WWW and HTML

Prof. Hyuk-Yoon Kwon

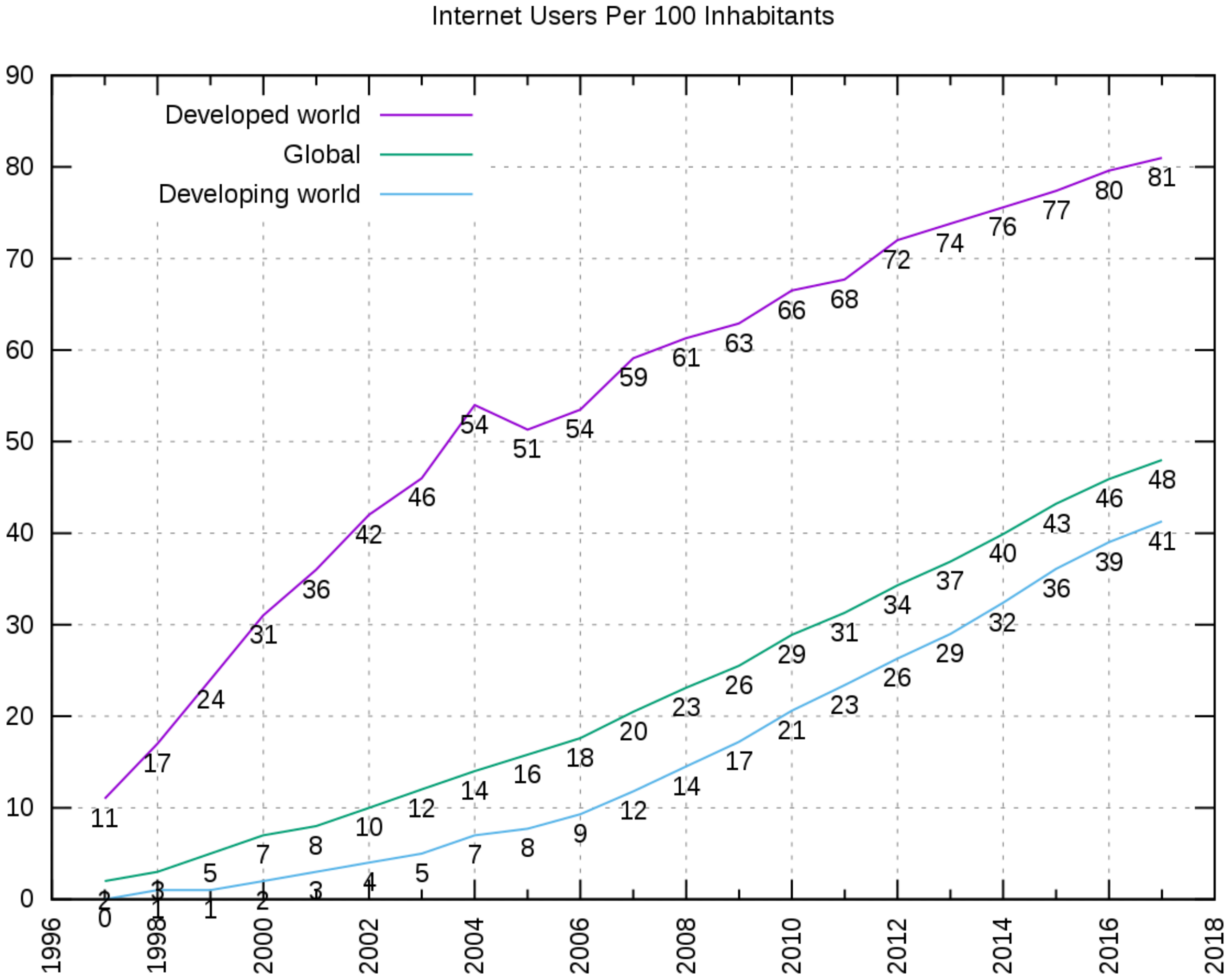
https://sites.google.com/view/seoultech-bigdata

Most parts are based on slides used in (http://ce.sharif.edu/~zarrabi/courses/2013/ce419/notes/)

