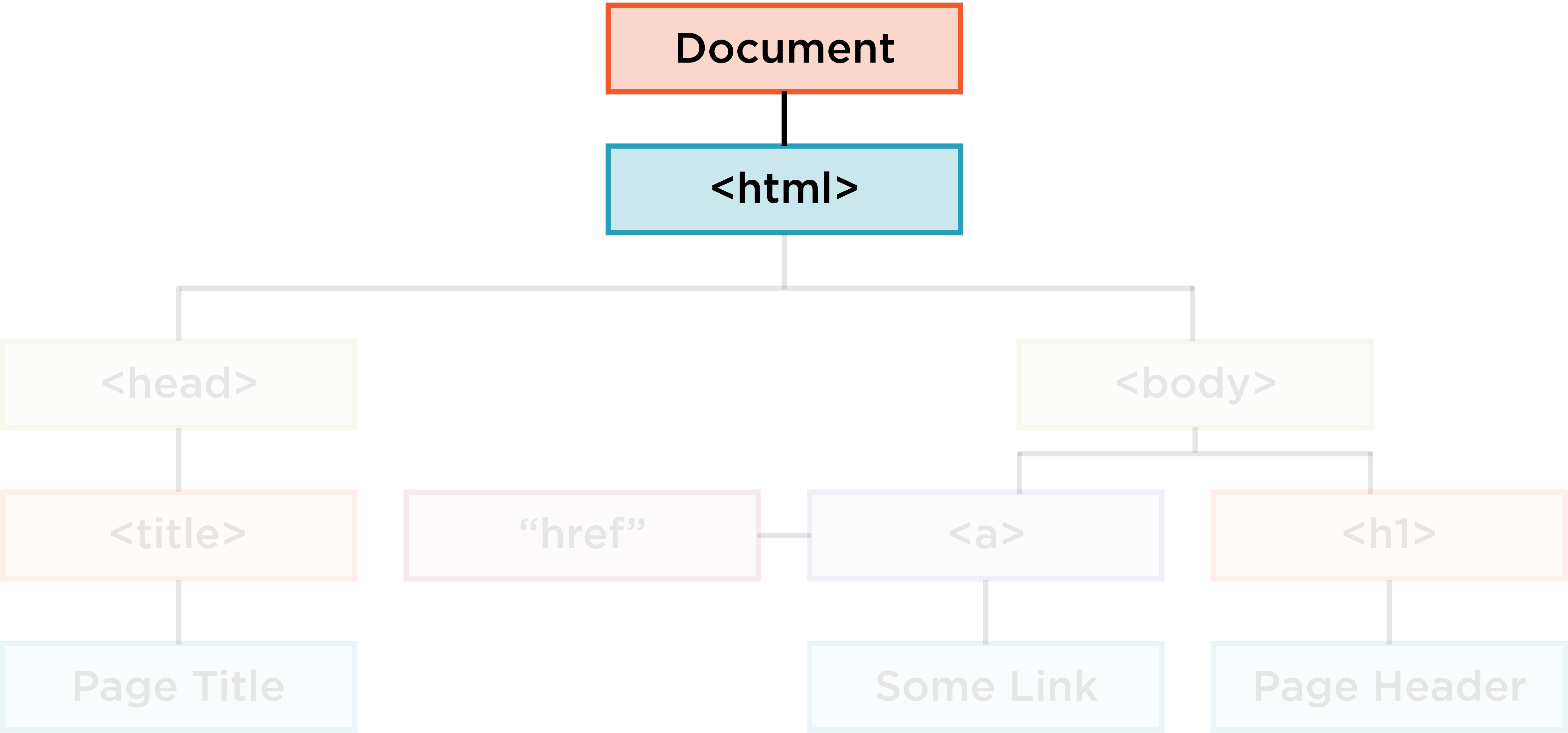
```
<html>

  <head>
    <title>Page Title</title>
  </head>

  <body>
    <h1>Page Header</h1>
    <a href=“https:/www.pluralsight.com”>Some Link</a>
  </body>

</html>
```

```
<html>

  <head>
    <title>Page Title</title>
  </head>

  <body>
    <h1>Page Header</h1>
    <a href=“https:/www.pluralsight.com”>Some Link</a>
  </body>

</html>
```

