Networked Programs

Chapter 12

Python for Informatics: Exploring Information
www.py4inf.com

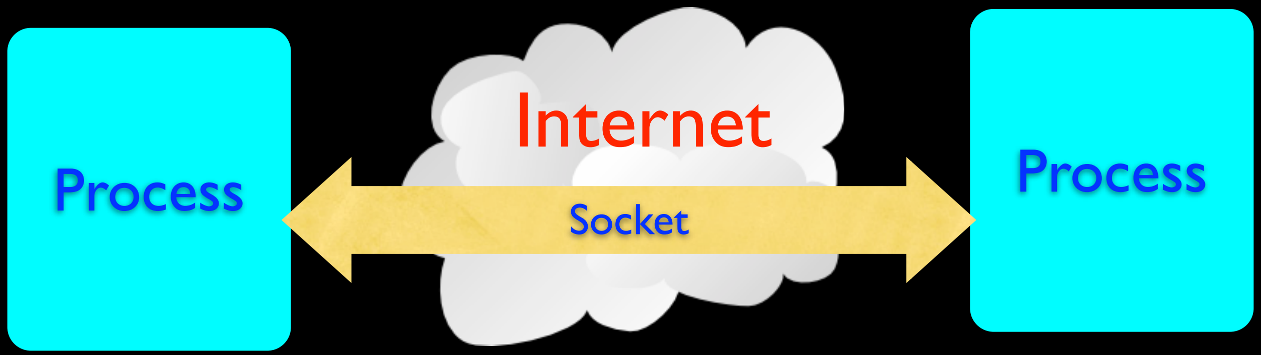
open.michigan

UNIVERSITY OF MICHIGAN

school of information

TCP Connections / Sockets

"In computer networking, an Internet socket or network socket is an endpoint of a bidirectional inter-process communication flow across an Internet Protocol-based computer network, such as the Internet."

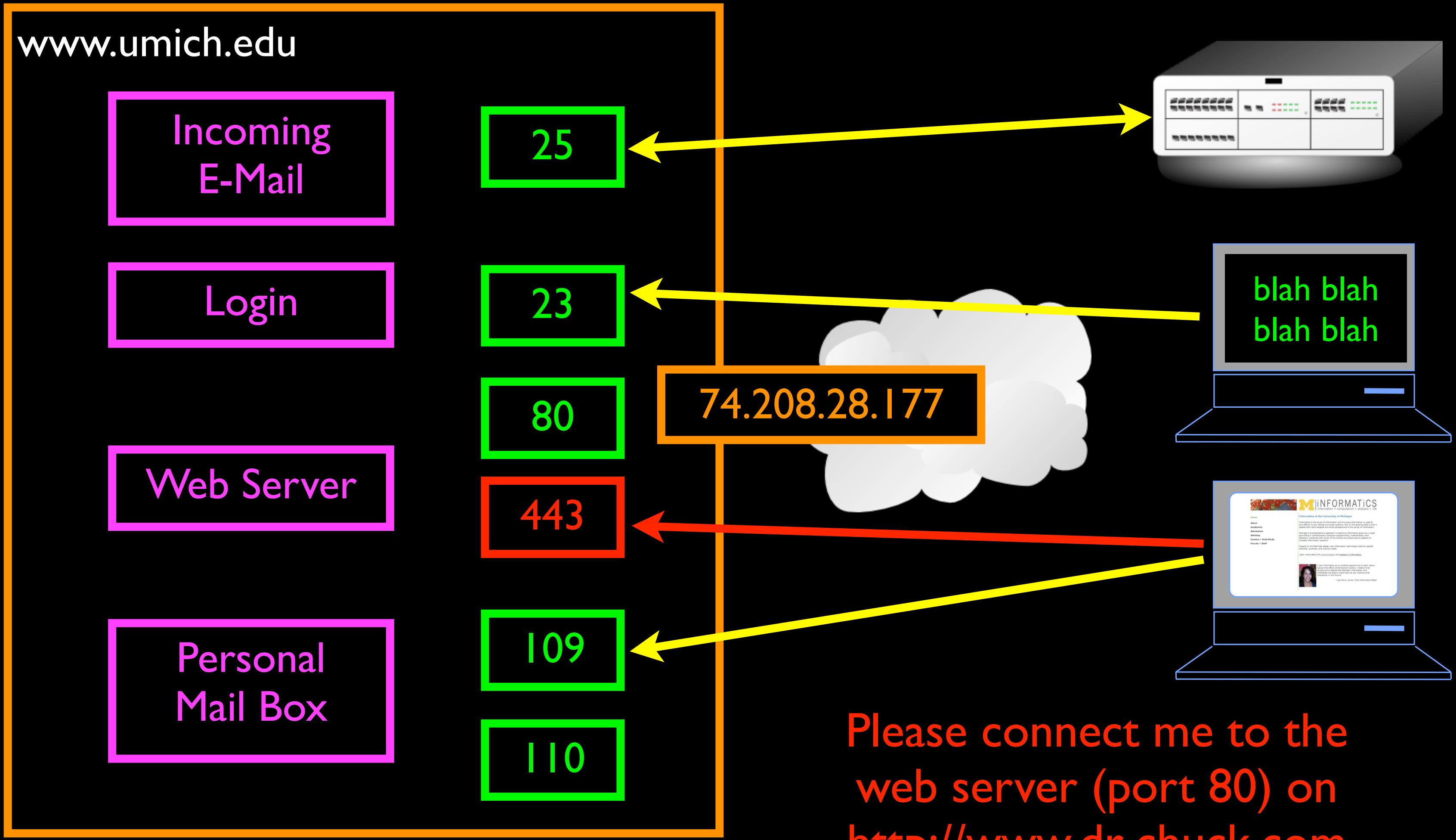


http://en.wikipedia.org/wiki/Internet_socket

TCP Port Numbers

- A port is an application-specific or process-specific software communications endpoint
- It allows multiple networked applications to coexist on the same server.
- There is a list of well-known TCP port numbers