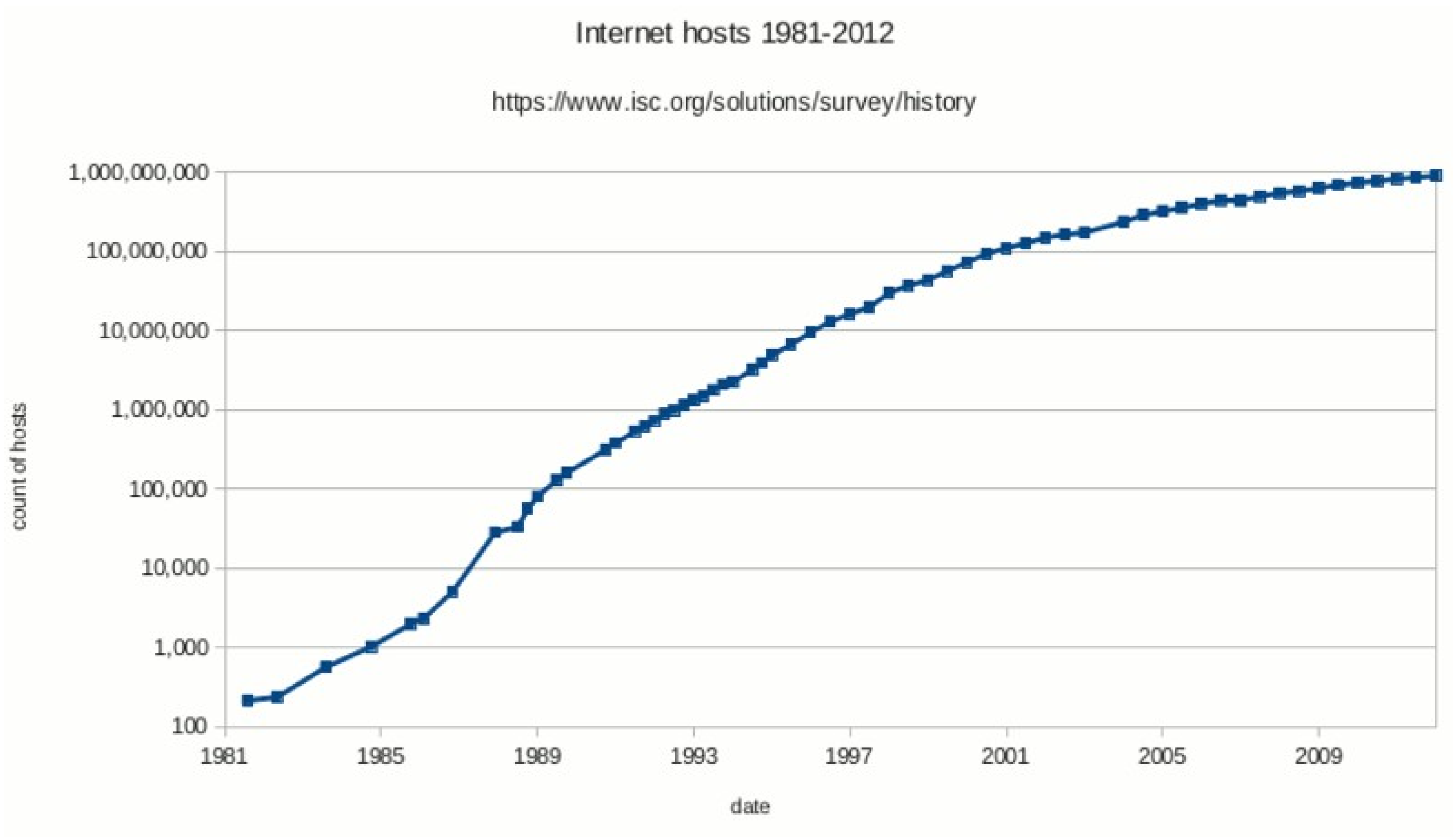
Outline

- **The HTTP Protocol**
  - HTTP Requests
  - HTTP Responses
- **Summary**

Internet Users

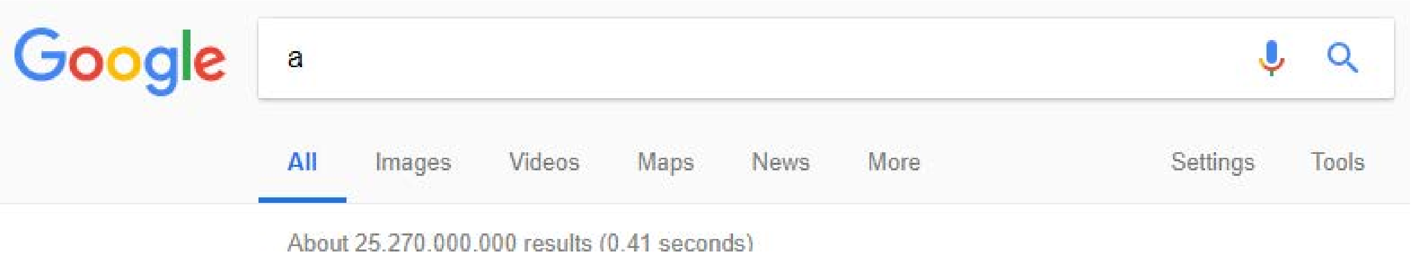


Internet Hosts



Statistics

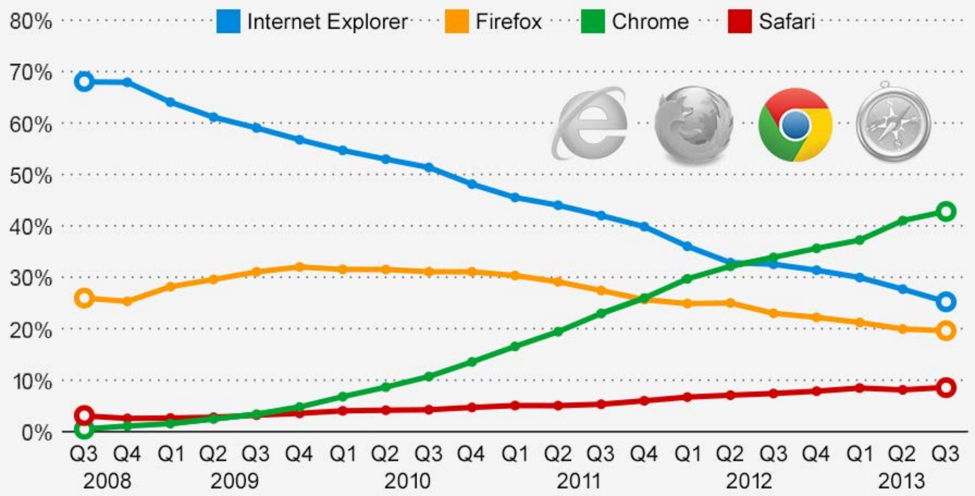
- The number of web users was expected to reach five billion in 2022
- There was a massive number, over 550 billion, of documents on the Web



a

About 25.270.000.000 results (0.41 seconds)

- Most popular web browsers

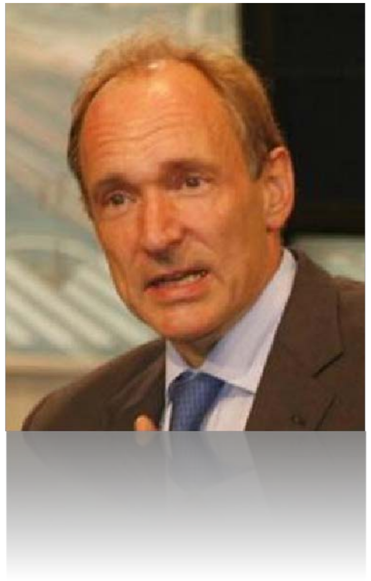


- Most widely-used Internet protocol: HTTP

The Hypertext Transfer Protocol

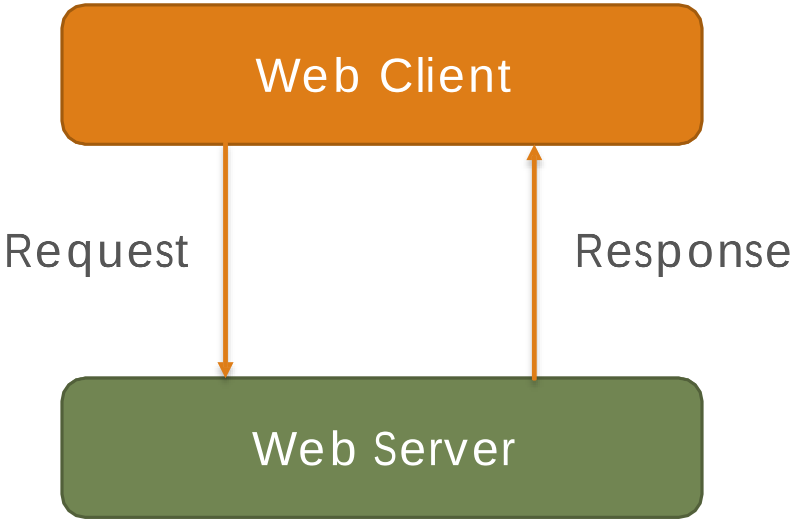
History of HTTP

- Initial version by Tim Berners-Lee (CERN) and implemented in the World's first web browser/server (Dec. 1990)
- HTTP 0.9 defined in 1991
- HTTP 1.0 defined in 1996
- HTTP 1.1 defined in 1997



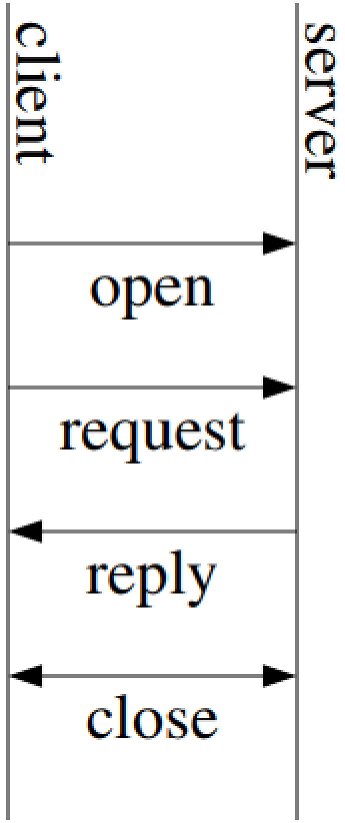
HTTP Protocol Summary

- A typical HTTP transaction:



