Overview of the World Wide Web

UT CS361S – Network Security and Privacy

Spring 2021

Lecture Notes

What is the World Wide Web?

- ***Internet*** - globally interconnected network system
- ***World Wide Web*** - HTTP-based content, apps, "ecosystem"

Key Tech: *Domain Name System (DNS)*

- IPv4 addresses were hard to remember/use
- IPv6 are worse
- Humans need semantically meaningful addresses
- DNS maps IP addresses to ***domain names***

Basic Idea

Where is "google.com?"

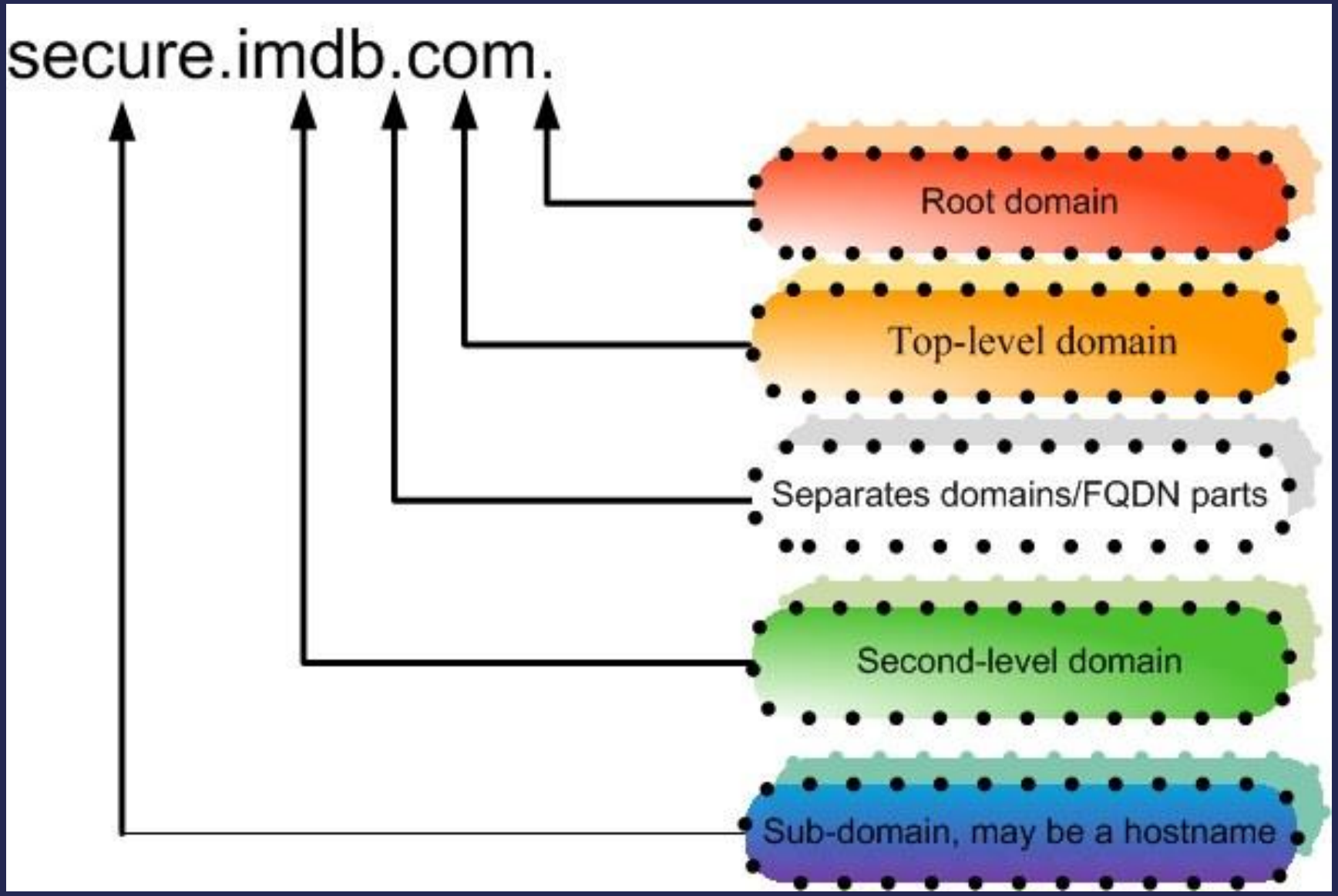
142.250.138.138

DNS SERVER

GET / HTTP/1.1

Google

142.250.138.138
(google.com)


