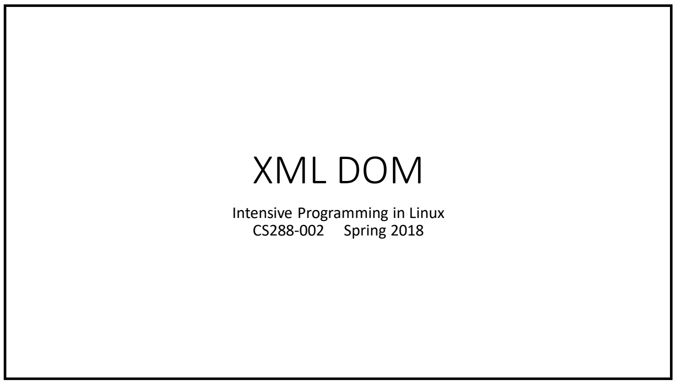
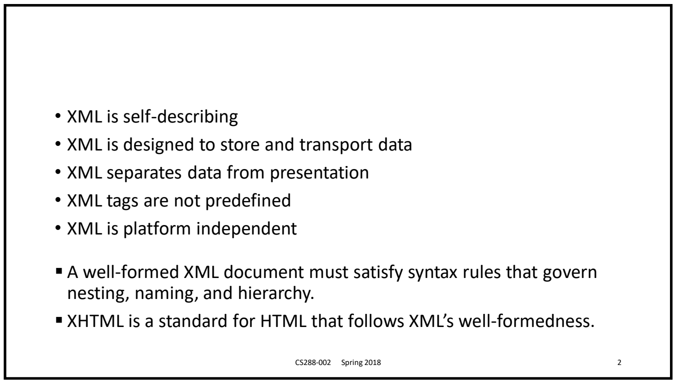


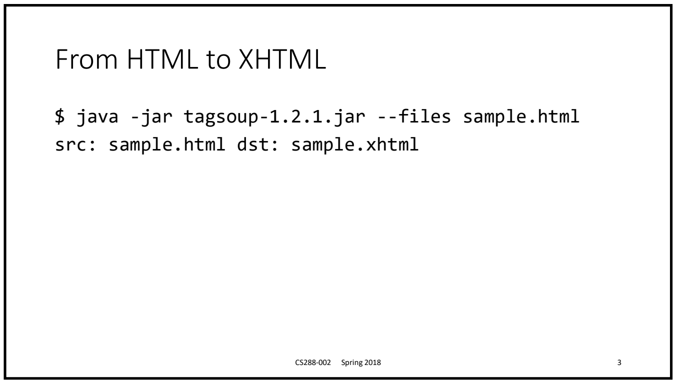
XML DOM

Intensive Programming in Linux
CS288-002 Spring 2018



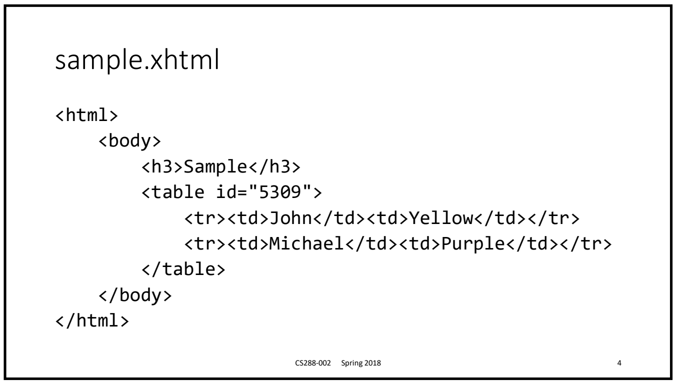
- XML is self-describing
- XML is designed to store and transport data
- XML separates data from presentation
- XML tags are not predefined
- XML is platform independent

- A well-formed XML document must satisfy syntax rules that govern nesting, naming, and hierarchy.
- XHTML is a standard for HTML that follows XML's well-formedness.