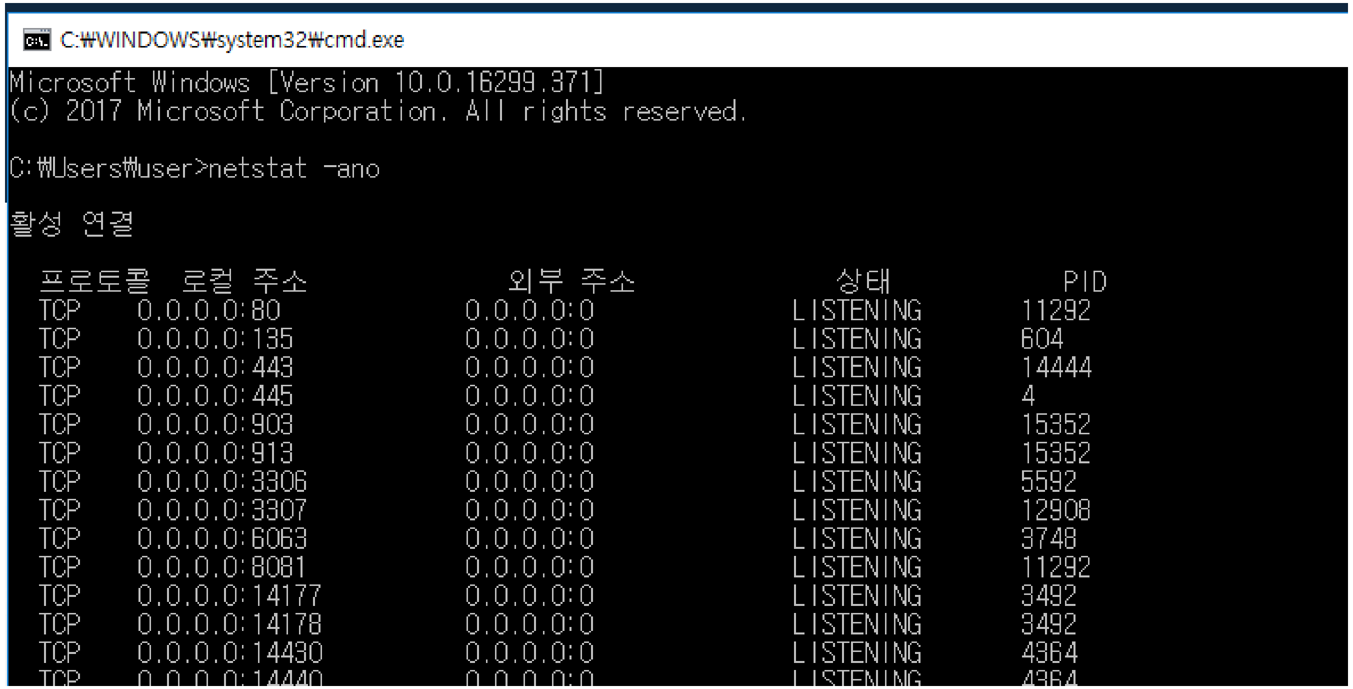
A Typical HTTP transaction

- Client (browser) open connection to server
- Client sends request to server
- Server processes request
- Server replies to client
- Server closes connection



Check Your Connection to Web Server

- Open browser (Internet Explorer, Edge, or Chrome)
- Access a Web site (www.google.com)
- Open command windows (execute cmd.exe)
- Type “netstat –ano”



C:₩WINDOWS₩system32₩cmd.exe

```
Microsoft Windows [Version 10.0.16299.371]
(c) 2017 Microsoft Corporation. All rights reserved.

C:₩Users₩user>netstat -ano

활성 연결

  프로토콜  로컬 주소              외부 주소              상태            PID
  TCP    0.0.0.0:80             0.0.0.0:0              LISTENING       11292
  TCP    0.0.0.0:135            0.0.0.0:0              LISTENING       604
  TCP    0.0.0.0:443            0.0.0.0:0              LISTENING       14444
  TCP    0.0.0.0:445            0.0.0.0:0              LISTENING       4
  TCP    0.0.0.0:903            0.0.0.0:0              LISTENING       15352
  TCP    0.0.0.0:913            0.0.0.0:0              LISTENING       15352
  TCP    0.0.0.0:3306           0.0.0.0:0              LISTENING       5592
  TCP    0.0.0.0:3307           0.0.0.0:0              LISTENING       12908
  TCP    0.0.0.0:6063           0.0.0.0:0              LISTENING       3748
  TCP    0.0.0.0:8081           0.0.0.0:0              LISTENING       11292
  TCP    0.0.0.0:14177          0.0.0.0:0              LISTENING       3492
  TCP    0.0.0.0:14178          0.0.0.0:0              LISTENING       3492
  TCP    0.0.0.0:14430          0.0.0.0:0              LISTENING       4364
  TCP    0.0.0.0:14440          0.0.0.0:0              LISTENING       4364
```

Uniform Resource Locator

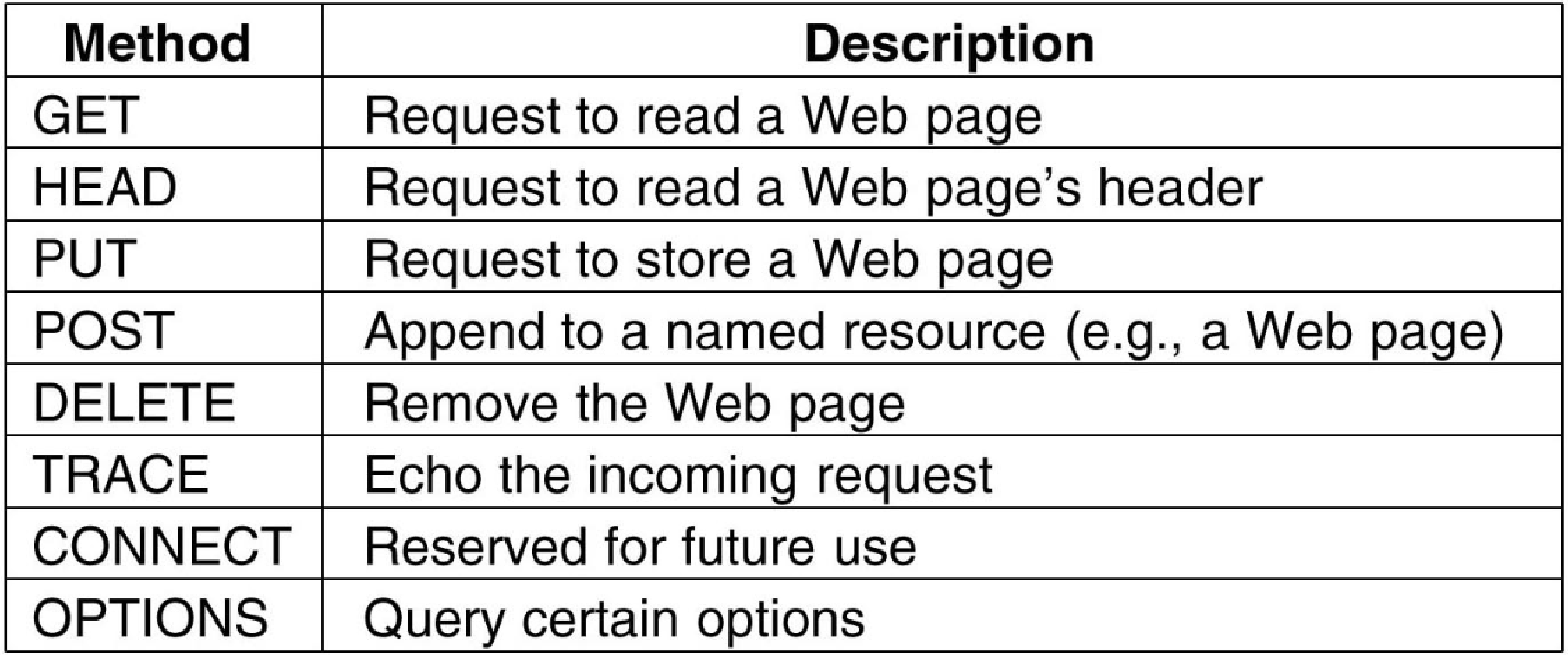
- **URL (Uniform Resource Locator)**
  - A Subset of Uniform Resource Identifiers (URIs)
- **Format:**
  - protocol://host:port/path?query#frag
- **Examples:**
  - http://ce.sharif.edu/courses/92-93/1/
  - https://www.google.com/?q=test

The HTTP Request

- Request includes a header and optional body

```
GET /courses/ HTTP/1.1
Host: ce.sharif.edu
Connection: keep-alive
Cache-Control: max-age=0
Accept: text/html,application/xml;q=0.9,*/*;q=0.8
User-Agent: Chrome/29.0.1547.66 Safari/537.36
Accept-Encoding: gzip,deflate,sdch
Accept-Language: en-GB,en-US;q=0.8,en;q=0.6
Cookie: PHPSESSID=91ef778df57dd25aab1520845ed7076d;
```

Main HTTP Request Types



Method	Description
GET	Request to read a Web page
HEAD	Request to read a Web page’s header
PUT	Request to store a Web page
POST	Append to a named resource (e.g., a Web page)
DELETE	Remove the Web page
TRACE	Echo the incoming request
CONNECT	Reserved for future use
OPTIONS	Query certain options