http://en.wikipedia.org/wiki/TCP_and_UDP_port



www.umich.edu

Incoming E-Mail

25

Login

23

80

Web Server

443

Personal Mail Box

109

110

74.208.28.177

blah blah
blah blah

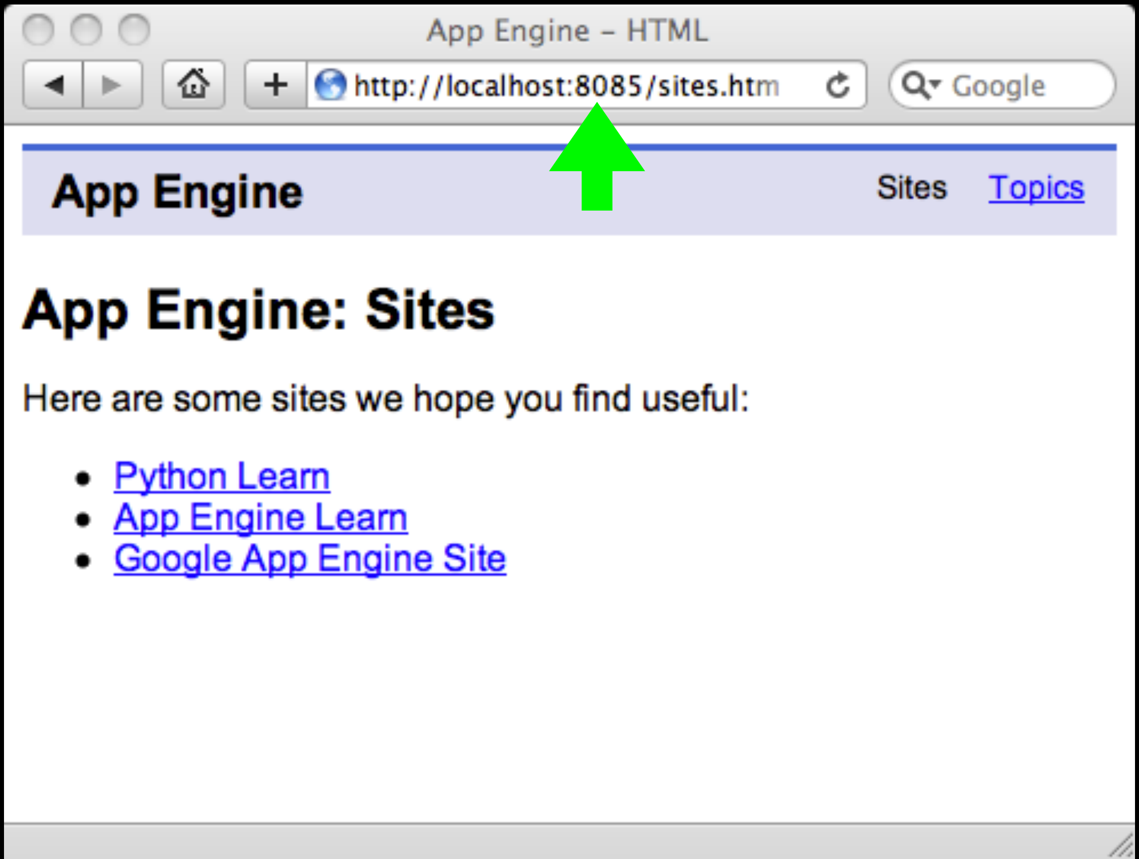
Please connect me to the web server (port 80) on http://www.dr-chuck.com

Clipart: http://www.clker.com/search/networksym/1

Common TCP Ports

- Telnet (23) - Login
- SSH (22) - Secure Login
- HTTP (80)
- HTTPS (443) - Secure
- SMTP (25) (Mail)
- IMAP (143/220/993) - Mail Retrieval
- POP (109/110) - Mail Retrieval
- DNS (53) - Domain Name
- FTP (21) - File Transfer

http://en.wikipedia.org/wiki/List_of_TCP_and_UDP_port_numbers



App Engine: Sites

Here are some sites we hope you find useful:

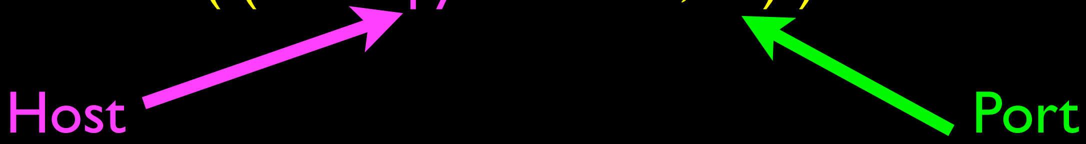
- Python Learn
- App Engine Learn
- Google App Engine Site

Sometimes we see the port number in the URL if the web server is running on a "non-standard" port.

Sockets in Python

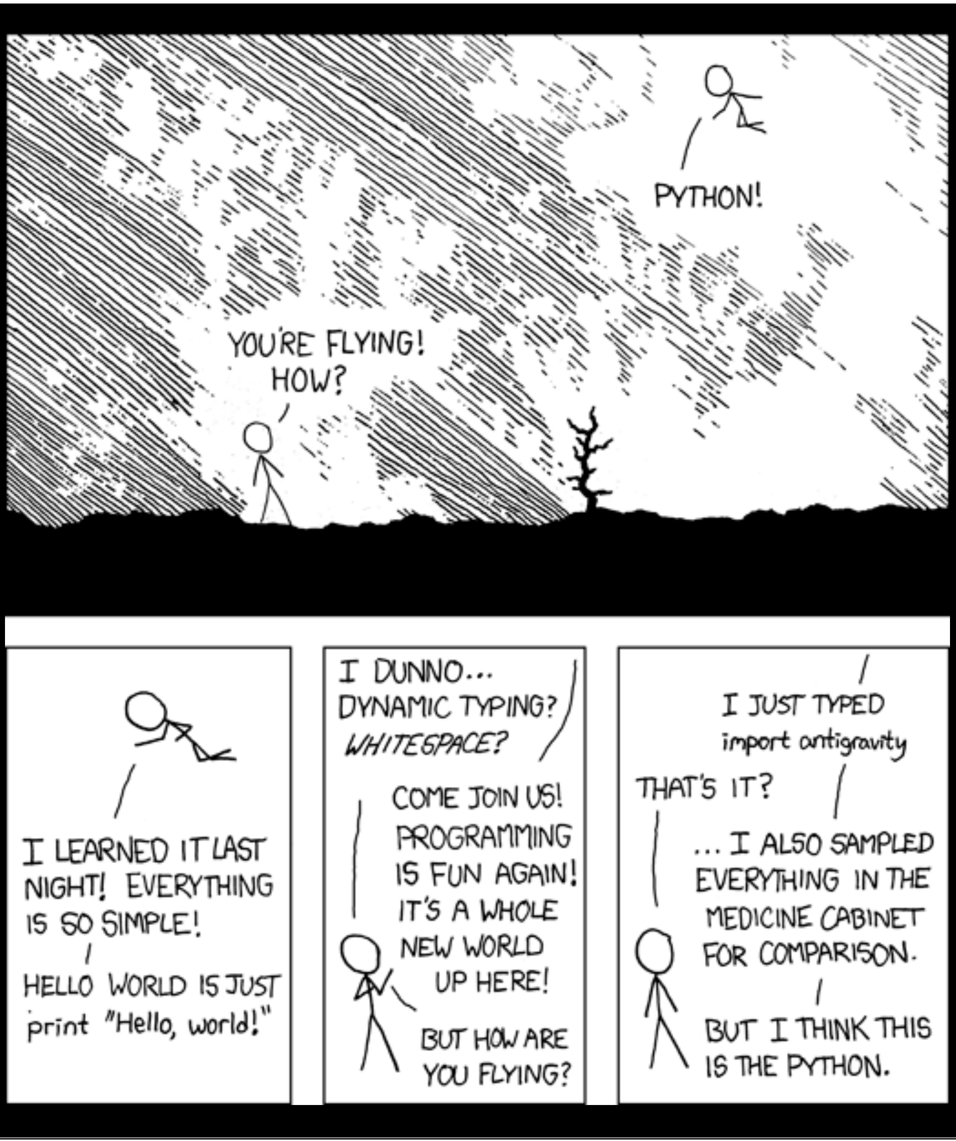
- Python has built-in support for TCP Sockets

```
import socket

mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect( ('www.py4inf.com', 80) )
```


