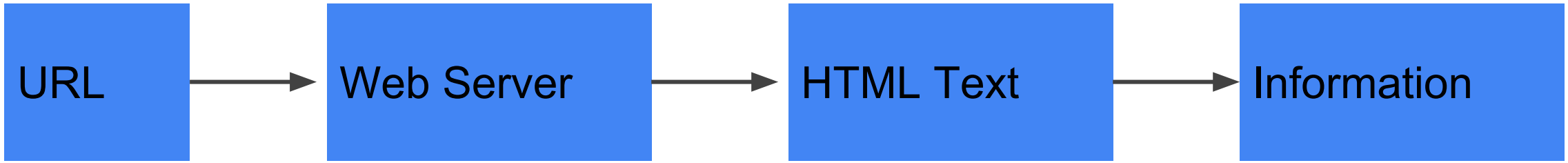
# Web Data Crawling

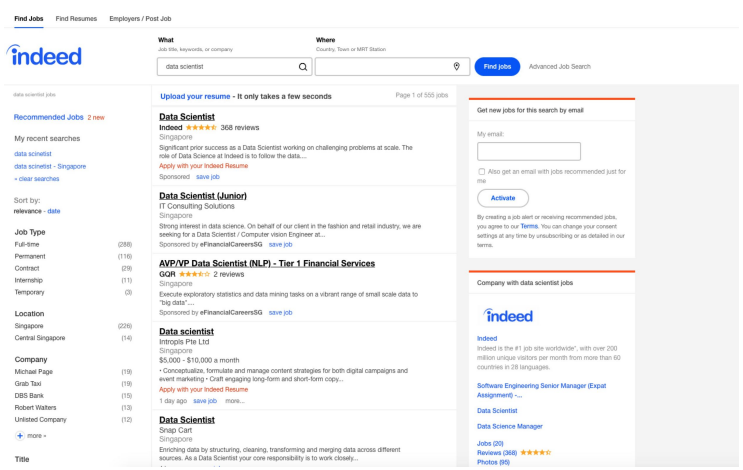

URL
Web Server
HTML Text
Information

# Agenda

- What is HTML
- Hands-On

# What is HTML?

- HTML: **H**yper**T**ext **M**arkup **L**anguage
  - a computer language that is used to create documents on the World Wide Web
  - simple and logical
  - a mark-up language that uses **<Tags>** instead of programming language
- All websites over the internet are plain text files that consist of HTML Tags.




# Tags

- Tags are instructions to markup the text shown on your Web browser.
- All tags are in the format **<Tags>**
- Each tag must be accompanied by a closing tag **</Tags>**
- Elements are made up of two tags (start one and end one) and the element content.

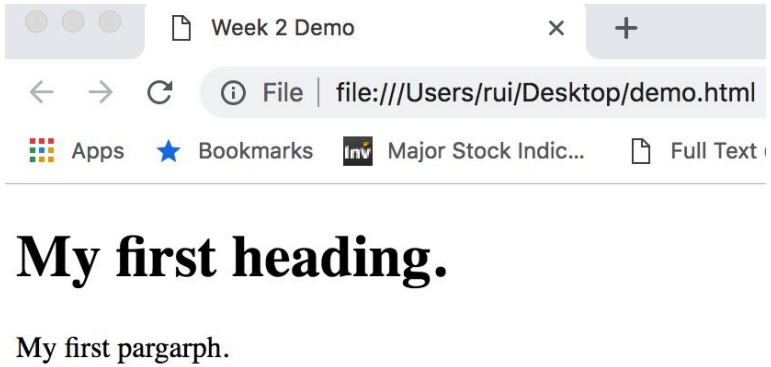
<title>Business Analytics</title>

# Toy Example

```
<!DOCTYPE html>
<html>
<head>
<title>Week 2 Demo</title>
</head>
<body>

<h1>My first heading.</h1>
<p>My first pargarph.</p>

</body>
</html>
```




- Browser use HTML tags to decide how to display the document.
  - **<html>** root element of an HTML page
  - **<head>** contains elements that are about the document which are not displayed in the page itself. **<title>** is one of such element
  - **<body>** is the web page itself
  - <h1> defines a large heading and <p> defines a paragraph

# Beautiful Soup

- Beautiful Soup is a Python library for parsing HTML documents (including having malformed markup), whose name is derived more from the unrelated “tag soup”.
- Help you pull data out of HTML and XML files.