secure.imdb.com.
Root domain
Top-level domain
Separates domains/FQDN parts
Second-level domain
Sub-domain, may be a hostname

Top Level Domains (TLDs)

- Generic Top Level Domain (gTLD) - .com, .net, et
- Country code Top Level Domain (ccTLD) - .uk

TLD Name Management

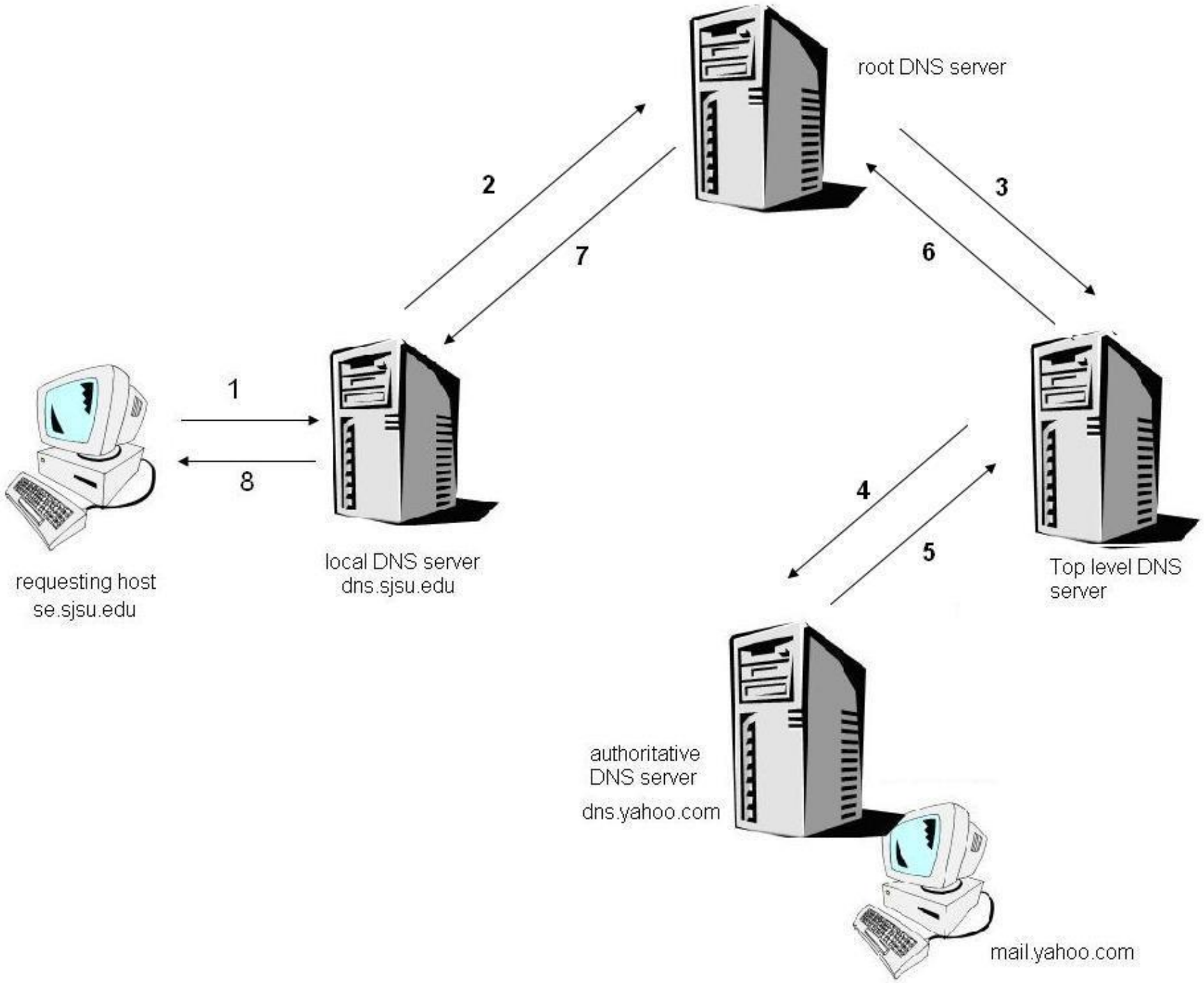
- Registrars administer TLDs
- For gTLDs, this is a business with pros and cons
- Registrars authorize “domain name registrars”