Host

Port

http://www.pythonlearn.com/code/socket1.py


PYTHON!
YOU'RE FLYING! HOW?
I LEARNED IT LAST NIGHT! EVERYTHING IS SO SIMPLE!
HELLO WORLD IS JUST print "Hello, world!"
I DUNNO... DYNAMIC TYPING? WHITESPACE?
COME JOIN US! PROGRAMMING IS FUN AGAIN! IT'S A WHOLE NEW WORLD UP HERE!
BUT HOW ARE YOU FLYING?
I JUST TYPED import antigravity
THAT'S IT?
... I ALSO SAMPLED EVERYTHING IN THE MEDICINE CABINET FOR COMPARISON.
BUT I THINK THIS IS THE PYTHON.

http://xkcd.com/353/

HTTP

- The HyperText Transport Protocol is the set of rules to allow browsers to retrieve web documents from servers over the Internet

What is a Protocol?

- A set of rules that all parties follow for so we can predict each other's behavior
- And not bump into each other
  - On two-way roads in USA, drive on the right hand side of the road
  - On two-way roads in the UK drive on the left hand side of the road

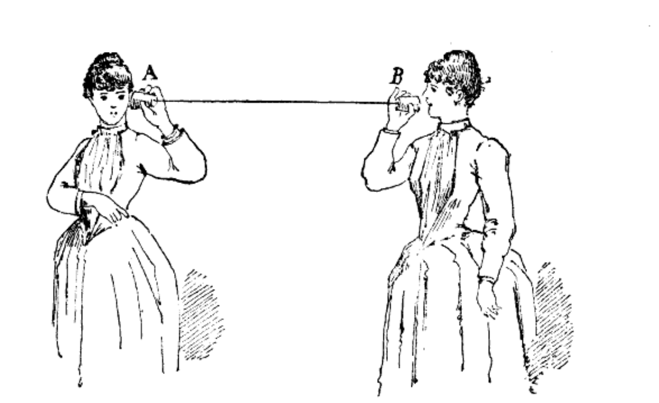


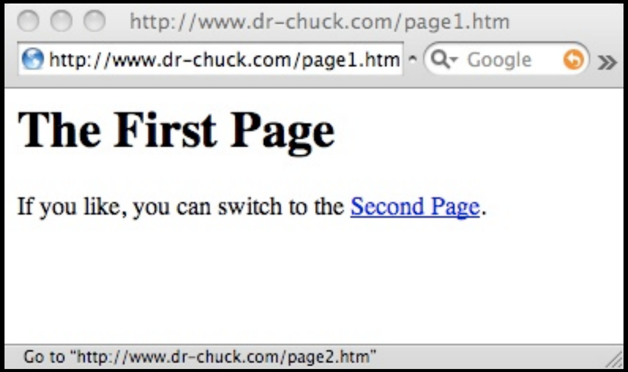
Fig. 76. Trådtelefon.

http://www.flickr.com/photos/kitcowan/2103850699/

http://en.wikipedia.org/wiki/Tin_can_telephone

Getting Data From The Server

- Each the user clicks on an anchor tag with an href= value to switch to a new page, the browser makes a connection to the web server and issues a "GET" request - to GET the content of the page at the specified URL
- The server returns the HTML document to the Browser which formats and displays the document to the user.



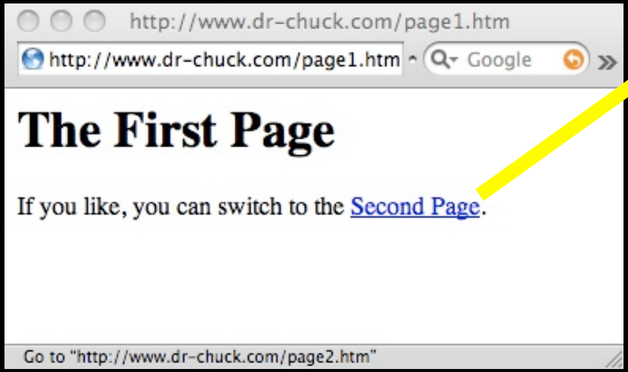
http://www.dr-chuck.com/page1.htm
http://www.dr-chuck.com/page1.htm
Google

The First Page

If you like, you can switch to the Second Page.

