



Dr Gabriele Salciute Civiliene

gabriele.salciute-civiliene@kcl.ac.uk

Coding & the Humanities

Week 8 | Part 1_ **HTML DOCUMENTS**

30/11/2020

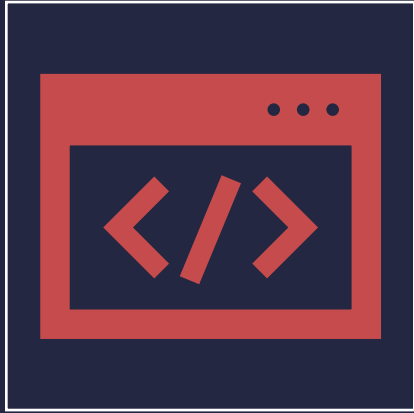
In this part ...

- You'll learn about the difference between scraping and crawling
- You'll learn what makes up the basic structure of HTML docs



Data (web) scraping vs. Data (web) crawling

Scraping	Crawling
Extracting data from sources, including the web	Iterative finding & fetching of web links from seed URLs
Any scale	Large scale
Can be done without crawling	Involves a degree of web scraping
Can be manual or programmatic	Needs crawler or bot



HTML

(HyperText Markup
Language)

Mark up language to structure content for web pages

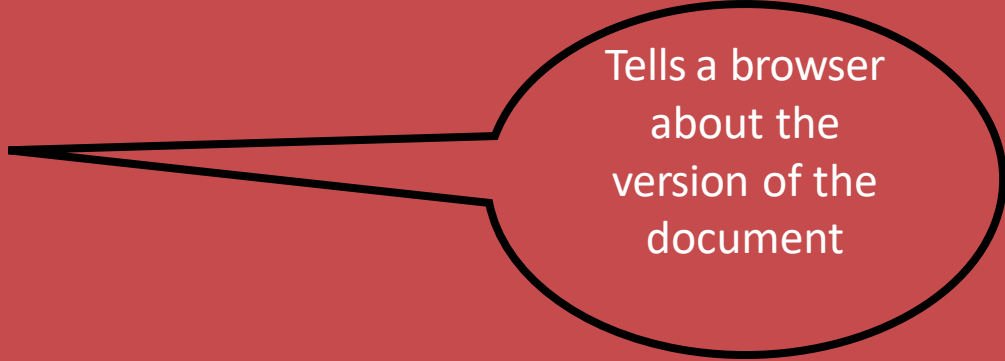
Can do a bit of formatting, but CSS is the major language to add style to structured content

HTML documents are highly structured hierarchies with nested tags that wrap up content