Domain Name Registration

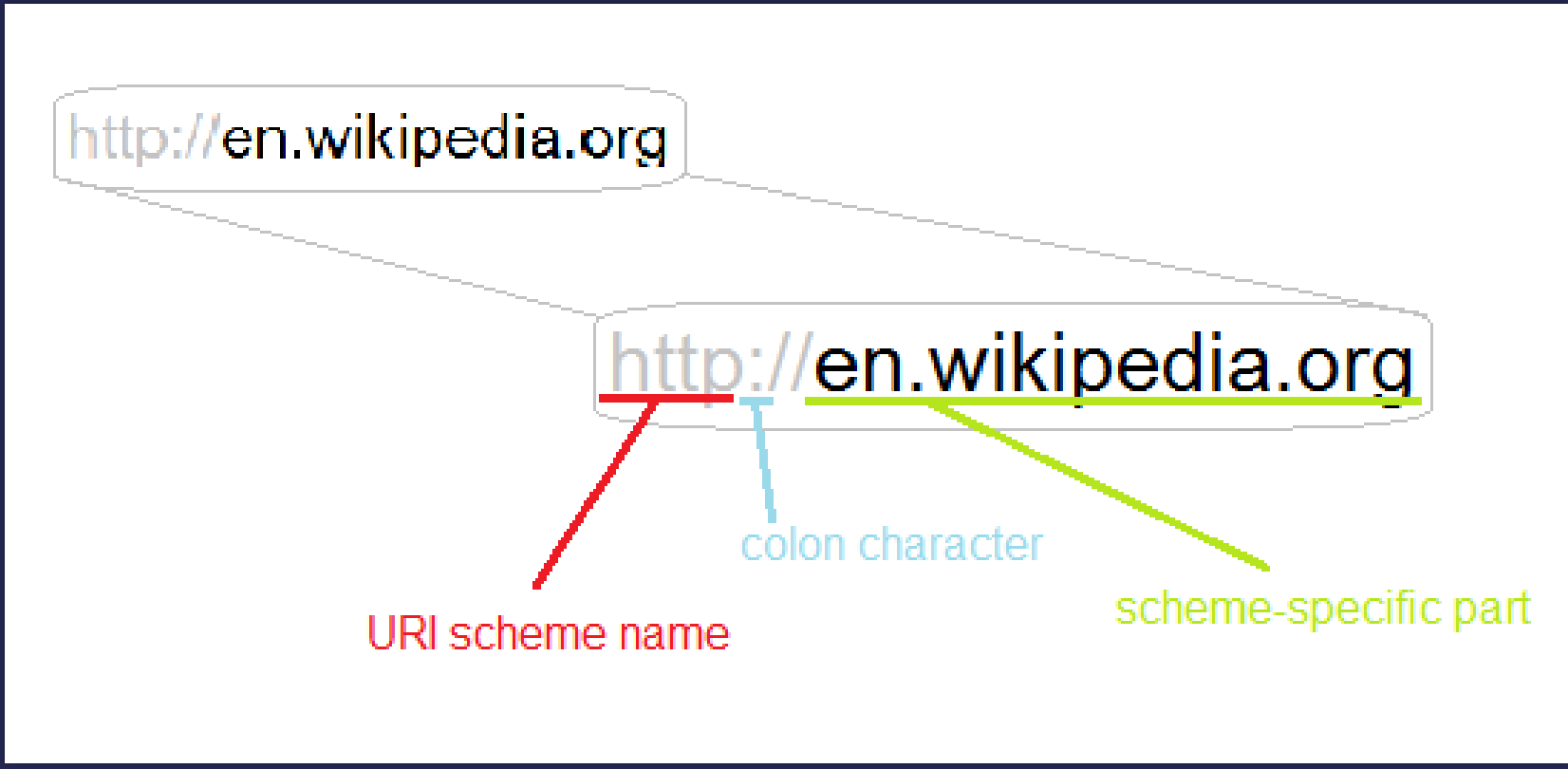
- Party requests SLD + TLD from domain name reseller
- Party submits “whois” information (contact info)
- Registrar verifies that name is available
- Registrar stores relevant data in registry and DNS servers