Go to "http://www.dr-chuck.com/page2.htm"

Browser

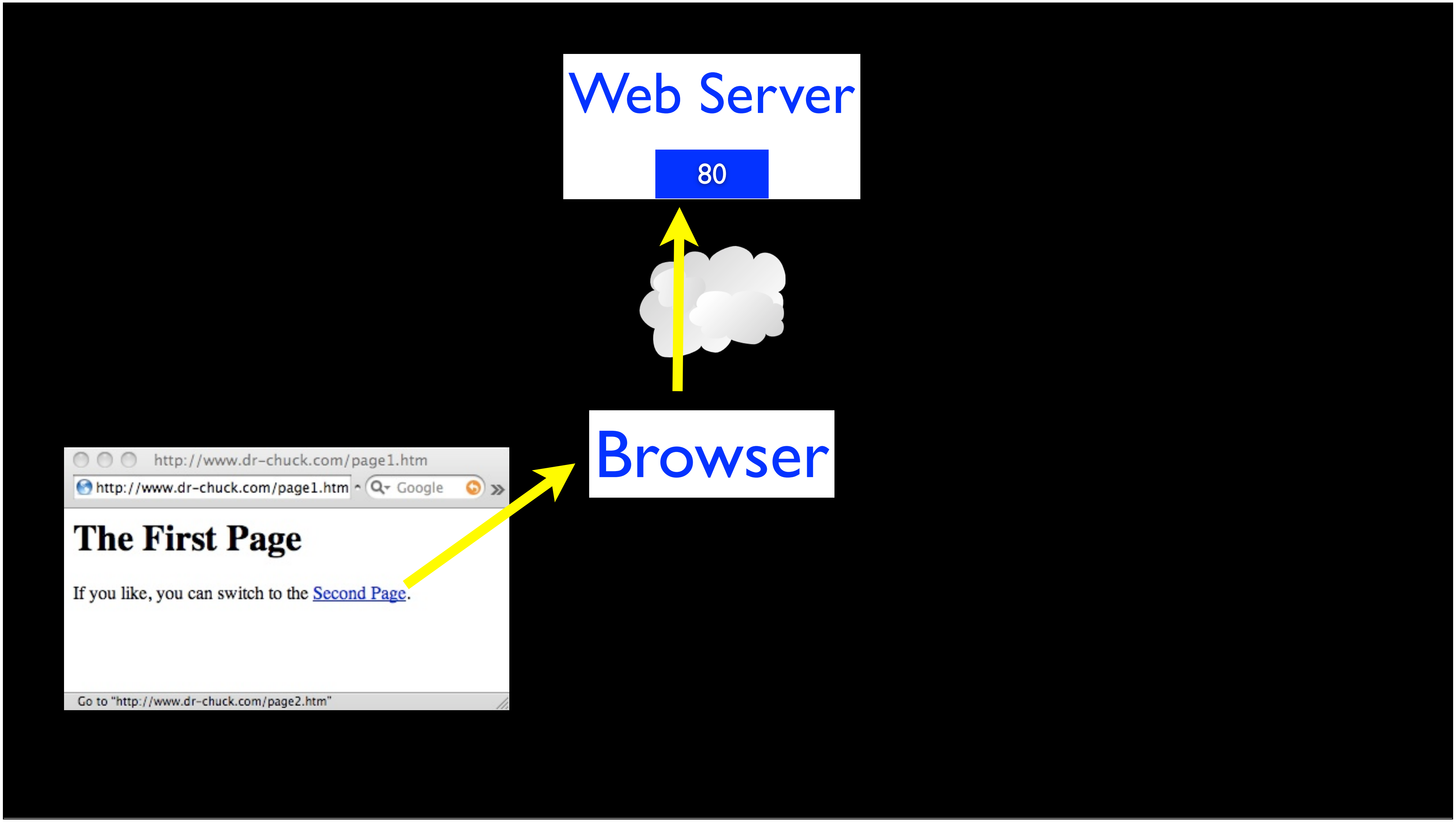


http://www.dr-chuck.com/page1.htm
http://www.dr-chuck.com/page1.htm
Google

The First Page

If you like, you can switch to the Second Page.

Go to "http://www.dr-chuck.com/page2.htm"