Element Tags in HTML5

<!DOCTYPE html>

<html>



Tells a browser
about the
version of the
document

</html>

Element Tags in HTML5

`<!DOCTYPE html>`

Tells a browser
about the version
of HTML, but is
not an HTML tag

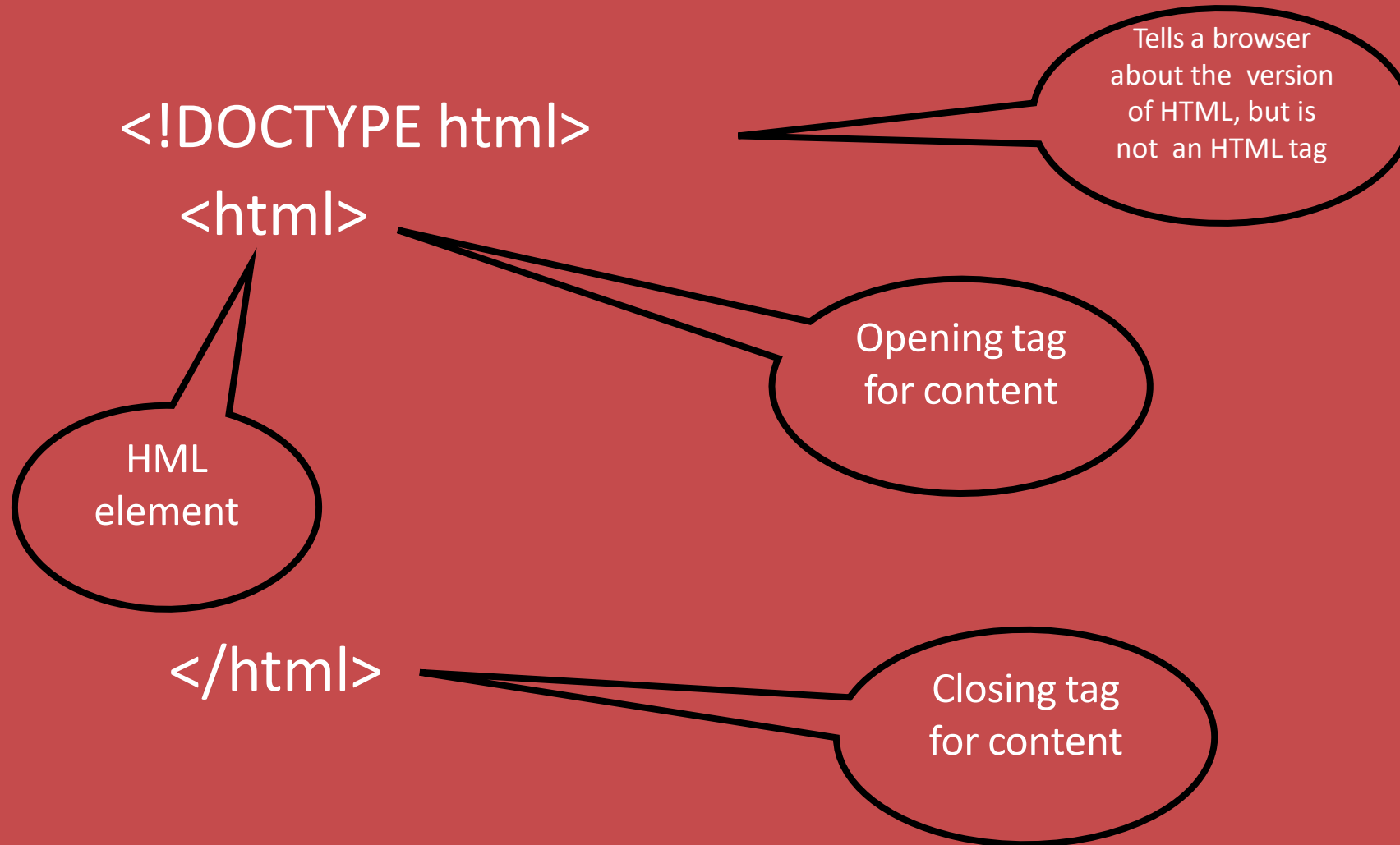
`<html>`

Opening tag
for content

`</html>`

Closing tag
for content

Element Tags in HTML5



Element Tags in HTML5

`<!DOCTYPE html>`

Tells a browser
about the version
of HTML, but is
not an HTML tag

`<html>`

Opening tag
for content

**CONTENT &
METADATA**

`</html>`

Closing tag
for content

Top-level Structure of HTML5

```
<!DOCTYPE html>
```

```
<html>
```

```
<head>
```

```
</head>
```

```
<body>
```

```
</body>
```

```
</html>
```



The diagram illustrates the top-level structure of an HTML5 document. It shows a sequence of HTML tags: `<!DOCTYPE html>`, `<html>`, `<head>`, `</head>`, `<body>`, `</body>`, and `</html>`. The `<body>` and `</body>` tags are enclosed in a black oval. A callout line extends from this oval to a speech bubble containing the text "CONTENT DISPLAYABLE IN A BROWSER".

CONTENT
DISPLAYABLE
IN A
BROWSER

Top-level Structure of HTML5

<!DOCTYPE html>

<html>

<head>

</head>

<body>

</body>

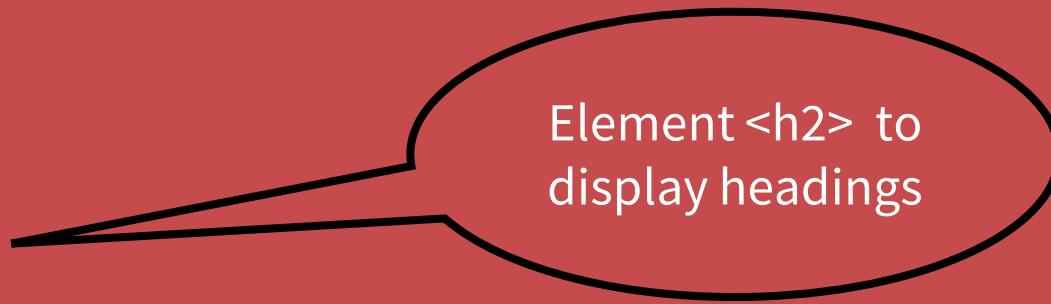
</html>

METADATA
THAT SEARCH
ENGINES CAN
'SEE'