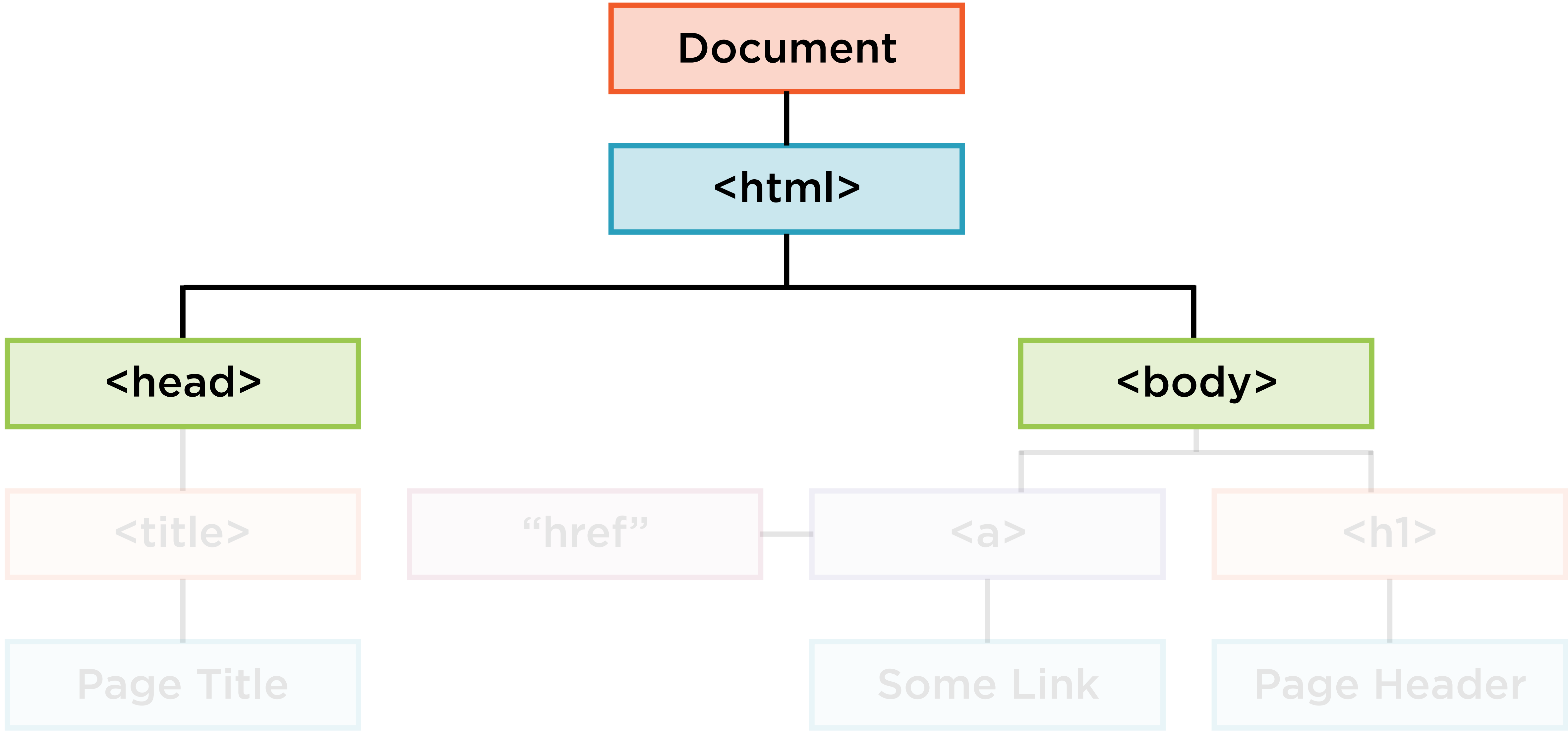
HTML Tree Structure



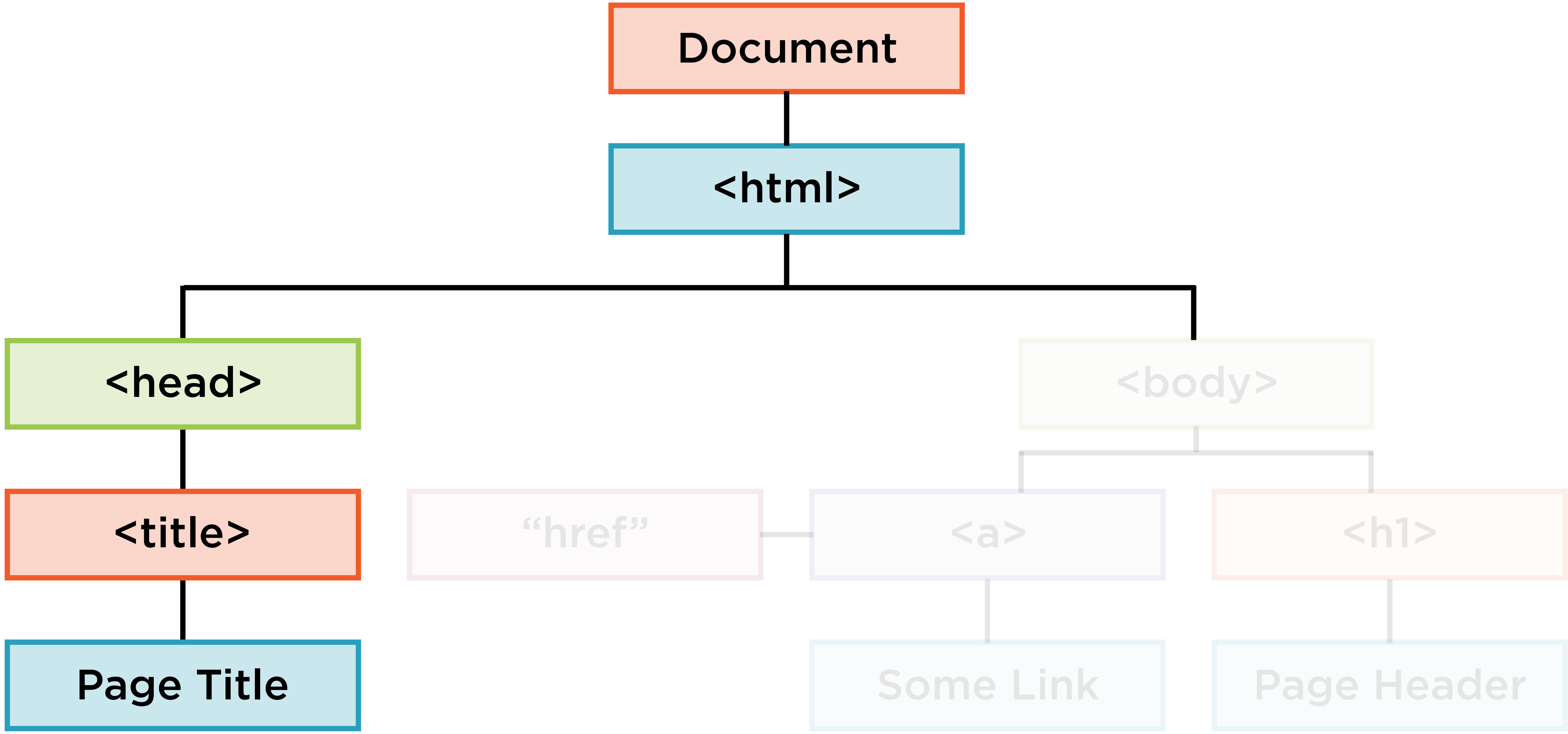
HTML Tree Structure



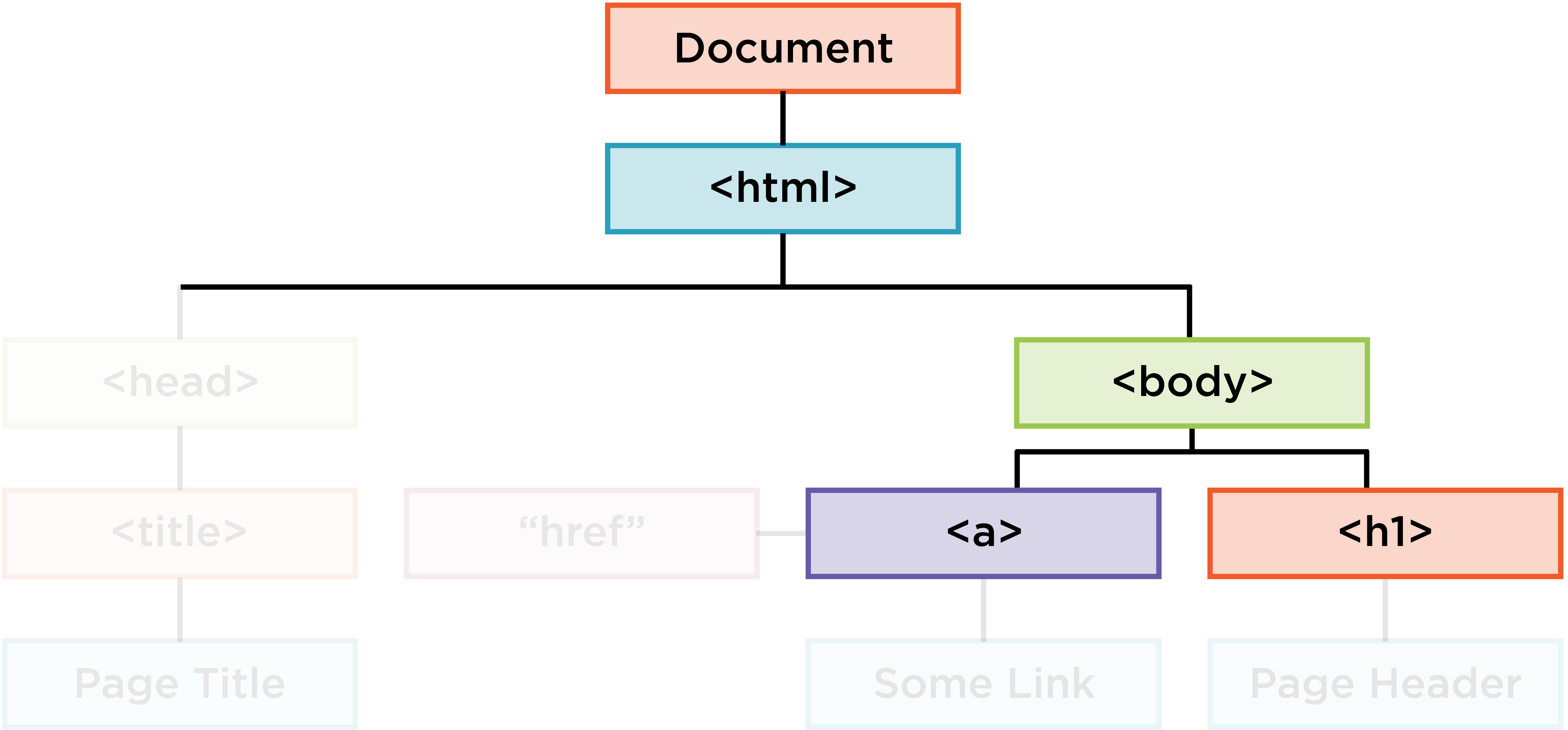
HTML Tree Structure



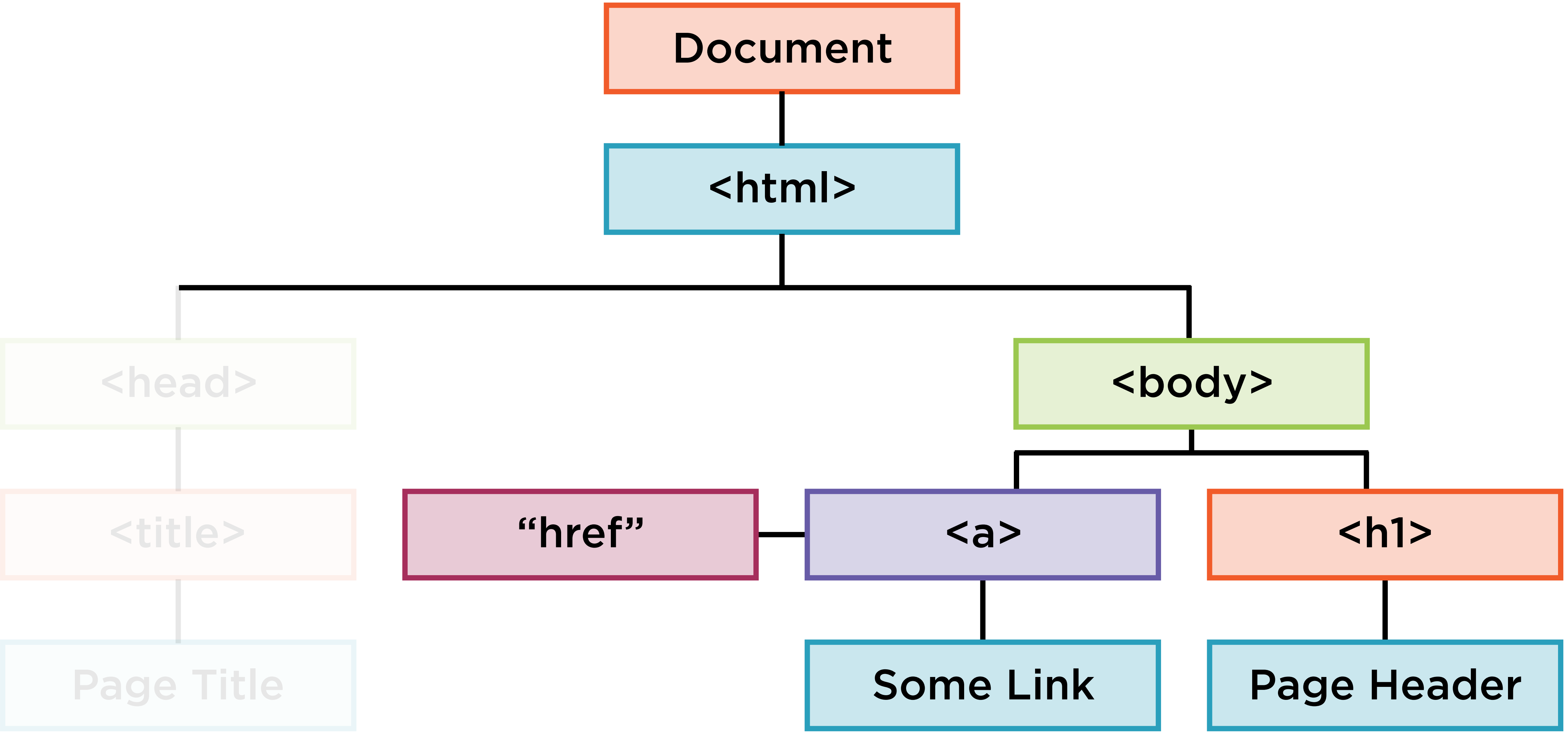
HTML Tree Structure



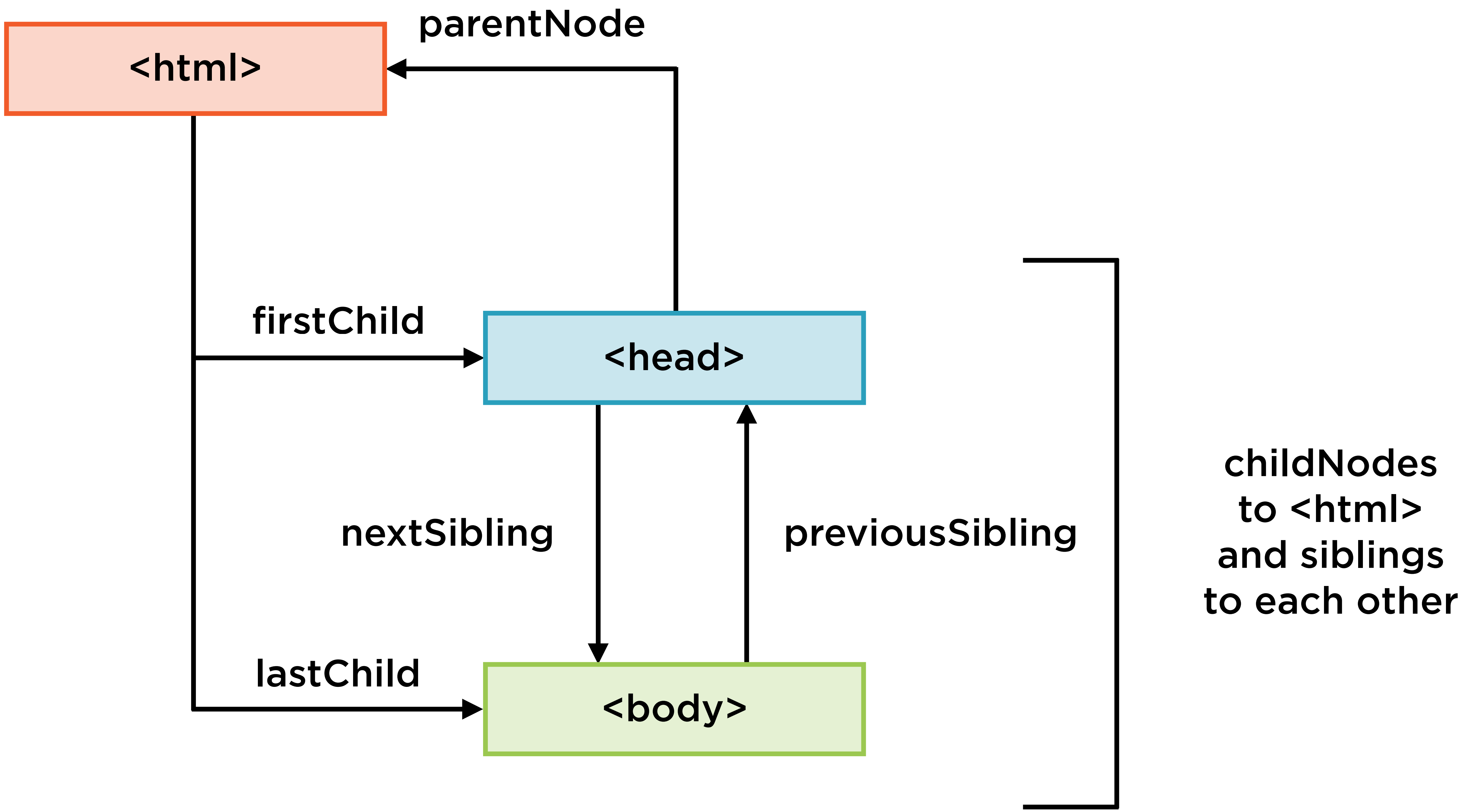