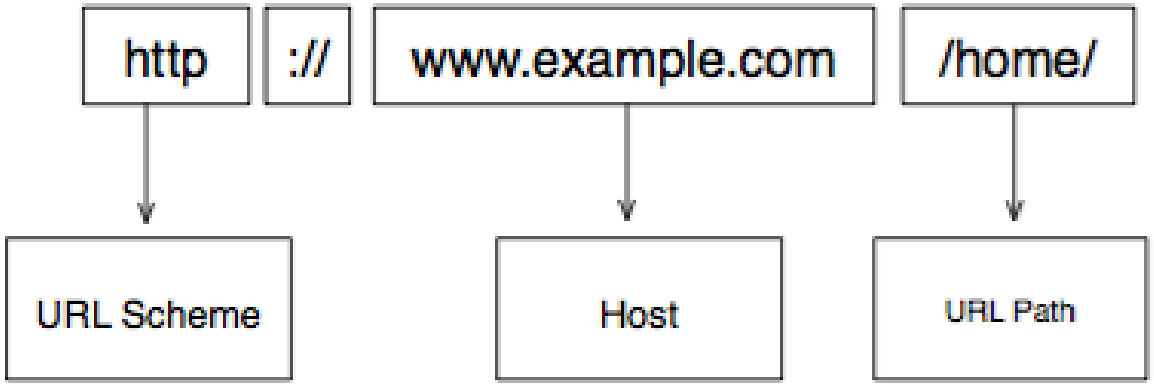
DNS and Address Resolution

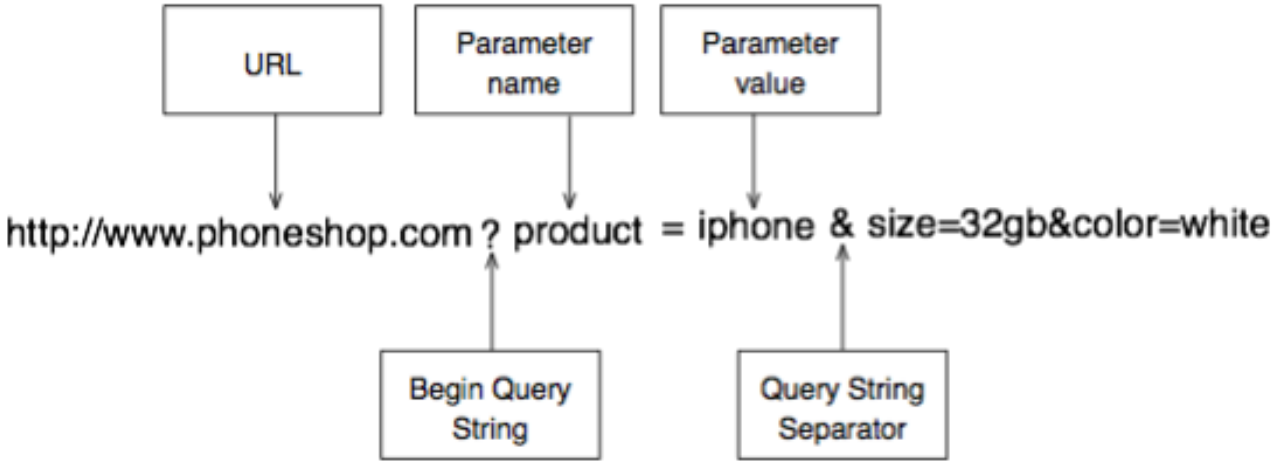
- DNS is a ***recursive*** and ***hierarchical*** process
- Recursive – DNS server searches another DNS server
- Hierarchical –
  - Root Domain to TLD
  - TLD to Subdomain