Web Server

80

Browser

http://www.dr-chuck.com/page1.htm

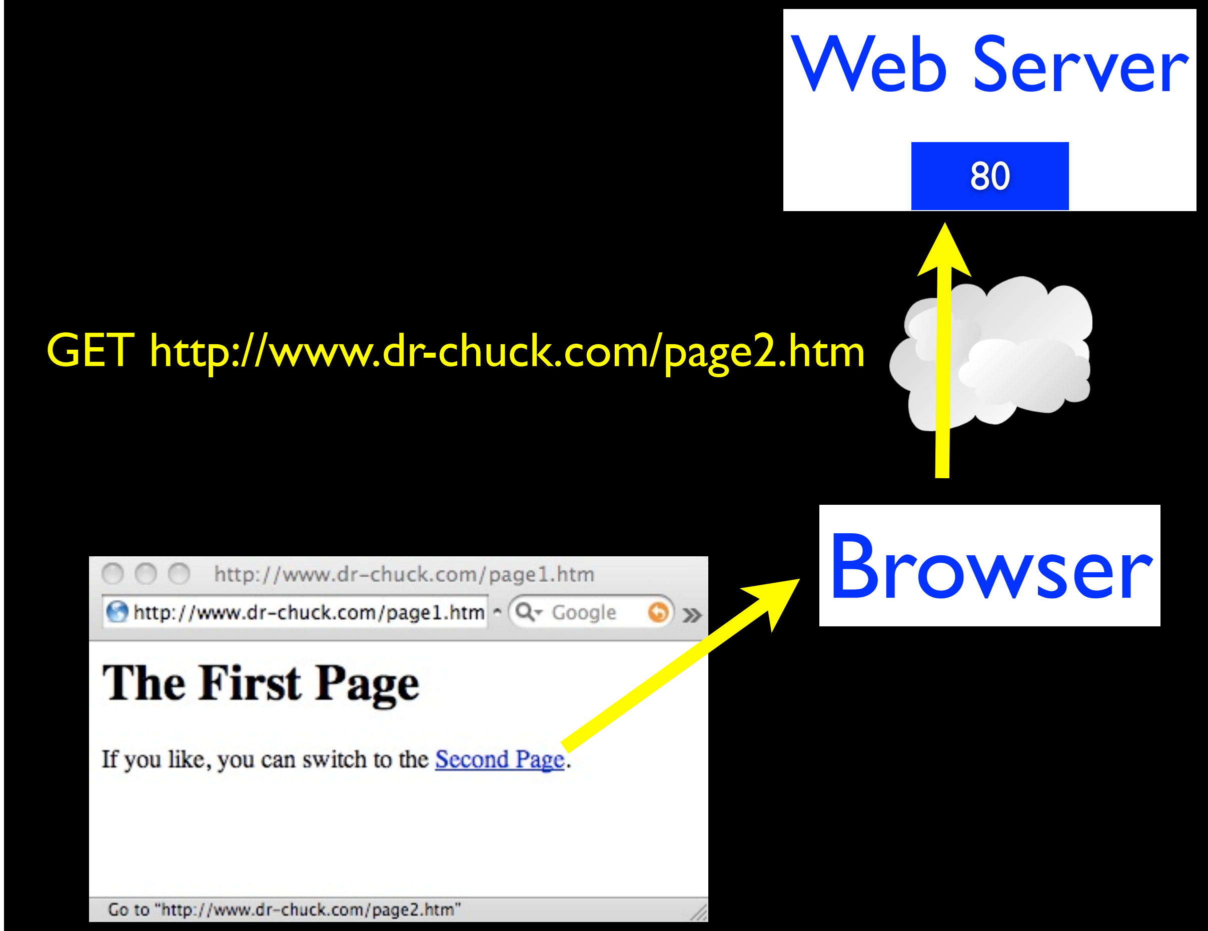
http://www.dr-chuck.com/page1.htm

Google

The First Page

If you like, you can switch to the Second Page.

Go to "http://www.dr-chuck.com/page2.htm"



Web Server

80

GET http://www.dr-chuck.com/page2.htm

Browser

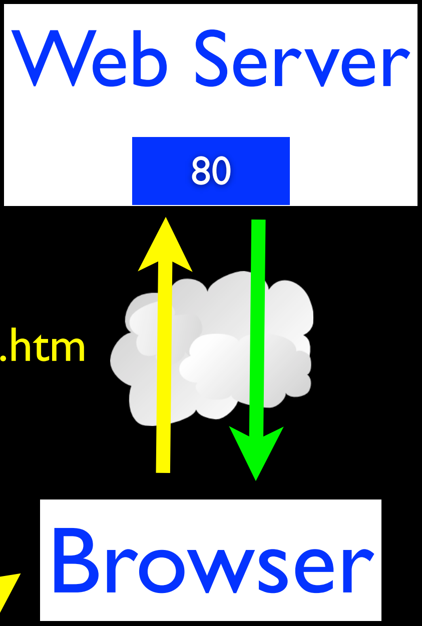
http://www.dr-chuck.com/page1.htm

http://www.dr-chuck.com/page1.htm

The First Page

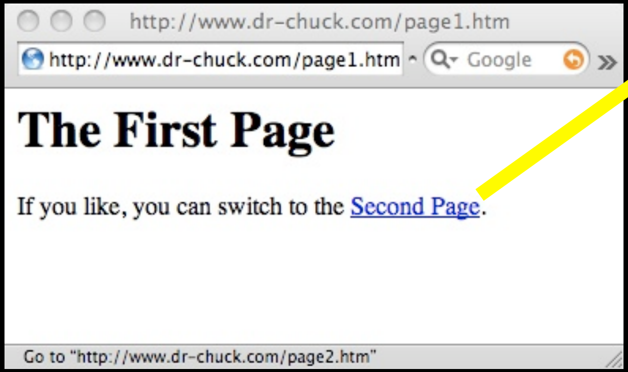
If you like, you can switch to the Second Page.

Go to "http://www.dr-chuck.com/page2.htm"



GET http://www.dr-chuck.com/page2.htm

```
<h1>The Second Page</h1>
<p>
If you like, you can switch
back to the
<a href="page1.htm">
First Page</a>.
</p>
```



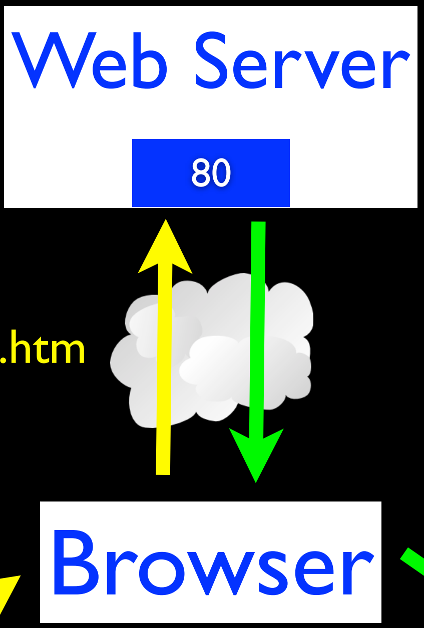
http://www.dr-chuck.com/page1.htm

http://www.dr-chuck.com/page1.htm

The First Page

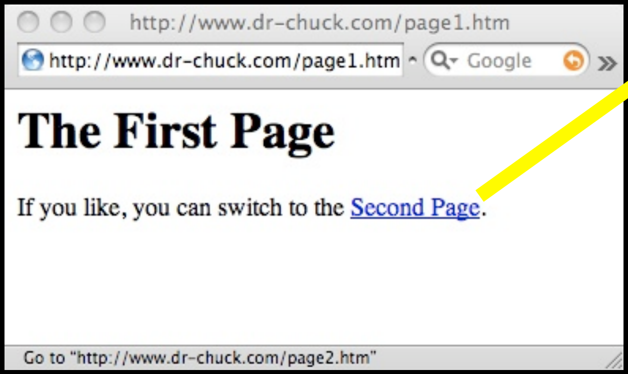
If you like, you can switch to the Second Page.

Go to "http://www.dr-chuck.com/page2.htm"



GET http://www.dr-chuck.com/page2.htm

```
<h1>The Second Page</h1>
<p>
If you like, you can switch
back to the
<a href="page1.htm">
First Page</a>.
</p>
```

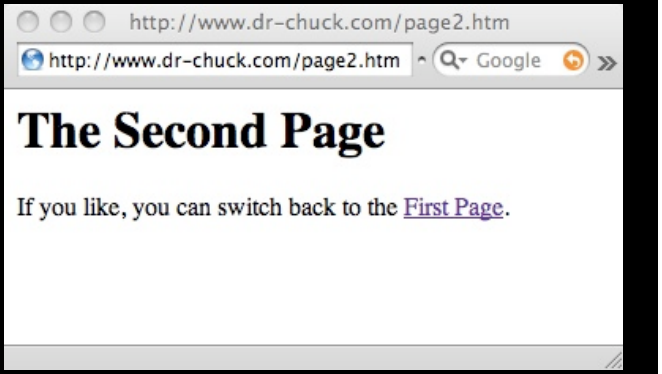


http://www.dr-chuck.com/page1.htm

The First Page

If you like, you can switch to the Second Page.