More Structure: Other Elements

<body>

<h2> This is a heading </h2>



Element <h2> to
display headings

<p> This is a good day to learn HTML and web scraping</p>

</body>

More Structure: Other Elements

<body>

<h2> This is a heading </h2>

<p> This is a good day to learn HTML and web scraping</p>

</body>



Element <p> to
display
paragraphs

More Structure: Other Elements

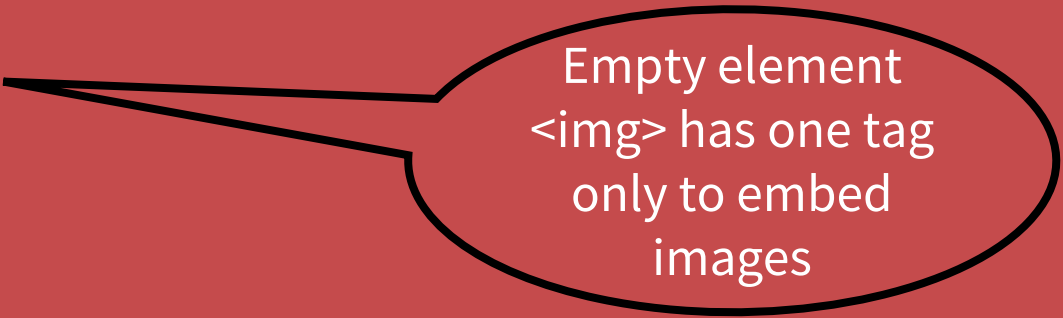
```
<body>
```

```
<h2> This is a heading </h2>
```

```
<p> This is a good day to learn HTML and web scraping</p>
```

```

```



Empty element
 has one tag
only to embed
images

```
</body>
```

More Structure: Other Elements

<body>

<h2> This is a heading </h2>

<p> This is a good day to learn HTML and web scraping</p>



Attribute
src

</body>

More Structure: Other Elements

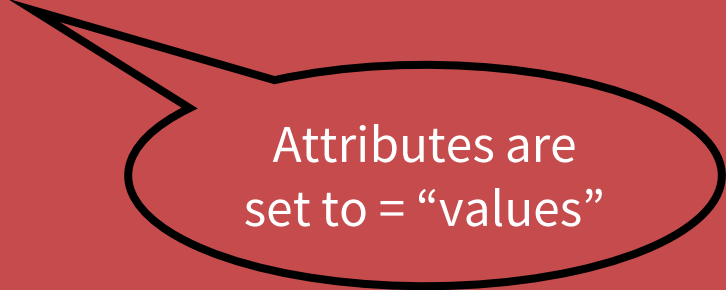
```
<body>
```

```
<h2> This is a heading </h2>
```

```
<p> This is a good day to learn HTML and web scraping</p>
```

```

```



Attributes are
set to = "values"

```
</body>
```

Attributes & Values

```
<!DOCTYPE html>
```

```
<html>
```

```
<head>
```

```
</head>
```

```
<body>
```

```
<a href = "https://www.w3schools.com"> Visit W3Schools.com!</a>
```

```
</body>
```

```
</html>
```

HTML element

<a>

Attribute href

Value of an
attribute



How do we access HTML content?

We can scrape web content with Python or another programming languages by:

