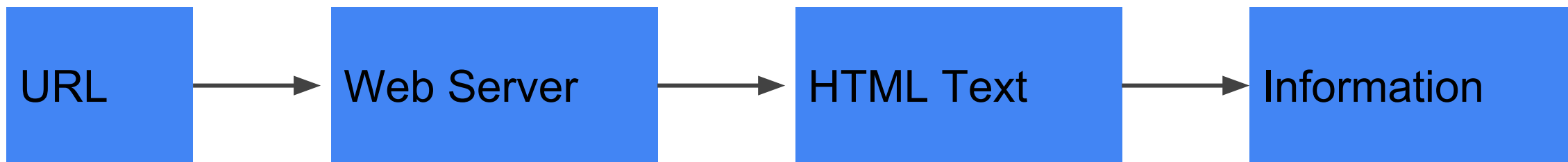


Web Data Crawling

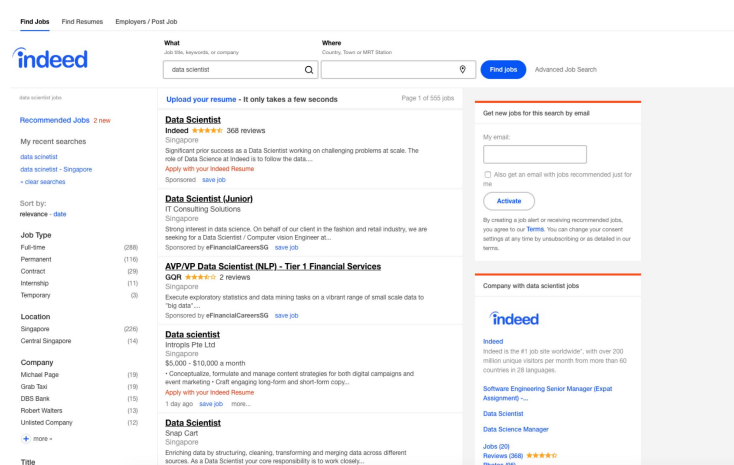


Agenda

- What is HTML
- Hands-On

What is HTML?

- **HTML: HyperText Markup Language**
 - 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.



```
<!DOCTYPE html>
<html lang="en" dir="ltr">
<head>
<meta http-equiv="content-type" content="text/html; charset=UTF-8">
<script type="text/javascript" src="//d3fw5vhllyvee.cloudfront.net/s/af4998f/en_5G.js"></script>
<link href="//d3fw5vhllyvee.cloudfront.net/s/978d98c/jobsearch_all.css" rel="stylesheet" type="text/css">
<link rel="alternate" type="application/rss+xml" title="Data Scientist Jobs, careers" href="http://www.indeed.com.sg/rss?q=data-scientist">
<link rel="alternate" media="only screen and (max-width: 640px)" href="/m/jobs?q=data-scientist">
<script type="text/javascript">

    if (typeof window['closureReadyCallbacks'] == 'undefined') {
        window['closureReadyCallbacks'] = [];
    }

    function call_when_jsall_loaded(cb) {
        if (window['closureReady']) {
            cb();
        } else {
            window['closureReadyCallbacks'].push(cb);
        }
    }
</script>
<meta name="ppstripist" content="1">
<script type="text/javascript" src="//d3fw5vhllyvee.cloudfront.net/s/4c9f4c8/jobsearch-all-compiled_en_5G.js"></script>

var searchUID = '1ctkqc5ku7g6m888';
var tk = '1ctkqc5ku7g6m888';

var loggedIn = false;
var dcnpayload = 'jobse0;jobal0;viewj0;savej0;0232301';
var myindeed = true;
var userEmail = '';
var tellFriendEmail = '';
var globalLoginURL = 'https://www.indeed.com.sg/account/login?dest=k2fjobs%3Fq%3Ddata%2Bscientist%26l%3D';
var globalRegisterURL = 'https://www.indeed.com.sg/account/register?dest=k2fjobs%3Fq%3Ddata%2Bscientist%26l%3D';
var searchKey = 'f5281a4aeef1eal';
var searchState = 'q=data-scientist&pl=';
var searchQS = 'q=data-scientist';
var eventType = 'jobsearch';
var locale = 'en_5G';
function clickId (var a = document.getElementById(id); var hr = a.href; var si = a.href.indexOf('Gjsa='); if (si > 0) return;
function sjondId (var a = document.getElementById(id); var hr = a.href; var ocs = hr.indexOf('&oc=1'); if (ocs < 0) return;
function etatId call) { var a = document.getElementById(id); a href = a href + '&oc=1'; if (ocs < 0) return;
```

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.

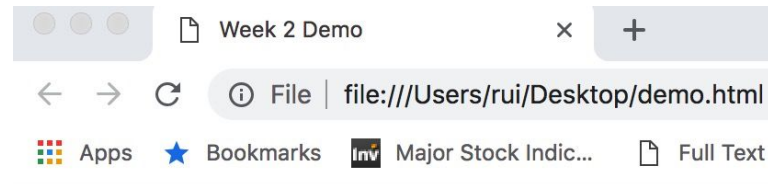
`<title>Business Analytics</title>`

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>
```



My first heading.

My first pargarph.

- 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.