The HTTP Response

- **Response includes a header and optional body, separated by a blank line**

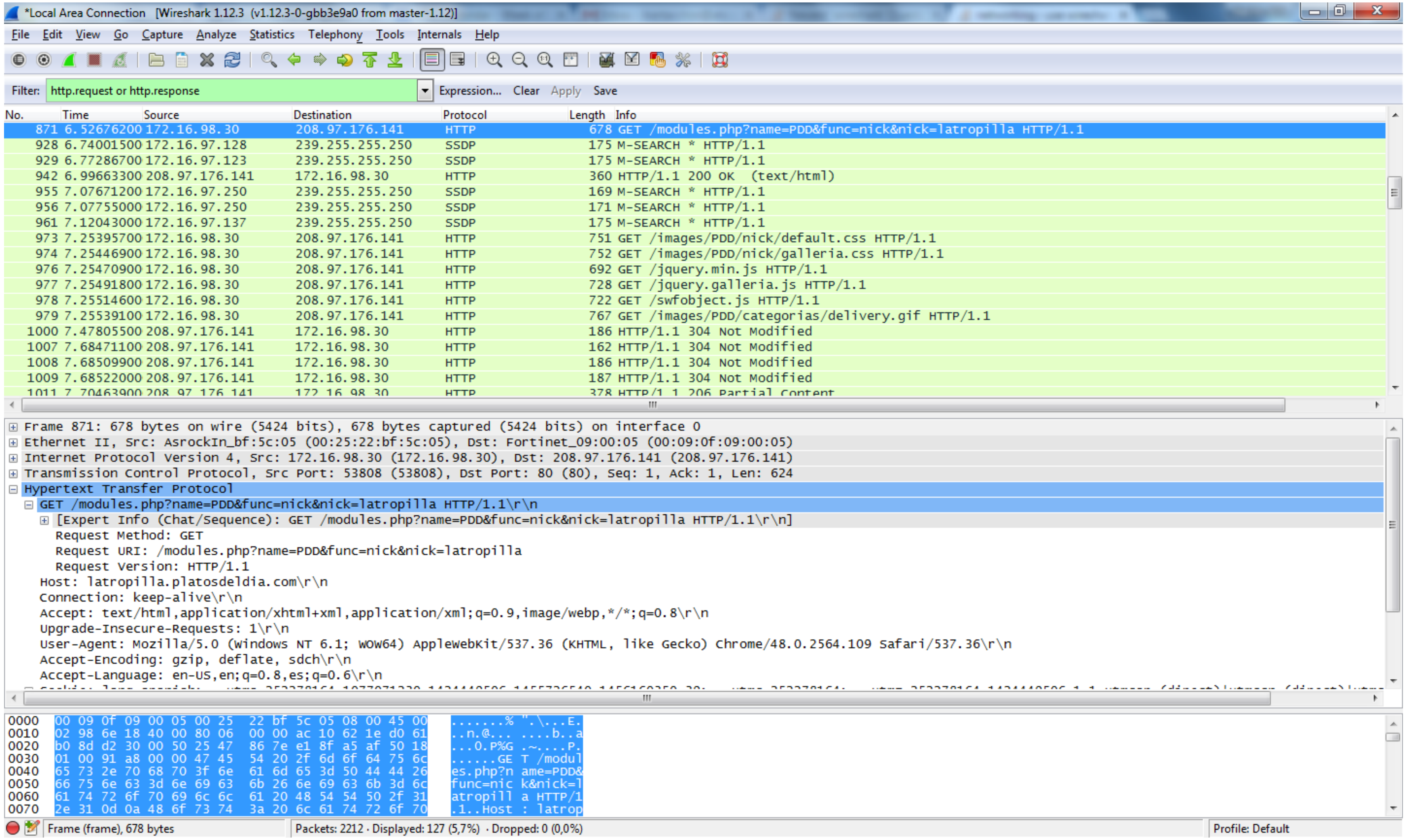
```
HTTP/1.1 200 OK
Date: Sat, 14 Sep 2013 23:09:14 GMT
Server: Apache/2.2.16 (Debian)
Accept-Ranges: bytes
Vary: Accept-Encoding
Content-Encoding: gzip
Content-Length: 1719
Keep-Alive: timeout=15, max=99
Connection: Keep-Alive
Content-Type: text/html

<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
...
```

Common Response Codes

- **200 OK**
  - Success
- **404 Not Found**
  - The specified resource does not exist
- **403 Forbidden**
  - The specified resource exists, but can not be accessed
- **301 & 302 Document Moved**
  - The resource is at the new (specified) location

Monitoring HTTP Websever

Wireshark



*Local Area Connection [Wireshark 1.12.3 (v1.12.3-0-gbb3e9a0 from master-1.12)]
File Edit View Go Capture Analyze Statistics Telephony Tools Internals Help
Expression... Clear Apply Save
Profile: Default

Filter: http.request or http.response

```
No.   Time        Source          Destination      Protocol  Length Info
 871 6.52676200 172.16.98.30    208.97.176.141   HTTP       678 GET /modules.php?name=PDD&func=nick&nick=latropilla HTTP/1.1
 928 6.74001500 172.16.97.128   239.255.255.250  SSDP       175 M-SEARCH * HTTP/1.1
 929 6.77286700 172.16.97.123   239.255.255.250  SSDP       175 M-SEARCH * HTTP/1.1
 942 6.99663300 208.97.176.141  172.16.98.30     HTTP       360 HTTP/1.1 200 OK  (text/html)
 955 7.07671200 172.16.97.250   239.255.255.250  SSDP       169 M-SEARCH * HTTP/1.1
 956 7.07755000 172.16.97.250   239.255.255.250  SSDP       171 M-SEARCH * HTTP/1.1
 961 7.12043000 172.16.97.137   239.255.255.250  SSDP       175 M-SEARCH * HTTP/1.1
 973 7.25395700 172.16.98.30    208.97.176.141   HTTP       751 GET /images/PDD/nick/default.css HTTP/1.1
 974 7.25446900 172.16.98.30    208.97.176.141   HTTP       752 GET /images/PDD/nick/galleria.css HTTP/1.1
 976 7.25470900 172.16.98.30    208.97.176.141   HTTP       692 GET /jquery.min.js HTTP/1.1
 977 7.25491800 172.16.98.30    208.97.176.141   HTTP       728 GET /jquery.galleria.js HTTP/1.1
 978 7.25514600 172.16.98.30    208.97.176.141   HTTP       722 GET /swfobject.js HTTP/1.1
 979 7.25539100 172.16.98.30    208.97.176.141   HTTP       767 GET /images/PDD/categorias/delivery.gif HTTP/1.1
1000 7.47805500 208.97.176.141  172.16.98.30     HTTP       186 HTTP/1.1 304 Not Modified
1007 7.68471100 208.97.176.141  172.16.98.30     HTTP       162 HTTP/1.1 304 Not Modified
1008 7.68509900 208.97.176.141  172.16.98.30     HTTP       186 HTTP/1.1 304 Not Modified
1009 7.68522000 208.97.176.141  172.16.98.30     HTTP       187 HTTP/1.1 304 Not Modified
1011 7.70463900 208.97.176.141  172.16.98.30     HTTP       378 HTTP/1.1 206 Partial Content
```