HTML Tree Structure



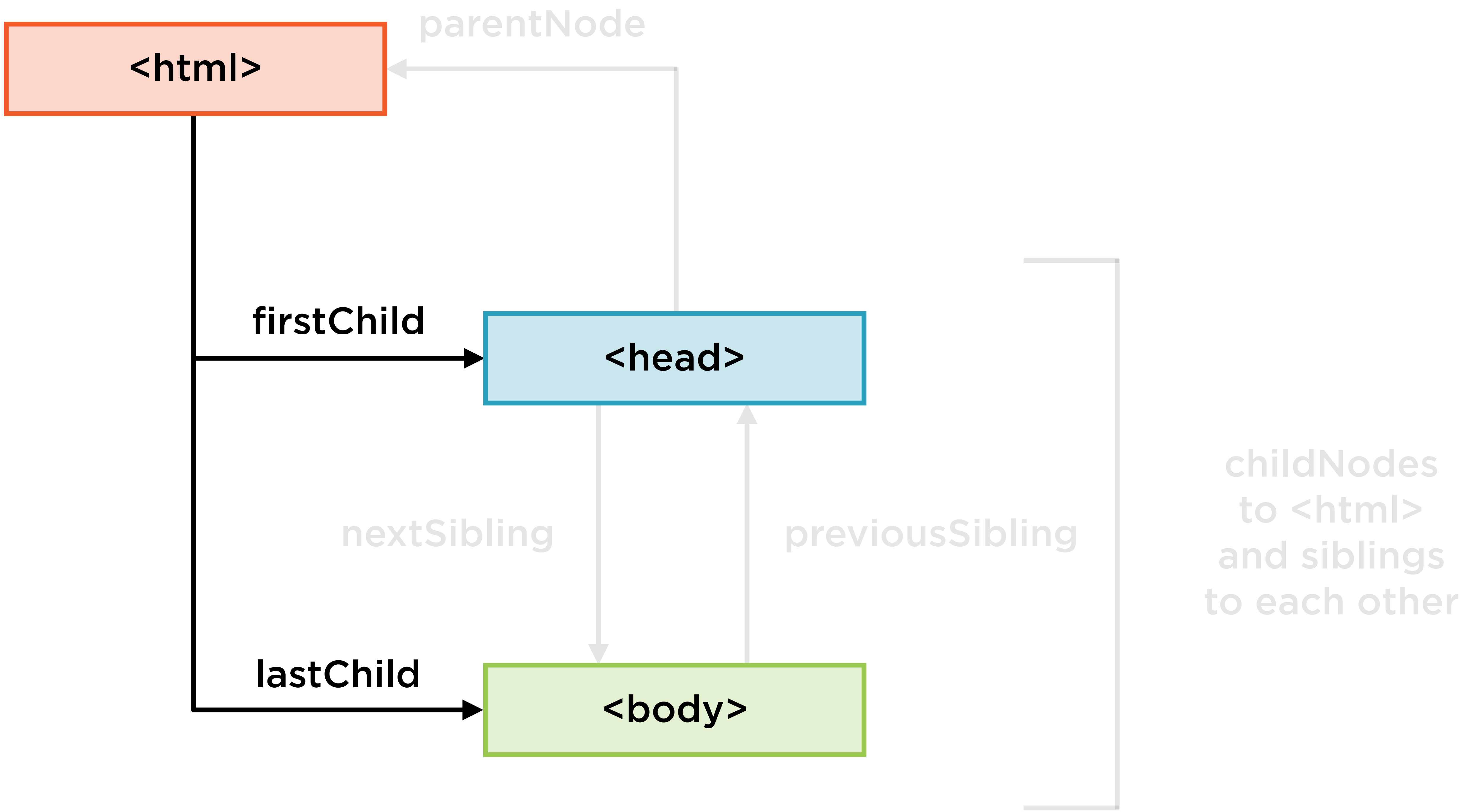
HTML Tree Structure



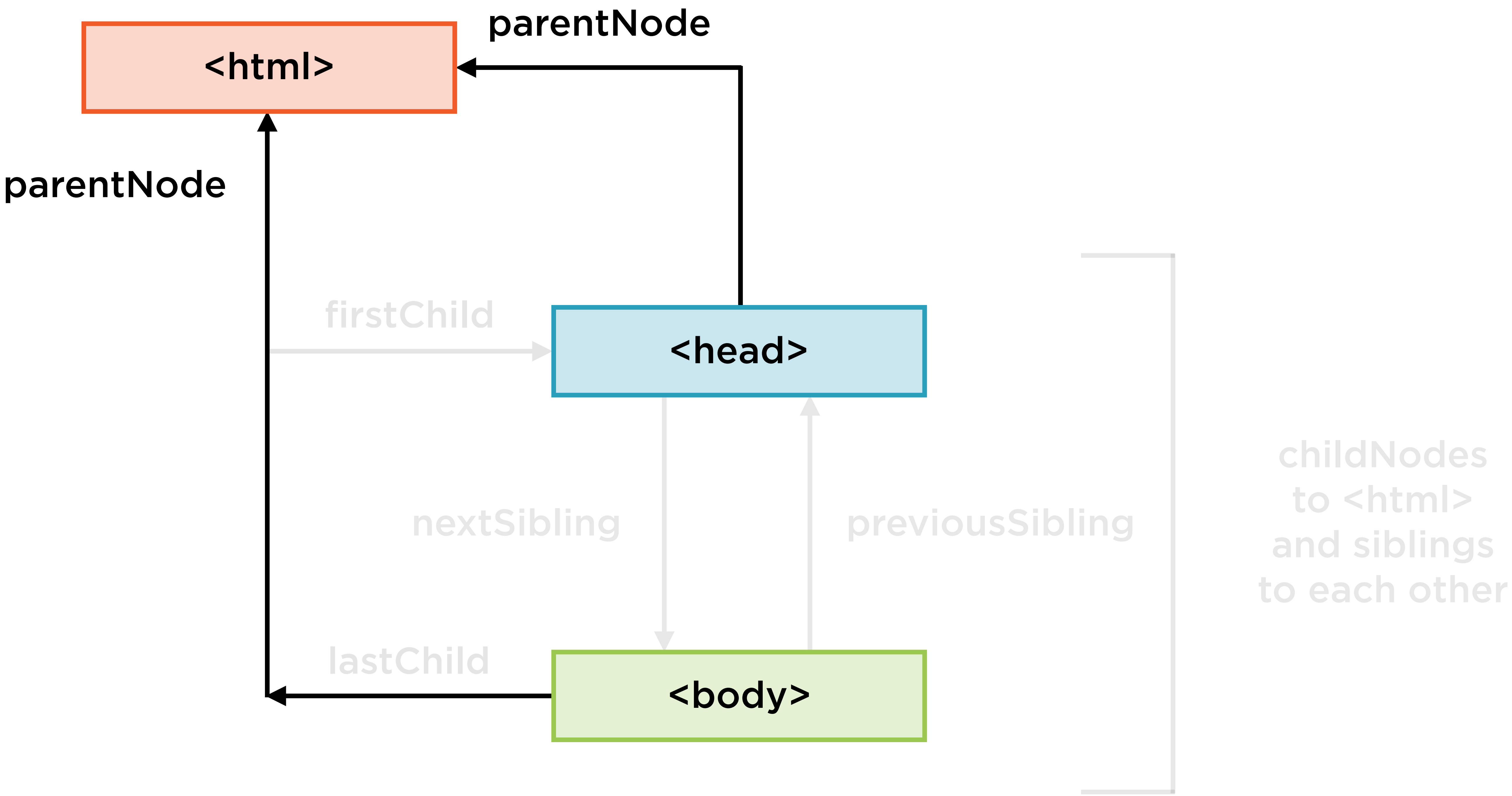
Hierarchical Relationship



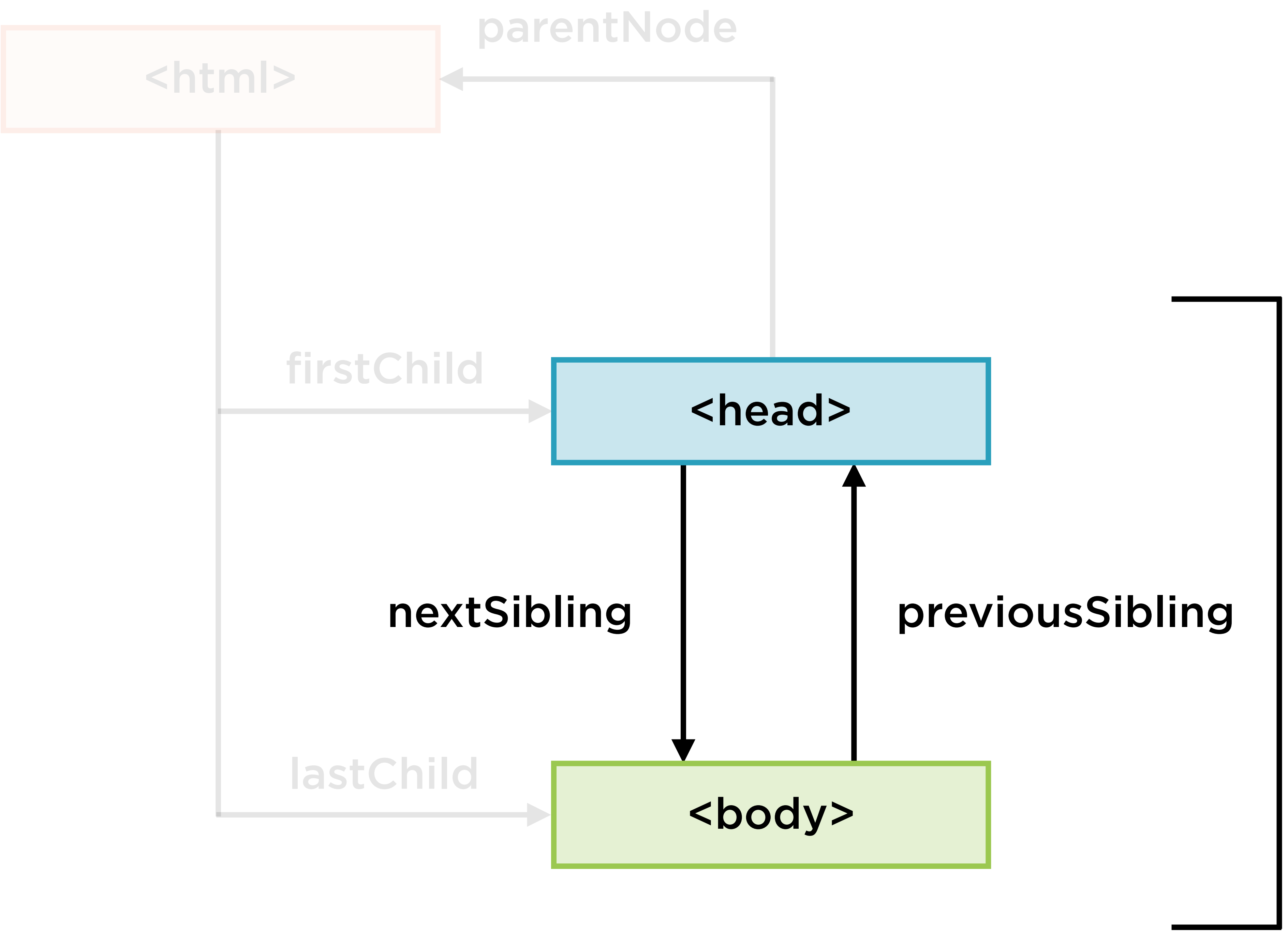
Hierarchical Relationship



Hierarchical Relationship

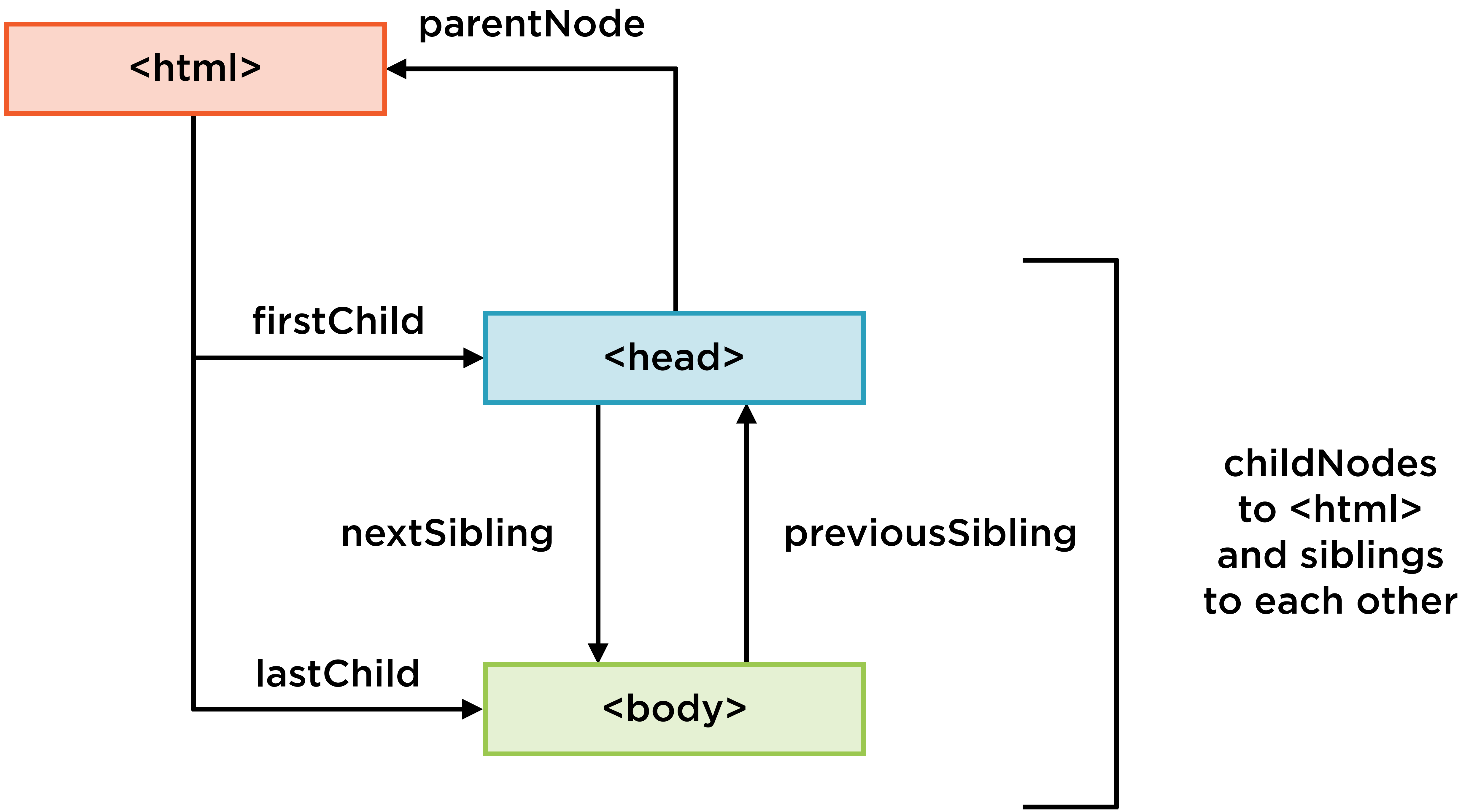