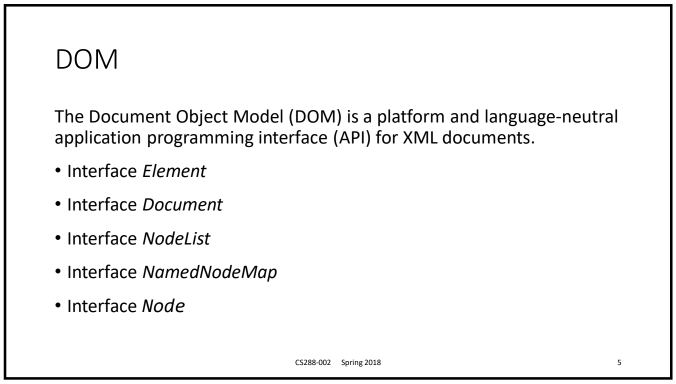
From HTML to XHTML

```
$ java -jar tagsoup-1.2.1.jar --files sample.html
src: sample.html dst: sample.xhtml
```



sample.xhtml

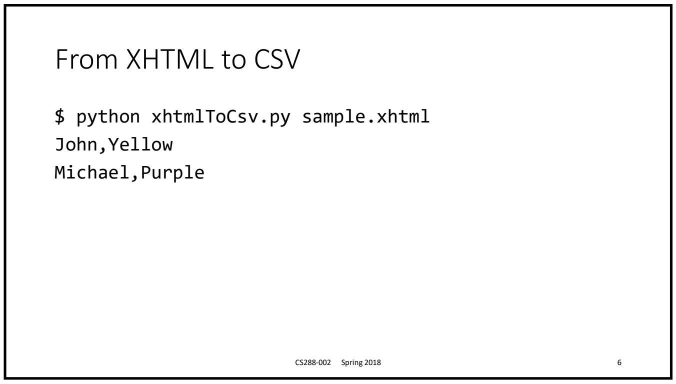
```
<html>
    <body>
        <h3>Sample</h3>
        <table id="5309">
            <tr><td>John</td><td>Yellow</td></tr>
            <tr><td>Michael</td><td>Purple</td></tr>
        </table>
    </body>
</html>
```



DOM

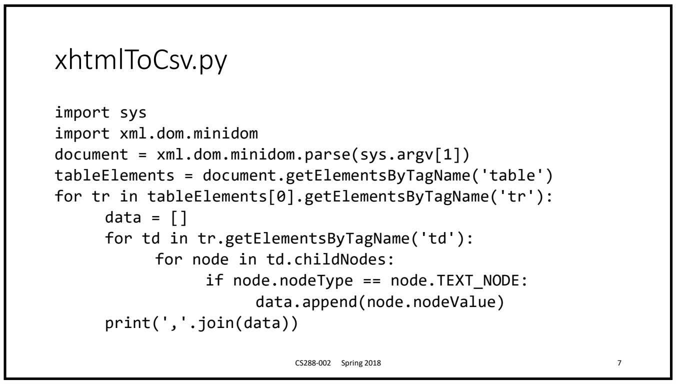
The Document Object Model (DOM) is a platform and language-neutral application programming interface (API) for XML documents.

- Interface *Element*
- Interface *Document*
- Interface *NodeList*
- Interface *NamedNodeMap*
- Interface *Node*



From XHTML to CSV

```
$ python xhtmlToCsv.py sample.xhtml
John,Yellow
Michael,Purple
```



xhtmlToCsv.py

```
import sys
import xml.dom.minidom
document = xml.dom.minidom.parse(sys.argv[1])
tableElements = document.getElementsByTagName('table')
for tr in tableElements[0].getElementsByTagName('tr'):
    data = []
    for td in tr.getElementsByTagName('td'):
        for node in td.childNodes:
            if node.nodeType == node.TEXT_NODE:
                data.append(node.nodeValue)
    print(','.join(data))
```