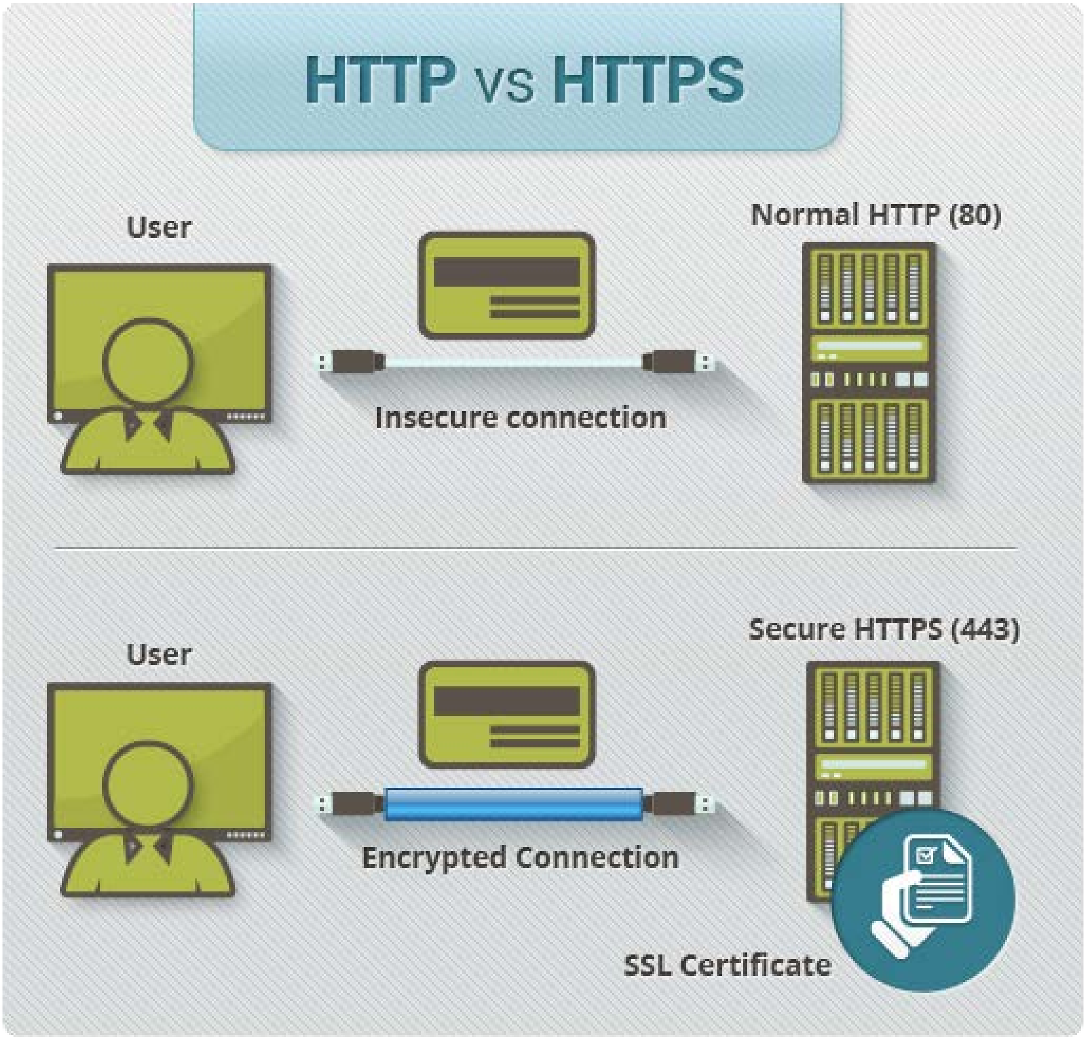
```
⊞ Frame 871: 678 bytes on wire (5424 bits), 678 bytes captured (5424 bits) on interface 0
⊞ Ethernet II, Src: AsrockIn_bf:5c:05 (00:25:22:bf:5c:05), Dst: Fortinet_09:00:05 (00:09:0f:09:00:05)
⊞ Internet Protocol Version 4, Src: 172.16.98.30 (172.16.98.30), Dst: 208.97.176.141 (208.97.176.141)
⊞ Transmission Control Protocol, Src Port: 53808 (53808), Dst Port: 80 (80), Seq: 1, Ack: 1, Len: 624
⊟ Hypertext Transfer Protocol
  ⊟ GET /modules.php?name=PDD&func=nick&nick=latropilla HTTP/1.1\r\n
    ⊞ [Expert Info (Chat/Sequence): GET /modules.php?name=PDD&func=nick&nick=latropilla HTTP/1.1\r\n]
      Request Method: GET
      Request URI: /modules.php?name=PDD&func=nick&nick=latropilla
      Request Version: HTTP/1.1
    Host: latropilla.platosdeldia.com\r\n
    Connection: keep-alive\r\n
    Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8\r\n
    Upgrade-Insecure-Requests: 1\r\n
    User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/48.0.2564.109 Safari/537.36\r\n
    Accept-Encoding: gzip, deflate, sdch\r\n
    Accept-Language: en-US,en;q=0.8,es;q=0.6\r\n
```

```
0000  00 09 0f 09 00 05 00 25  22 bf 5c 05 08 00 45 00   .......% ".\...E.
0010  02 98 6e 18 40 00 80 06  00 00 ac 10 62 1e d0 61   ..n.@... ....b..a
0020  b0 8d d2 30 00 50 25 47  86 7e e1 8f a5 af 50 18   ...0.P%G .~....P.
0030  01 00 91 a8 00 00 47 45  54 20 2f 6d 6f 64 75 6c   ......GE T /modul
0040  65 73 2e 70 68 70 3f 6e  61 6d 65 3d 50 44 44 26   es.php?n ame=PDD&
0050  66 75 6e 63 3d 6e 69 63  6b 26 6e 69 63 6b 3d 6c   func=nic k&nick=l
0060  61 74 72 6f 70 69 6c 6c  61 20 48 54 54 50 2f 31   atropill a HTTP/1
0070  2e 31 0d 0a 48 6f 73 74  3a 20 6c 61 74 72 6f 70   .1..Host : latrop
```

Frame (frame), 678 bytes

Packets: 2212 · Displayed: 127 (5,7%) · Dropped: 0 (0,0%)

HTTP vs HTTPS



Practice: Network Monitoring

Install “Microsoft Network Monitor”

- Type “Microsoft network monitor” in Google
- Download and install “Microsoft network monitor 3.4”
- Reboot your computers

Monitor network access

- Execute “Microsoft Network Monitor”
- Click “new capture” – “Start”
- Access http://www.seoultech.ac.kr
- Find the access log in “Microsoft Network Monitor”

Check HTTPS connection

- Try https://www.seoultech.ac.kr
- Check the difference from the previous result

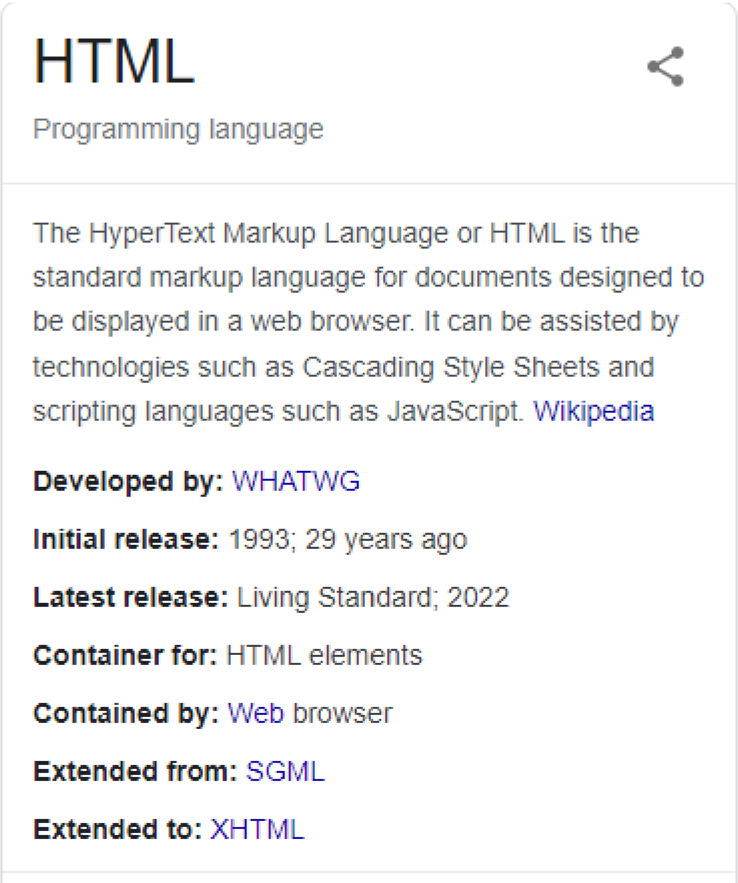
The Hypertext Markup Language (HTML)