HTML element, e.g. `<p>`

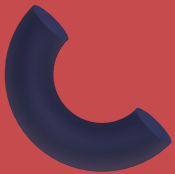
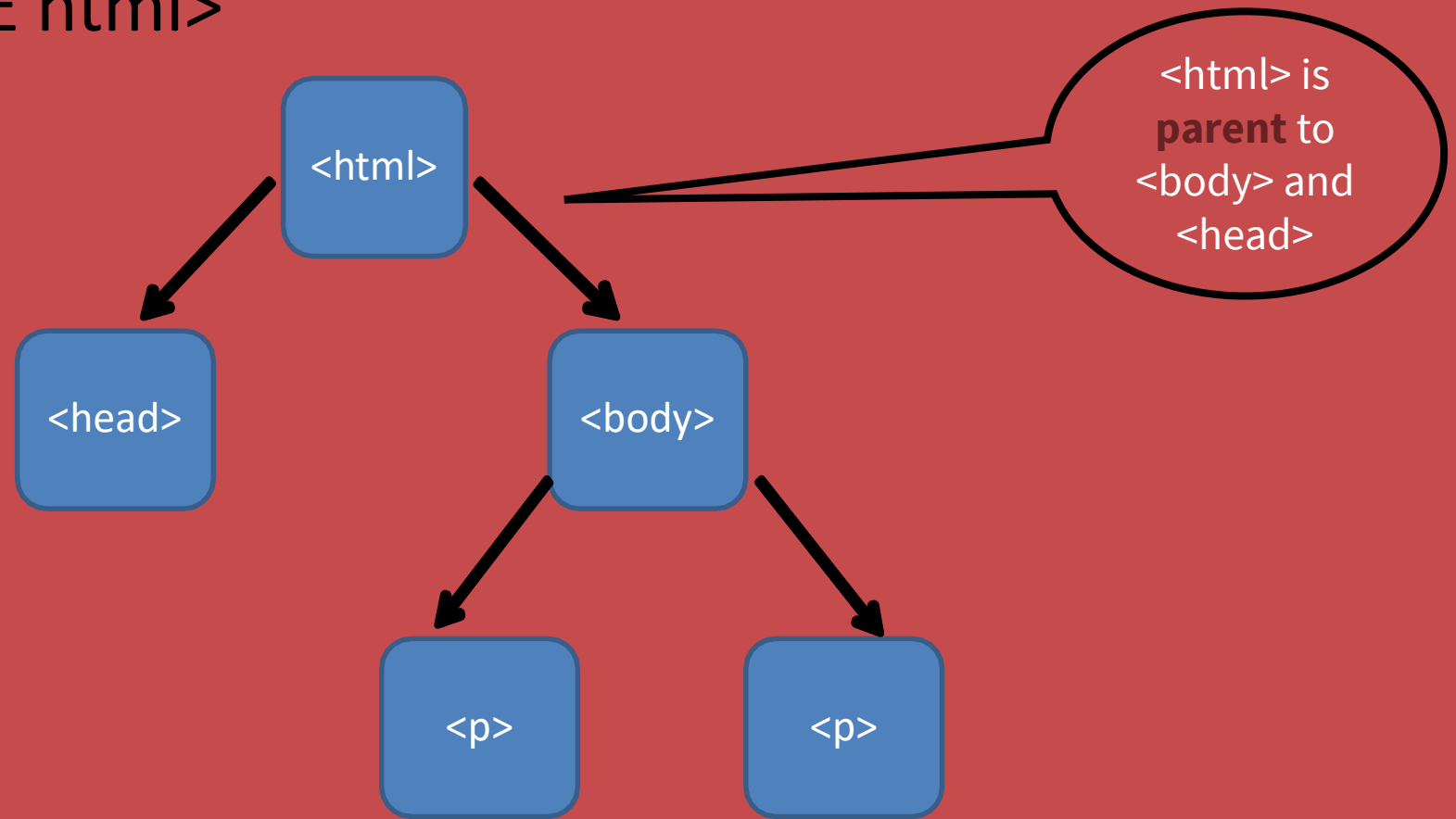
Element's attributes, e.g. `<p class="first">`

Attribute's values, e.g. `"first"`

Family relationships between HTML elements, e.g. `<p>` may be a `child` or `parent` to other elements