Go to "http://www.dr-chuck.com/page2.htm"



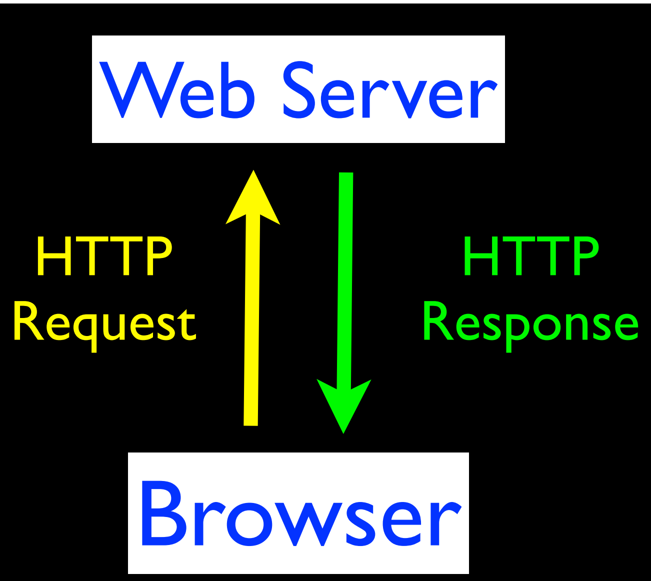
http://www.dr-chuck.com/page2.htm

The Second Page

If you like, you can switch back to the First Page.

Lets Write a Web Browser!

“Hacking” HTTP

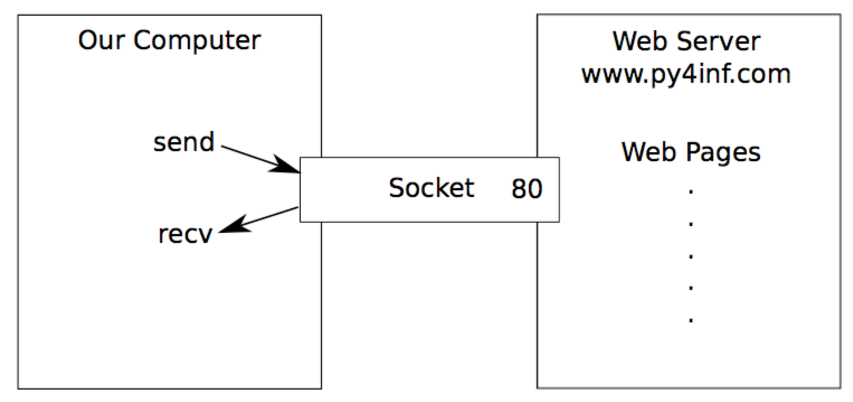


```
$ telnet www.dr-chuck.com 80
Trying 74.208.28.177...
Connected to www.dr-chuck.com.
Escape character is '^]'.
GET http://www.dr-chuck.com/page1.htm
<h1>The First Page</h1>
<p>
If you like, you can switch to the
<a href="http://www.dr-chuck.com/page2.htm">
Second Page</a>.
</p>
```

Port 80 is the non-encrypted HTTP port

An HTTP Request in Python

```
import socket

mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect(('www.py4inf.com', 80))
mysock.send('GET http://www.py4inf.com/code/romeo.txt HTTP/1.0\n\n')

while True:
    data = mysock.recv(512)
    if ( len(data) < 1 ) :
        break
    print data

mysock.close()
```



http://www.pythonlearn.com/code/socket2.py

```
HTTP/1.1 200 OK
Date: Sun, 14 Mar 2010 23:52:41 GMT
Server: Apache
Last-Modified: Tue, 29 Dec 2009 01:31:22 GMT
ETag: "143c1b33-a7-4b395bea"
Accept-Ranges: bytes
Content-Length: 167
Connection: close
Content-Type: text/plain

But soft what light through yonder window breaks
It is the east and Juliet is the sun
Arise fair sun and kill the envious moon
Who is already sick and pale with grief
```

HTTP Header

```
while True:
    data = mysock.recv(512)
    if ( len(data) < 1 ) :
        break
    print data
```

HTTP Body

http://www.pythonlearn.com/code/socket2.py

Making HTTP Easier With urllib

Using urllib in Python

- Since HTTP is so common, we have a library that does all the socket work for us and makes web pages look like a file

```
import urllib

fhand = urllib.urlopen('http://www.py4inf.com/code/romeo.txt')

for line in fhand:
    print line.strip()
```

http://www.pythonlearn.com/code/urllib1.py

urllib1.py

```
import urllib

fhand = urllib.urlopen('http://www.py4inf.com/code/romeo.txt')

for line in fhand:
    print line.strip()
```

```
But soft what light through yonder window breaks
It is the east and Juliet is the sun
Arise fair sun and kill the envious moon
Who is already sick and pale with grief
```

http://docs.python.org/library/urllib.html

urllib1.py

Like a file...

```
import urllib

fhand = urllib.urlopen('http://www.py4inf.com/code/romeo.txt')

counts = dict()
for line in fhand:
    words = line.split()
    for word in words:
        counts[word] = counts.get(word,0) + 1
print counts
```

urlwords.py

Reading Web Pages

```
import urllib

fhand = urllib.urlopen('http://www.dr-chuck.com/page1.htm')
for line in fhand:
    print line.strip()
```

```
<h1>The First Page</h1>
<p>
If you like, you can switch to the
<a href="http://www.dr-chuck.com/page2.htm">
Second Page</a>.
</p>
```

urllib1.py

Going from one page to another...

```
import urllib

fhand = urllib.urlopen('http://www.dr-chuck.com/page1.htm')
for line in fhand:
    print line.strip()
```

```
<h1>The First Page</h1>
<p>
If you like, you can switch to the
<a href="http://www.dr-chuck.com/page2.htm">
Second Page</a>.
</p>
```

Google

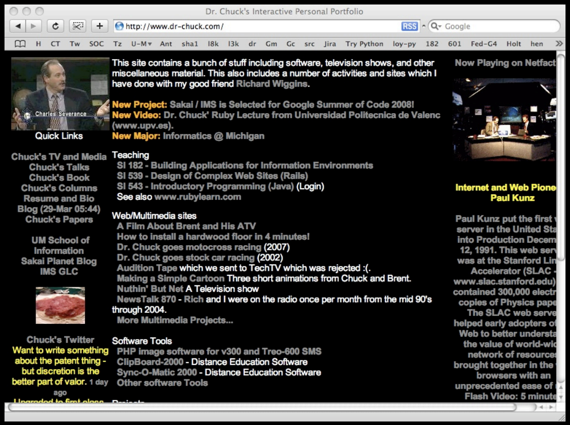
```
import urllib

fhand = urllib.urlopen('http://www.dr-chuck.com/page1.htm')
for line in fhand:
    print line.strip()
```

Parsing HTML
(a.k.a Web Scraping)

What is Web Scraping?