Outline

- History of HTML
- HTML Document Structure
  - Document Header
  - Document Body
- HTML Tags

HTML History

HTML	1991
HTML+	1993
HTML 2.0	1995
HTML 3.2	1997
HTML 4.01	1999
XHTML 1.0	2000
HTML5	2012



HTML

Programming language

The HyperText Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets and scripting languages such as JavaScript. Wikipedia

Developed by: WHATWG

Initial release: 1993; 29 years ago

Latest release: Living Standard; 2022

Container for: HTML elements

Contained by: Web browser

Extended from: SGML

Extended to: XHTML

HTML Example

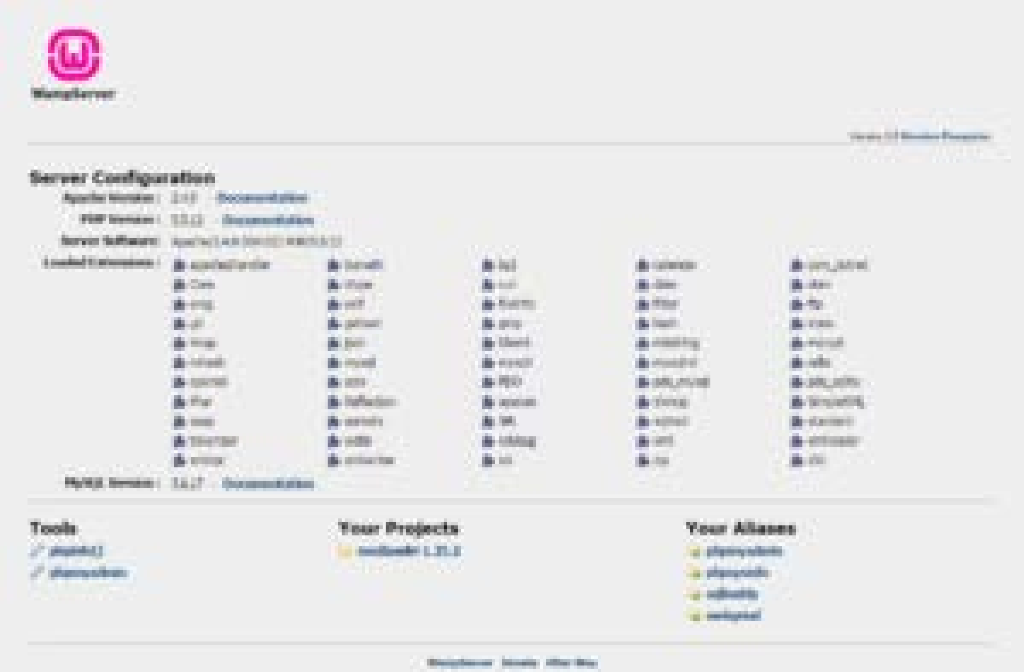
- Minimal HTML5 document

```
<!DOCTYPE html>
<html>
<head>
    <title>My Page Title</title>
</head>

<body>
    <p>Hello World!</p>
</body>
</html>
```

Practice: Start Your WAMP

Install WAMP

- A software stack for the Microsoft Windows operating system, consisting of the Apache web server, MySQL database, and PHP programming language
- https://bitnami.com/stack/wamp/installer



WampServer

Server Configuration

Apache Version : 2.4.9 Documentation

PHP Version : 5.5.12 Documentation

Server Software : [illegible]

Loaded Extensions : [illegible]

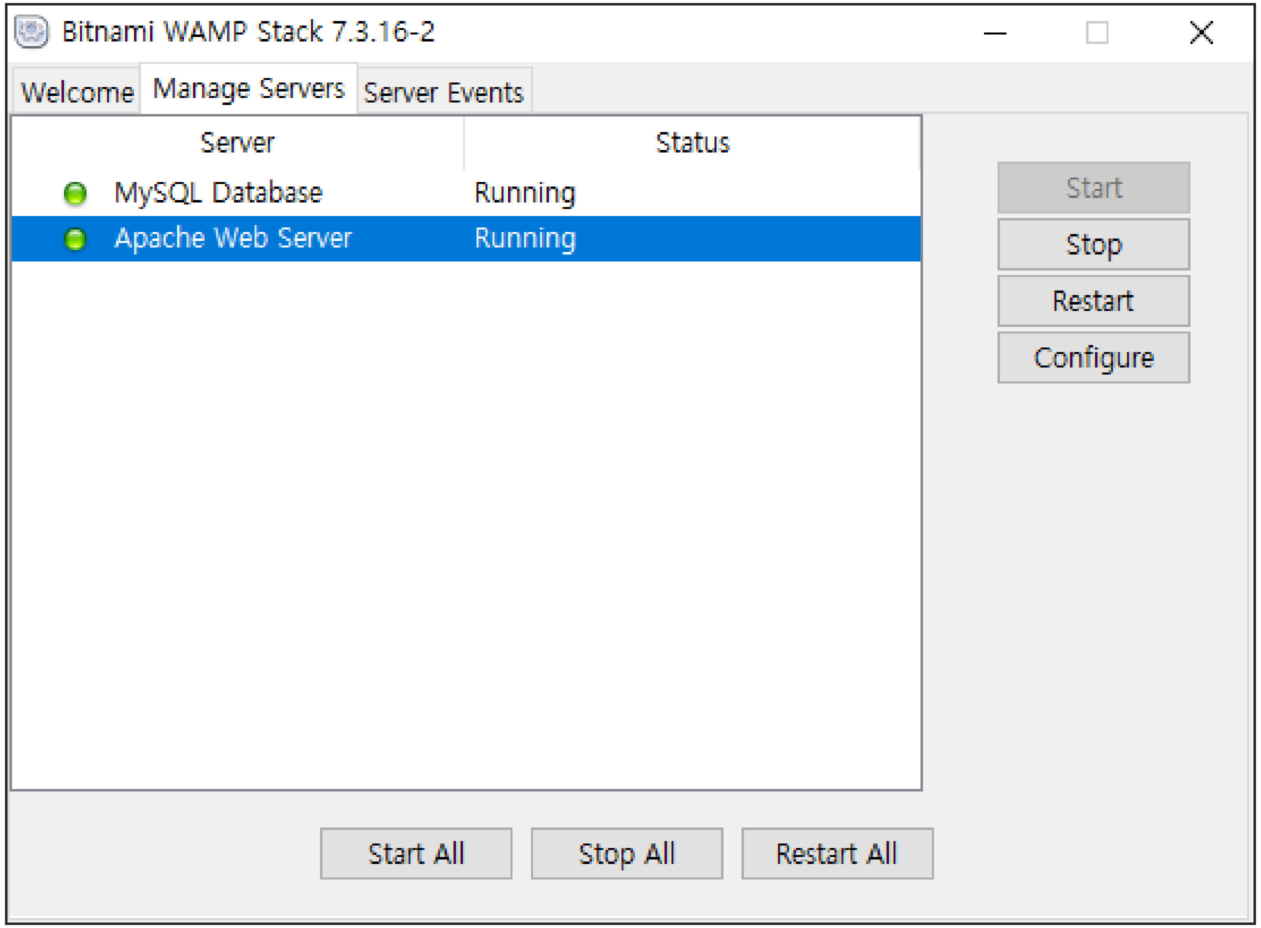
MySQL Version : 5.6.17 Documentation

Tools
- phpinfo()
- phpmyadmin

Your Projects
- mediawiki 1.25.2

Your Aliases
- phpmyadmin
- phpsysinfo
- sqlbuddy
- webgrind

■ Start WAMP



Server	Status
● MySQL Database	Running
● Apache Web Server	Running

- **Try to access your Web page**
  - localhost/index.html

  OR

  - 127.0.0.1/index.html

- **Make a new Web page having the following contents**

```
<!DOCTYPE html>
<html>
<head>
    <title>My Page Title</title>
</head>

<body>
    <p>Hello World!</p>
</body>
</html>
```

HTML Basics

- **HTML Tags**
  - HTML markup tags (keywords), like `<p>` and `</p>`
- **HTML Element**
  - Anything between a start and end tag, including the tags:
    - `<p>Hello World!</p>`
  - Void elements (single tags): `<br/>`
- **HTML Attributes**
  - Provide additional information about an element
    - `<a href="test.html">Link</a>`
  - Come in name/value pairs: name="value“

DOCTYPE

- **HTML5**
  - <!DOCTYPE html>
- **HTML 4.01**
  - <!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
- **XHTML 1.0**
  - <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1- transitional.dtd">

Document Head

- The HEAD section contains information about the document as opposed to the contents of the document
- Minimally, it should include a title
    - <title>My Page Title</title>
- The HEAD may also include metadata such as character encoding and information about how to format the document

Document Body

- The BODY contains the document itself, i.e., what the user sees
- Any text (also called character data or cdata) within the body must be contained within some other tag

Document Structure

- <h1> . . . <h6> (headings)
- <p> (paragraph)
- <br /> (line break)
- <hr /> (horizontal rule)
- <!-- This is a comment -->

Practice: Document Structure

1. Mark up the following text with appropriate tags:
   - "Universal Studios Presents" is the most important content.
   - "Jurassic Park" is the next most important content.
   - "About" is of lesser importance than Jurassic Park.