Hierarchical Relationship

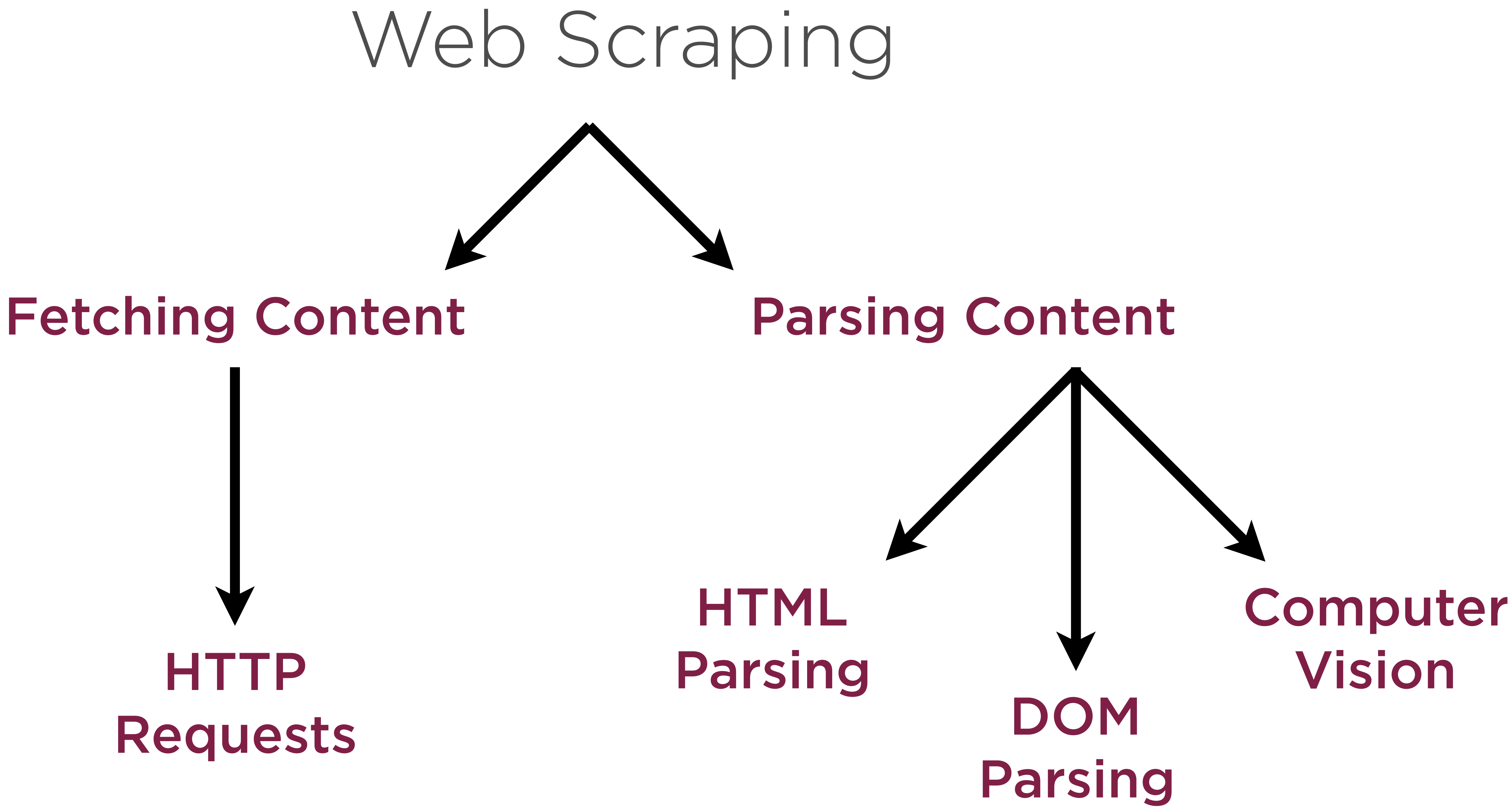


childNodes
to <html>
and siblings
to each other

Hierarchical Relationship



Parsing HTML Content



Web Scraping

Fetching Content

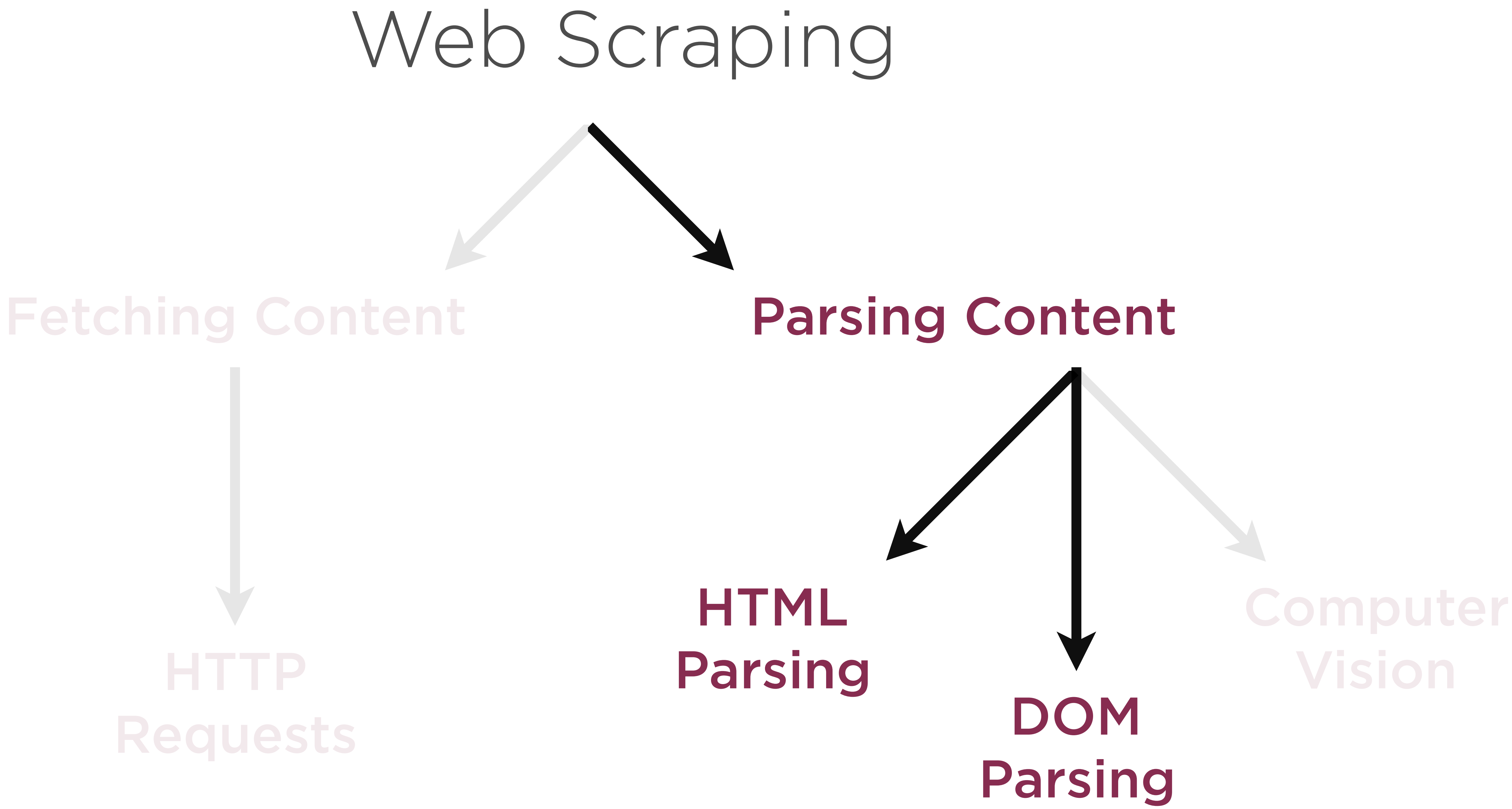
Parsing Content

HTTP Requests

HTML Parsing

DOM Parsing