root DNS server
2
3
7
6
1
8
4
5
requesting host
se.sjsu.edu
local DNS server
dns.sjsu.edu
Top level DNS
server
authoritative
DNS server
dns.yahoo.com
mail.yahoo.com

Uniform Resource Identifiers (URIs)




http
://
www.example.com
/home/
URL Scheme
Host
URL Path

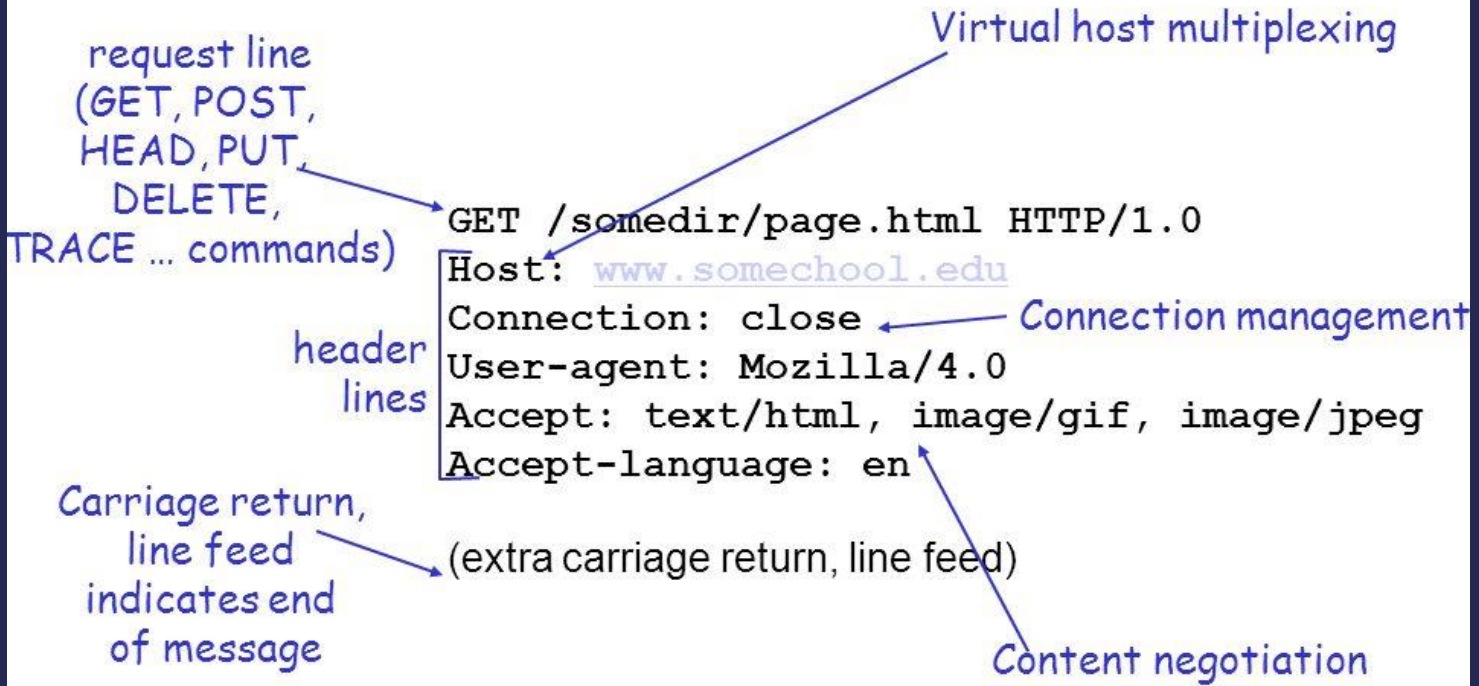