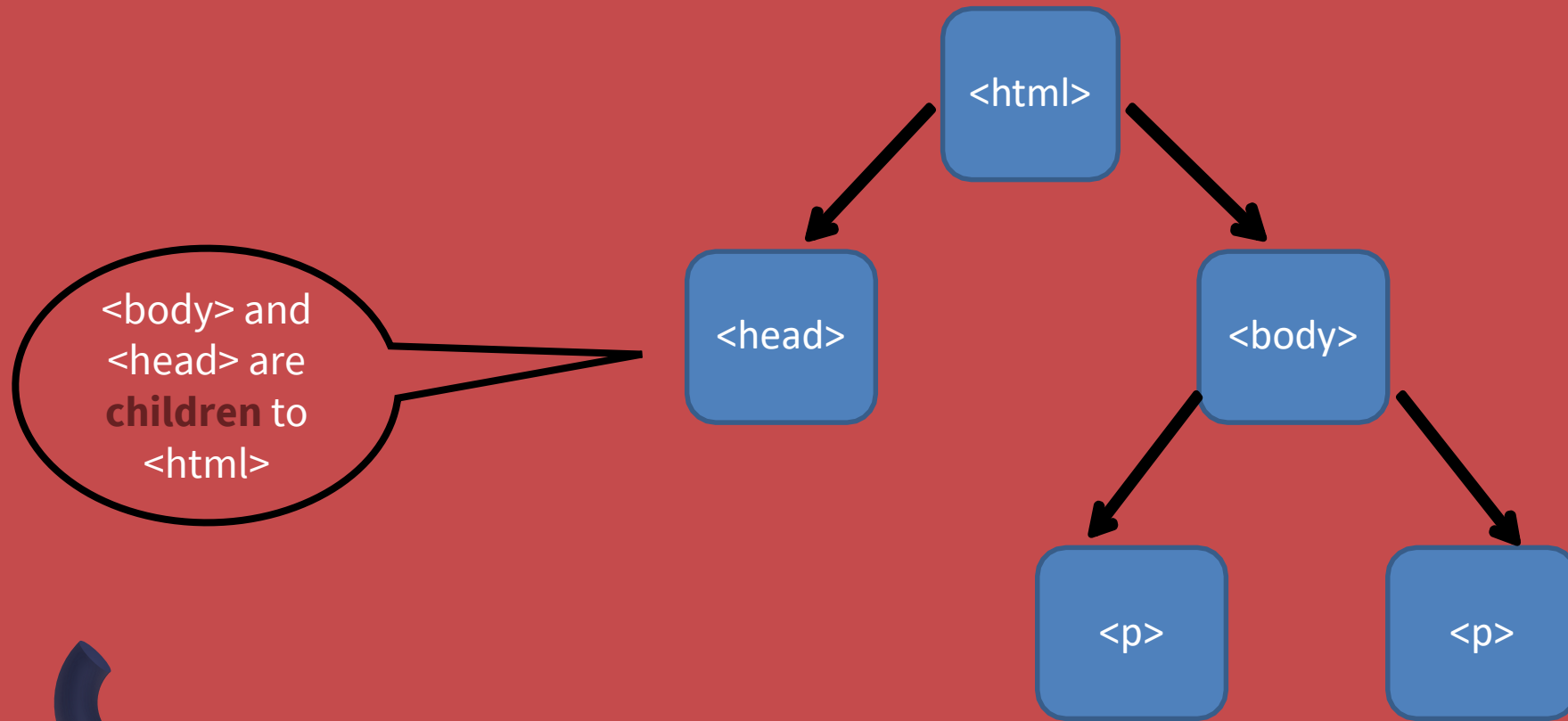
Parent-child-sibling relationships

<!DOCTYPE html>



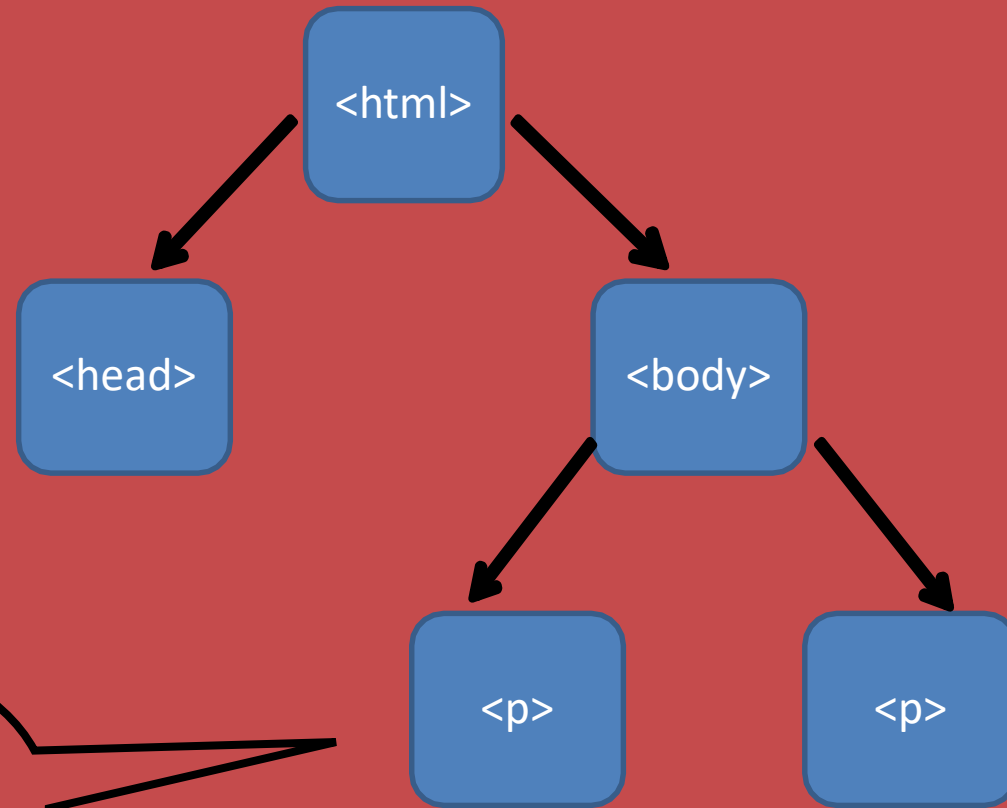
Parent-child-sibling relationships