Computer Vision



Web Scraping

Fetching Content

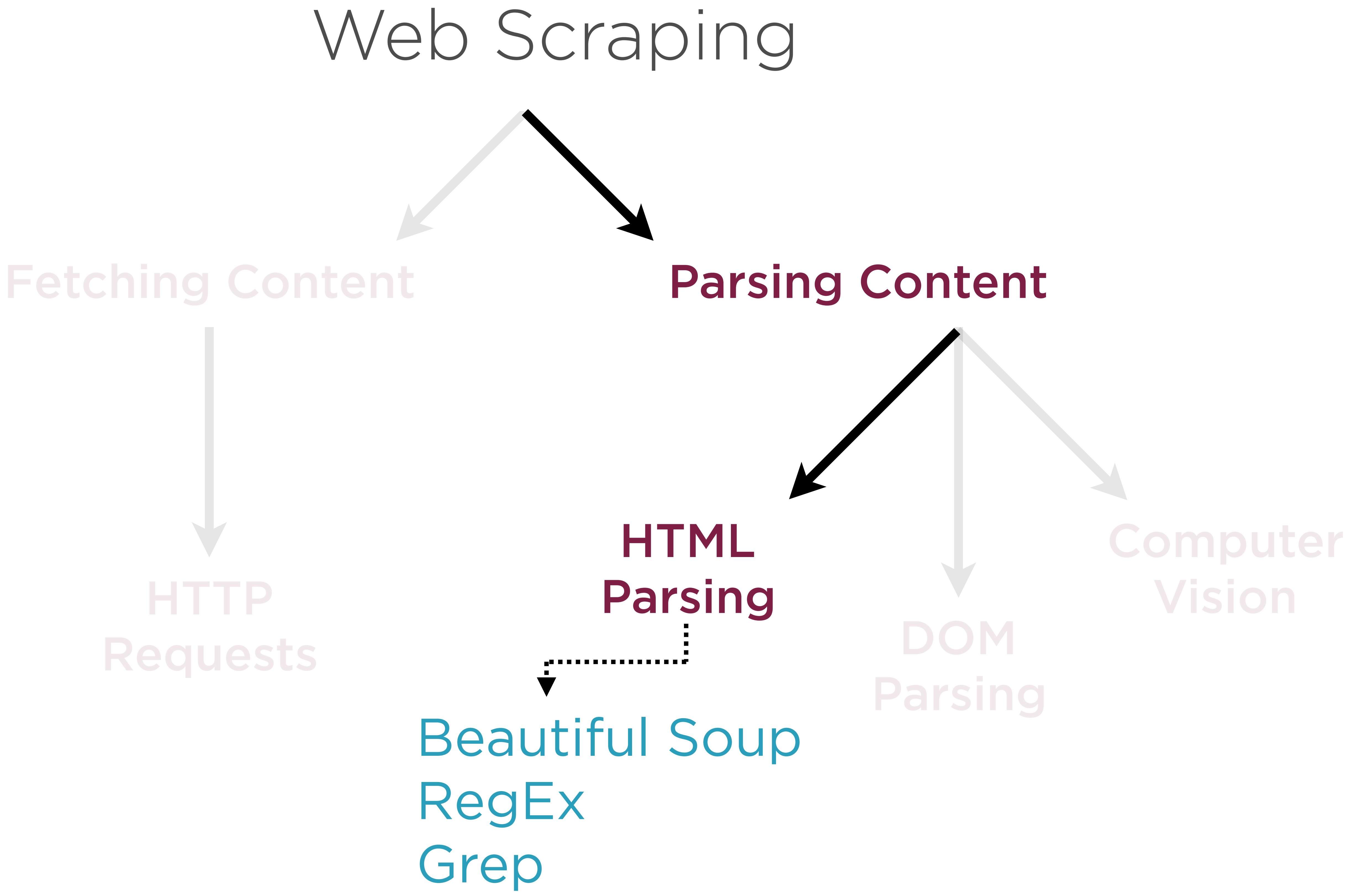
Parsing Content

HTTP Requests

HTML Parsing

DOM Parsing

Computer Vision



Web Scraping

Fetching Content

Parsing Content

HTTP Requests

HTML Parsing

DOM Parsing

Computer Vision

Beautiful Soup

RegEx

Grep

Beautiful Soup

Python package for parsing HTML and XML, including those with malformed markup such as missing tags.

Beautiful Soup

Mitigates weaknesses of regular expressions

- Global: Forms parse tree of entire HTML
- Relatively simple to use
- Robust to problems in markup being parsed

Demo

Introducing BeautifulSoup

Parsing HTML using multiple parsers

Demo

Extracting specific page elements with Beautiful Soup

Demo

Using find() and findAll() to search for and filter elements on an HTML page

Demo

Extracting links from a web page

Demo

Using SoupStrainer to parse a subset of a document

Summary

The HTML tree structure

Getting started with Beautiful Soup

Understanding tags, attributes, NavigableStrings and comments

Using filters with tags, attribute values, regular expressions and functions

Extracting links from documents

Using Soup Strainer to parse just parts of a document