- When a program or script pretends to be a browser and retrieves web pages, looks at those web pages, extracts information and then looks at more web pages.
- Search engines scrape web pages - we call this “spidering the web” or “web crawling”

http://en.wikipedia.org/wiki/Web_scraping
http://en.wikipedia.org/wiki/Web_crawler


Dr. Chuck's Interactive Personal Portfolio
http://www.dr-chuck.com/
RSS
Google
H CT Tw SOC Tz U-M Ant sha1 l8k l3k dr Gm Gc src Jira Try Python loy-py 182 601 Fed-G4 Holt hen
Charles Severance

Quick Links

Chuck's TV and Media
Chuck's Talks
Chuck's Book
Chuck's Columns
Resume and Bio
Blog (29-Mar 05:44)
Chuck's Papers

UM School of Information
Sakai Planet Blog
IMS GLC

Chuck's Twitter
Want to write something about the patent thing - but discretion is the better part of valor. 1 day ago

This site contains a bunch of stuff including software, television shows, and other miscellaneous material. This also includes a number of activities and sites which I have done with my good friend Richard Wiggins.

New Project: Sakai / IMS is Selected for Google Summer of Code 2008!
New Video: Dr. Chuck' Ruby Lecture from Universidad Politecnica de Valenc (www.upv.es).
New Major: Informatics @ Michigan

Teaching
SI 182 - Building Applications for Information Environments
SI 539 - Design of Complex Web Sites (Rails)
SI 543 - Introductory Programming (Java) (Login)
See also www.rubylearn.com

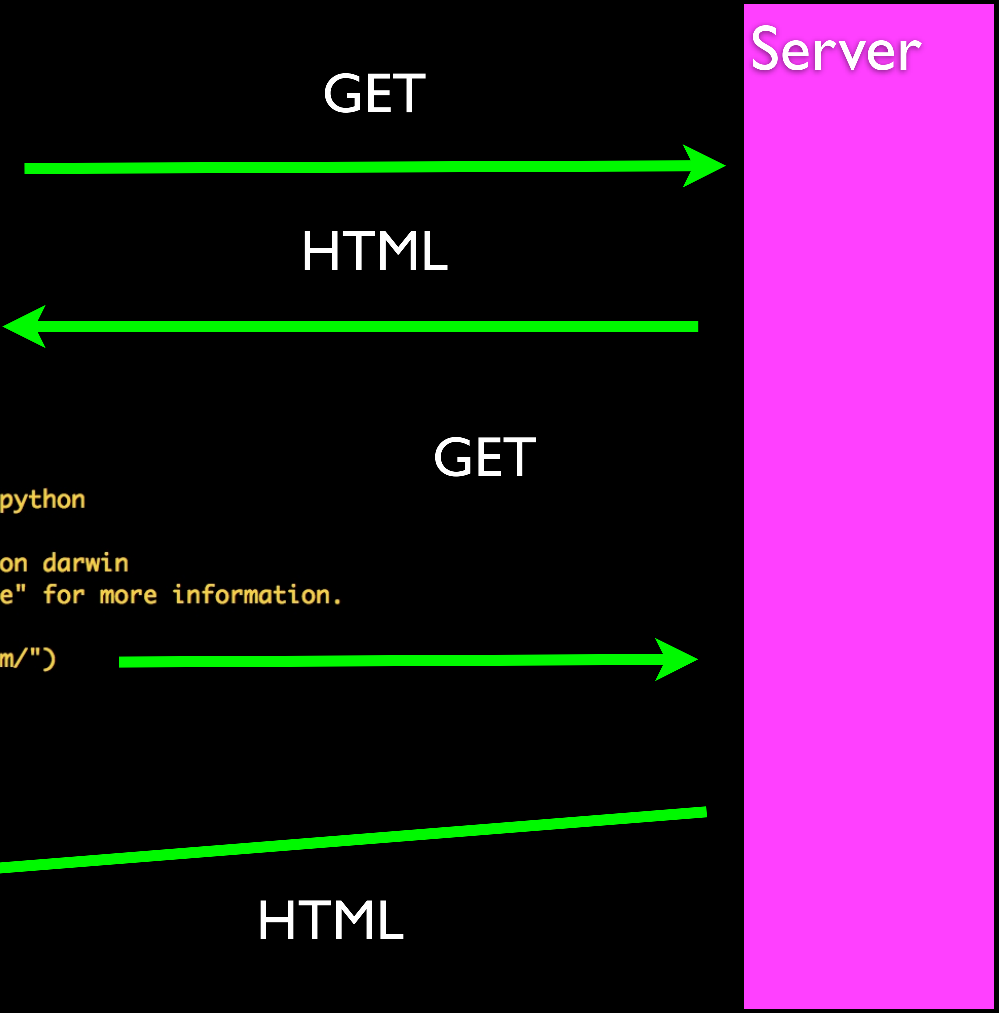
Web/Multimedia sites
A Film About Brent and His ATV
How to install a hardwood floor in 4 minutes!
Dr. Chuck goes motocross racing (2007)
Dr. Chuck goes stock car racing (2002)
Audition Tape which we sent to TechTV which was rejected :(.
Making a Simple Cartoon Three short animations from Chuck and Brent.
Nuthin' But Net A Television show
NewsTalk 870 - Rich and I were on the radio once per month from the mid 90's through 2004.
More Multimedia Projects...