<!DOCTYPE html>



Parent-child-sibling relationships

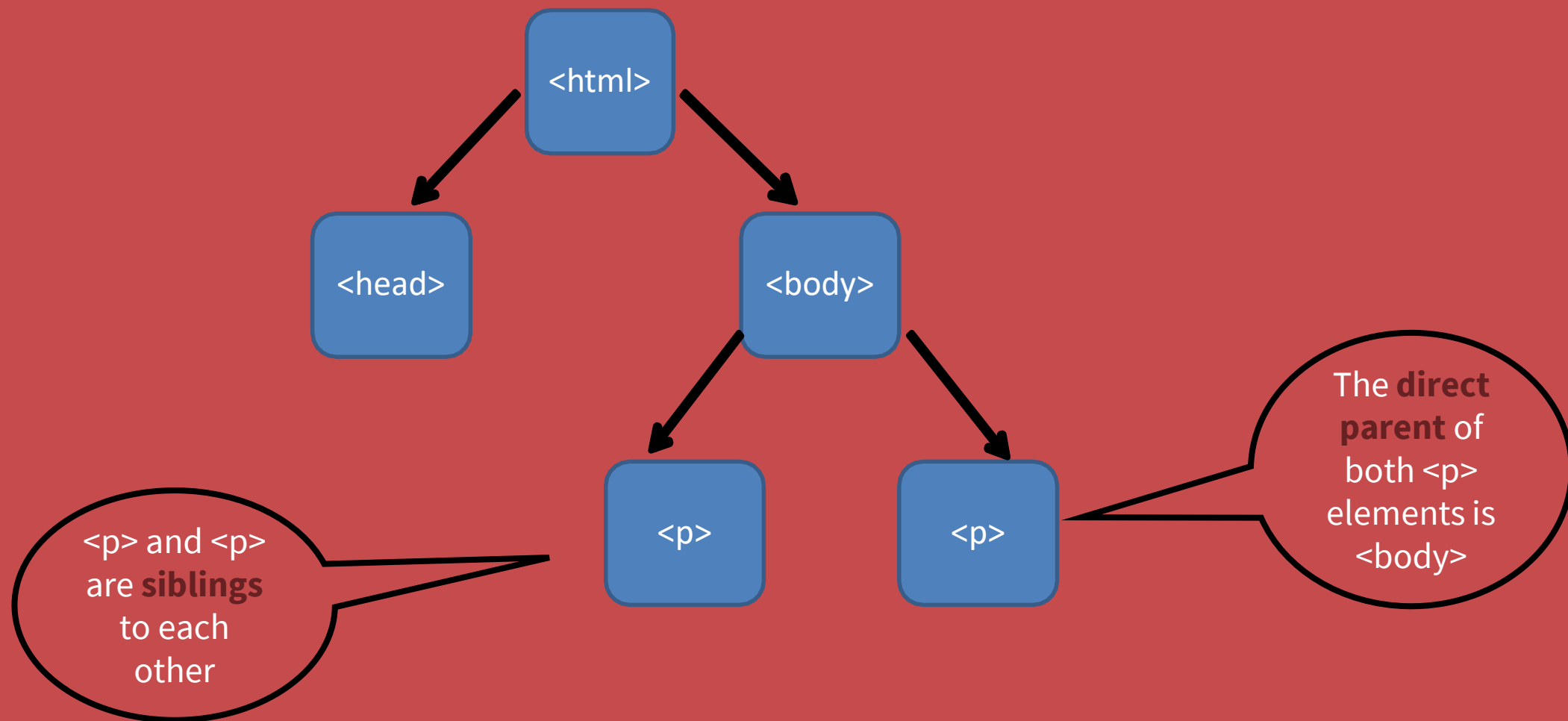
<!DOCTYPE html>



<p> and <p>
are **siblings**
to each
other

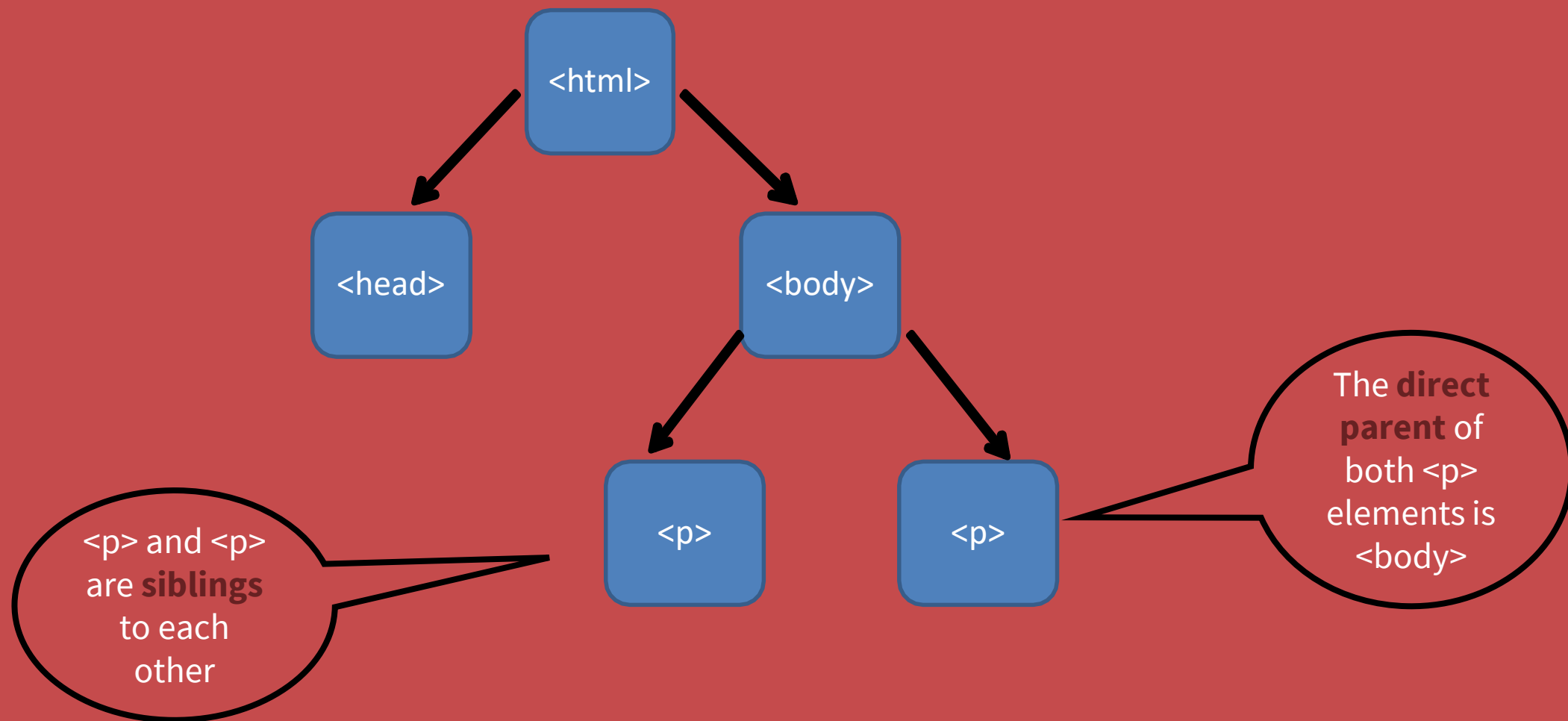
Parent-child-sibling relationships

<!DOCTYPE html>



Parent-child-sibling relationships

<!DOCTYPE html>



Parent-child-sibling relationships

<!DOCTYPE html>