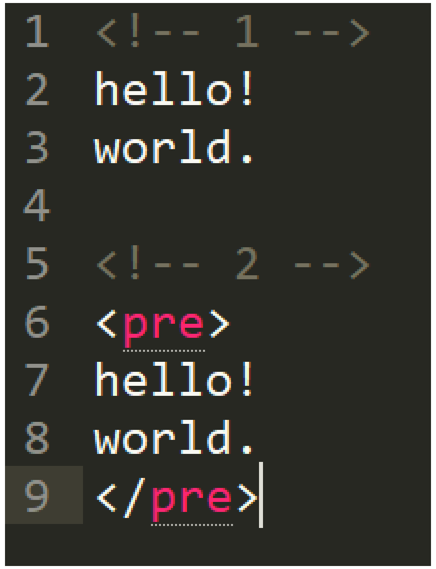
2. Display the poem over 4 lines using a line break for each sentence .
   - My Bonnie lies over the ocean.
   - My Bonnie lies over the sea.
   - My Bonnie lies over the ocean.
   - Oh, bring back my Bonnie to me.

3. Make each sentence as a paragraph. Add comment tags around the "Do not display this text in the browser"

- This is a paragraph
- Do not display this text in the browser
- This is another paragraph

Text Formatting

- <b>Bold text</b>
- <i>Italic text</i>
- <strong>Important text</strong>
- <em>Emphasized text</em>
- <small>Smaller text</small>
- <code>Computer code</code>
- <pre>Preformatted text</pre>



```
<!-- 1 -->
hello!
world.

<!-- 2 -->
<pre>
hello!
world.
</pre>
```

Text Formatting (cont'd)

- <del> (deleted text)
- <ins> (inserted text)
- <sub> (subscripted text)
- <sup> (superscripted text)
- <q> (short quotation)
- <blockquote> (long quotation)
- <abbr title=“”> (abbreviation)
  - Element: full name
  - Attribute: abbreviated name
- <address> (contact information)
- <center> (center arrangement)

Practice: Text Formatting

1. Extra importance to the word "degradation" in the paragraph.
    - WWF's mission is to stop the degradation of our planet's natural environment.

2. Emphasize the word "metropolitan" in the text below.
    - Tokyo is the capital of Japan, the center of the Greater Tokyo Area, and the most populous metropolitan area in the world.

3. Apply subscript formatting to the number "2" in the text below.
    - H2O is the scientific term for water.

4. Add a line through (strikeout) the letters "blue" in the text below.
    - My favorite color is blue red.

5. Print the squares of the numbers 1 - 5. Each number should be on a separate line, next to it the number 2 superscripted, an equal sign and the result. (Example: 5^2 = 25)

6. Print some preformatted text of your choosing. (hint: use the <pre> tag)

7. Print some deleted and inserted text of your choosing.

8. Print two addresses in the same format used on the front of envelopes (senders address in top left corner, receivers address in the center).

9. Print three abbreviations of your choosing, each separated by two lines.

Links

- <a href=“URL"> (page link)
  - Absolute or relative address
- <a href="mailto: ..."> (email link)
- <a name="name"> (anchor)
- <a href="#name"> (link to anchor)

Images

- Image: <img src="PATH" width="42" height="42" />
- Image Link: <a href="..."><img src="URL" /></a>

Lists

- Ordered Lists (numbered): <ol>
- Unordered Lists (bulleted): <ul>

```
<ul>
    <li>Cat</li>
    <li>Dog</li>
    <li>Mouse</li>
</ul>
```

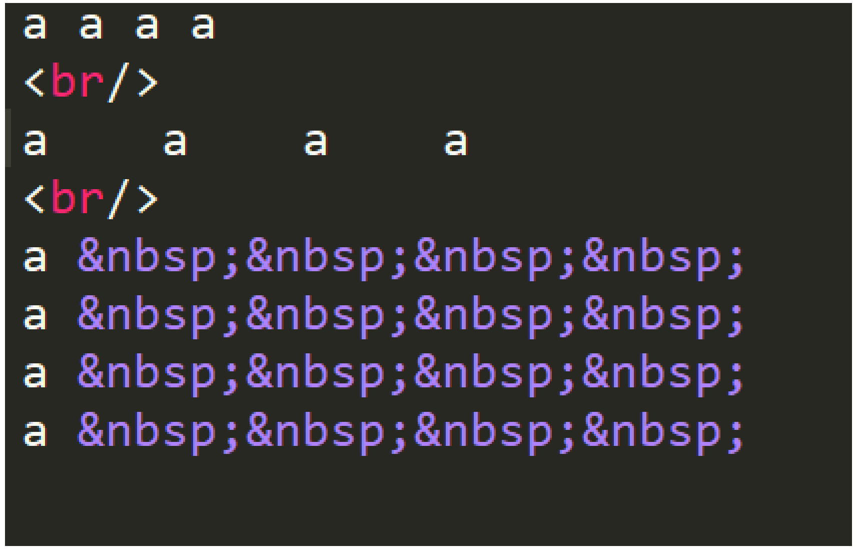
Definition Lists

- Definition Lists (dictionary like): <dl>
- Each item has a *term* <dt> and a *definition* <dd>

```
<dl>
  <dt>Item 1</dt>
  <dd>Describe item 1</dd>

  <dt>Item 2</dt>
  <dd>Describe item 2</dd>
</dl>
```

HTML Entities

- **HTML entities are used to display special characters in HTML**
- **Examples:**
  - &nbsp; (non-breaking space)
  - &lt; (less than: <)



```
a a a a
<br/>
a    a    a    a
<br/>
a &nbsp;&nbsp;&nbsp;&nbsp;
a &nbsp;&nbsp;&nbsp;&nbsp;
a &nbsp;&nbsp;&nbsp;&nbsp;
a &nbsp;&nbsp;&nbsp;&nbsp;
```

Useful Entities

- **Entities general format:**
  - &name;
  - &#number;

Character	Description	Entity Name	Entity Number
	non-breaking space	&nbsp;	&#160;
<	less than	&lt;	&#60;
>	greater than	&gt;	&#62;
&	ampersand	&amp;	&#38;
€	euro	&euro;	&#8364;
§	section	&sect;	&#167;
©	copyright	&copy;	&#169;

Practice

1. Create some links to various search engines (google, yahoo, Bing, lycos, etc).
2. Print two lists with any information you want. One list should be an ordered list, the other list should be an unordered list
3. Print two paragraphs that are both indented using the &nbsp;
4. Create a page with a link at the top of it that when clicked will jump all the way to the bottom of the page. At the bottom of the page there should be a link to jump back to the top of the page (Hint: Use anchor)

5. Display five different images. Skip two lines between each image. Store all the images in one directory.

6. Display an image that when clicked will link to a search engine of your choice

Summary

- **HTML is universal markup language of the web**
- **There are various types of HTML**
  - Use DOCTYPE to specify the document type
- **Basic HTML elements include:**
  - Paragraphs and Headings
  - Text Formatting
  - Links, Images, and Lists