Software Tools
PHP image software for v300 and Treo-600 SMS
ClipBoard-2000 - Distance Education Software
Sync-O-Matic 2000 - Distance Education Software
Other software Tools

Now Playing on Netfact

Internet and Web Pione
Paul Kunz

Paul Kunz put the first
server in the United Sta
into Production Decem
12, 1991. This web ser
was at the Stanford Lin
Accelerator (SLAC -
www.slac.stanford.edu)
contained 300,000 electr
copies of Physics pape
The SLAC web serve
helped early adopters of
Web to better understa
the value of world-wid
network of resources
brought together in the
browsers with an
unprecedented ease of
Flash Video: 5 minute



```
charles-severances-macbook-air:Scraping csev$ python
Python 2.5 (r25:51918, Sep 19 2006, 08:49:13)
[GCC 4.0.1 (Apple Computer, Inc. build 5341)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import urllib
>>> f = urllib.urlopen("http://www.dr-chuck.com/")
>>> contents = f.read()
>>> f.close()
>>> print len(contents)
95328
>>> print contents[0:30]
<html>
<head>
  <title>Dr. C
>>> 
```

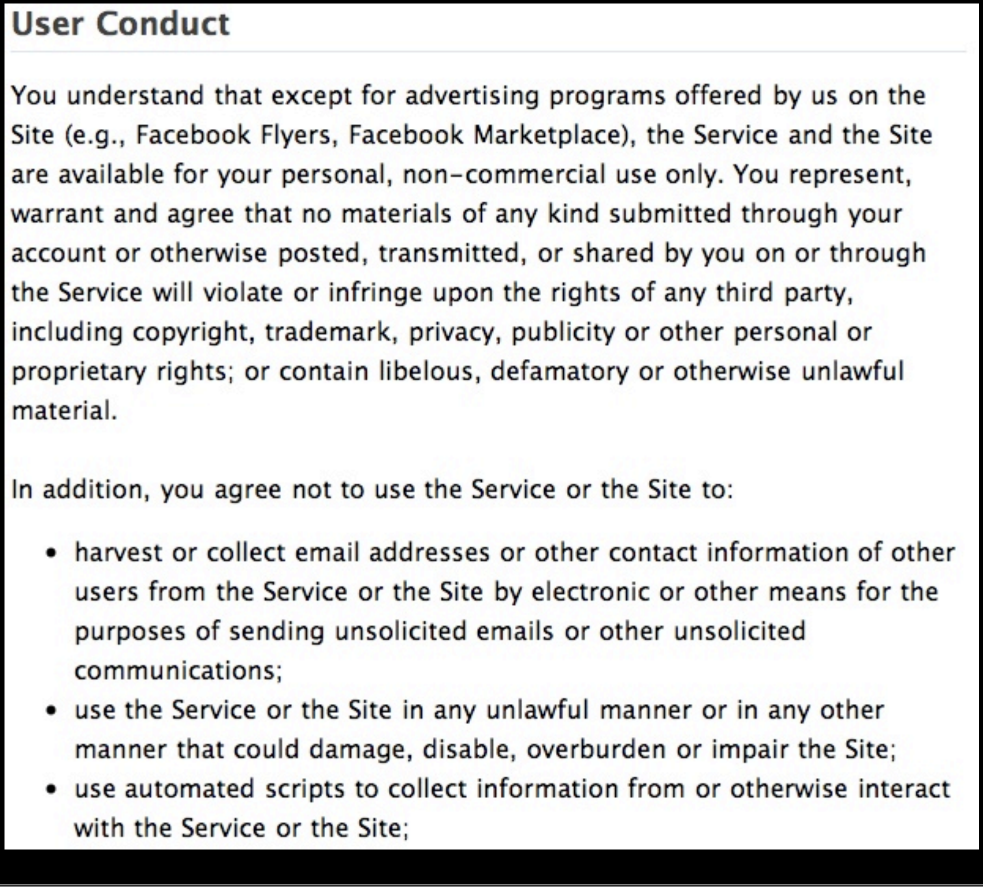
Why Scrape?

- Pull data - particularly social data - who links to who?
- Get your own data back out of some system that has no “export capability”
- Monitor a site for new information
- Spider the web to make a database for a search engine

Scraping Web Pages

- There is some controversy about web page scraping and some sites are a bit snippy about it.
    - Google: facebook scraping block
- Republishing copyrighted information is not allowed
- Violating terms of service is not allowed

http://www.facebook.com/terms.php



User Conduct

You understand that except for advertising programs offered by us on the Site (e.g., Facebook Flyers, Facebook Marketplace), the Service and the Site are available for your personal, non-commercial use only. You represent, warrant and agree that no materials of any kind submitted through your account or otherwise posted, transmitted, or shared by you on or through the Service will violate or infringe upon the rights of any third party, including copyright, trademark, privacy, publicity or other personal or proprietary rights; or contain libelous, defamatory or otherwise unlawful material.

In addition, you agree not to use the Service or the Site to:

- harvest or collect email addresses or other contact information of other users from the Service or the Site by electronic or other means for the purposes of sending unsolicited emails or other unsolicited communications;
- use the Service or the Site in any unlawful manner or in any other manner that could damage, disable, overburden or impair the Site;
- use automated scripts to collect information from or otherwise interact with the Service or the Site;

The Easy Way - Beautiful Soup

- You could do string searches the hard way
- Or use the free software called BeautifulSoup from www.crummy.com

http://www.crummy.com/software/BeautifulSoup/

Place the BeautifulSoup.py file in the same folder as your Python code...

```
import urllib
from BeautifulSoup import *

url = raw_input('Enter - ')
html = urllib.urlopen(url).read()
soup = BeautifulSoup(html)

# Retrieve a list of the anchor tags
# Each tag is like a dictionary of HTML attributes

tags = soup('a')

for tag in tags:
   print tag.get('href', None)
```

urllinks.py

```
<h1>The First Page</h1>
<p>
If you like, you can switch to the
<a href="http://www.dr-chuck.com/page2.htm">
Second Page</a>.
</p>
```

```
html = urllib.urlopen(url).read()
soup = BeautifulSoup(html)
tags = soup('a')
for tag in tags:
   print tag.get('href', None)
```

```
python urllinks.py
Enter - http://www.dr-chuck.com/page1.htm
http://www.dr-chuck.com/page2.htm
```

Summary

- The TCP/IP gives us pipes / sockets between applications
- We designed application protocols to make use of these pipes
- HyperText Transport Protocol (HTTP) is a simple yet powerful protocol
- Python has good support for sockets, HTTP, and HTML parsing