The Hypertext Markup Language (HTML) – Part II

Outline

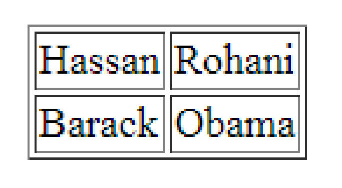
- **HTML Structures**
  - Tables
  - Forms
- **New HTML5 Elements**
- **Summary**

HTML Tables

Tables

- Tables are created with <table> tag
- Each table is divided into *rows* <tr>, containing *table data* <td>

```
 <table border="1">
<tr>
     <td>Hassan</td>
     <td>Rohani</td>
</tr>
<tr>
     <td>Barack</td>
     <td>Obama</td>
</tr>
</table>
```



Hassan	Rohani
Barack	Obama

Table Size and Alignment

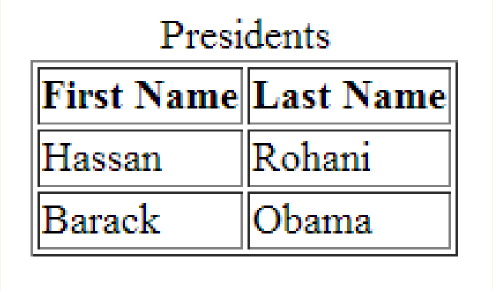
- Size of each column can be determined by width
- Alignment of each column can be determined by align

```
<table border="1">
<tr>
<td width=100 align=center>Hassan</td>
<td width=100 align=right>Rohani</td>
</tr>
<tr>
<td width=100 align=left>Barack</td>
<td width=100 align=center>Obama</td>
</tr>
</table>
```

Headers and Captions

- Tables can have *headers* and *captions*

```
<table border="1">
  <caption>Presidents</caption>
  <tr>
    <th>First Name</th>
    <th>Last Name</th>
  </tr>
  <tr>
    <td>Hassan</td>
    <td>Rohani</td>
  </tr>
  <tr>
    <td>Barack</td>
    <td>Obama</td>
  </tr>
</table>
```

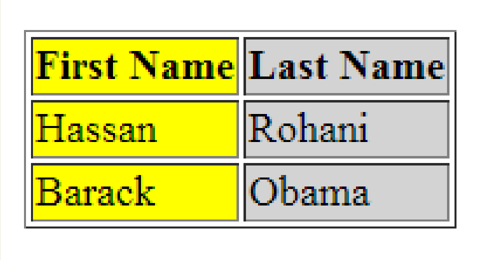


Presidents

First Name	Last Name
Hassan	Rohani
Barack	Obama

Column Groups

```
<table border="1">
  <colgroup>
    <col style="background-color: yellow" />
    <col style="background-color: lightgray" />
  </colgroup>
  <tr>
    <th>First Name</th>
    <th>Last Name</th>
  </tr>
  <tr>
    <td>Hassan</td>
    <td>Rohani</td>
  </tr>
  <tr>
    <td>Barack</td>
    <td>Obama</td>
  </tr>
</table>
```



First Name	Last Name
Hassan	Rohani
Barack	Obama

Span Columns or Rows

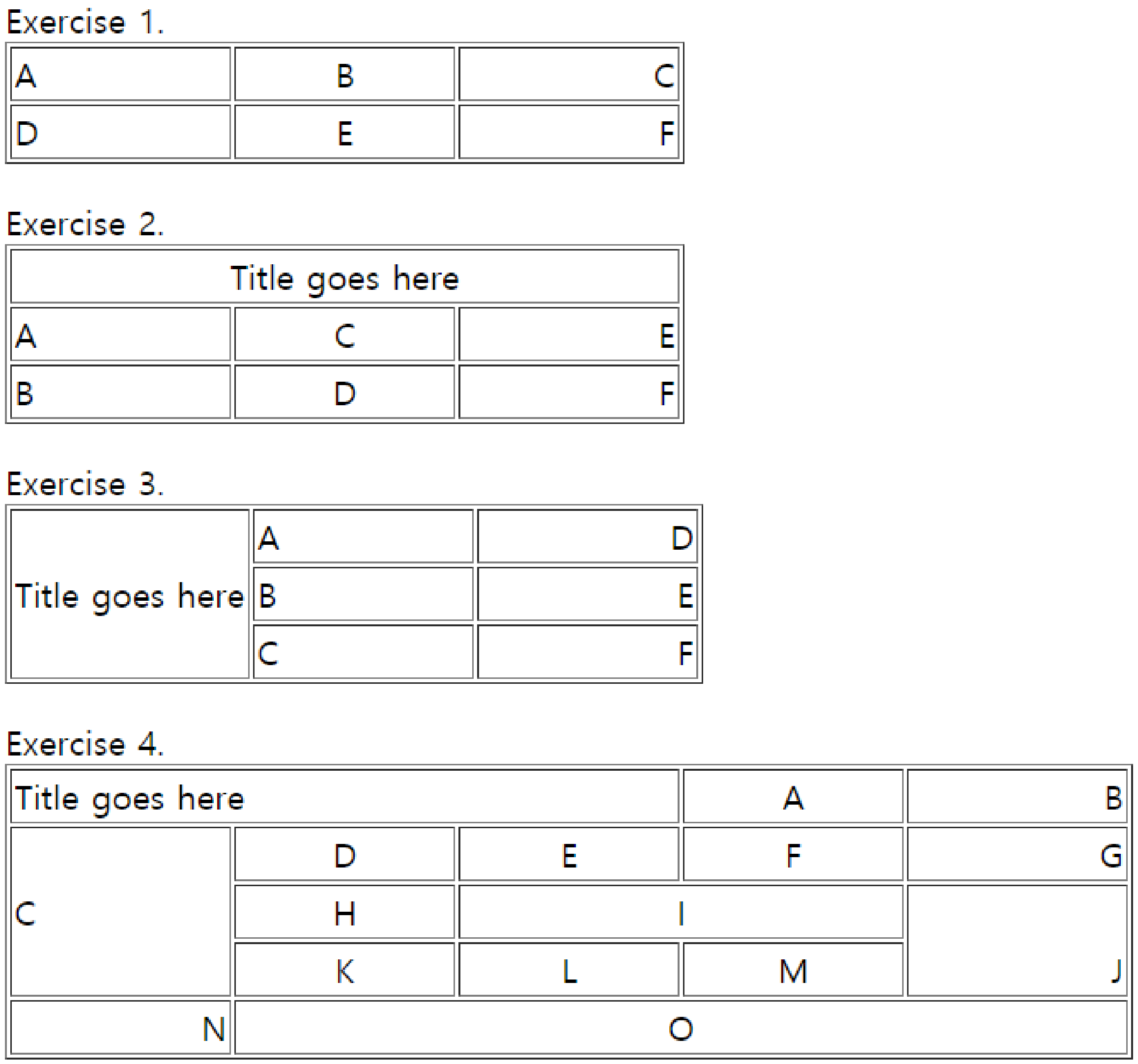
- colspan attribute can be used to make a cell span more than one column
- rowspan attribute can be used to make a cell span more than one row

```
<table border="1">
  <tr>
    <th>Name</th>
    <th colspan="2">Telephone</th>
  </tr>
  <tr>
    <td>Bill Gates</td>
    <td>55577854</td>
    <td>55577855</td>
  </tr>
</table>
```

```
<table border="1">
  <tr>
    <th>Name:</th>
    <td>Bill Gates</td>
  </tr>
  <tr>
    <th rowspan="2">Telephone:</th>
    <td>55577854</td>
  </tr>
  <tr>
    <td>55577855</td>
  </tr>
</table>
```

Practice

- **Make the following four tables**



Exercise 1.

A	B	C
D	E	F

Exercise 2.

	Title goes here	
A	C	E
B	D	F

Exercise 3.

Title goes here	A	D
	B	E
	C	F

Exercise 4.

Title goes here			A	B
C	D	E	F	G
	H	I		J
	K	L	M	
N	O			