URL
Parameter name
Parameter value
http://www.phoneshop.com ? product = iphone & size=32gb&color=white
Begin Query String
Query String Separator

Absolute vs Relative URI

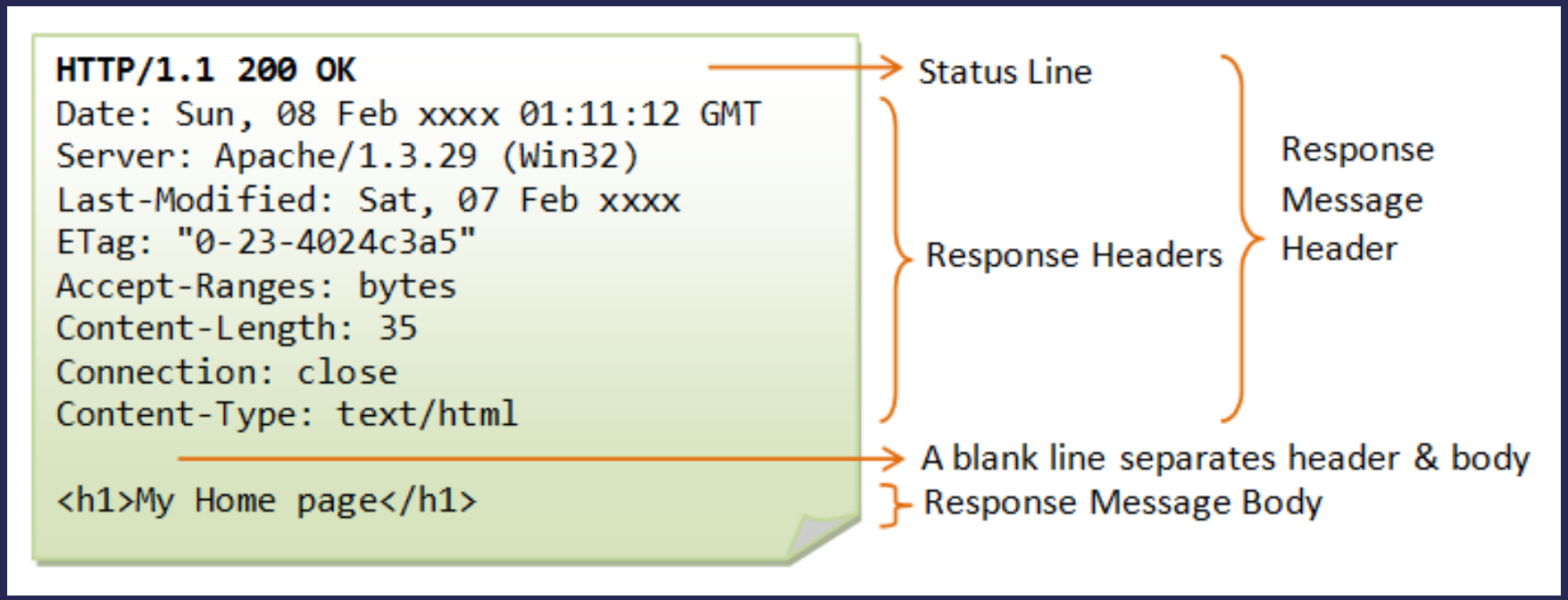
- ***Absolute*** paths begin with **<scheme>://host/**
  - e.g., *http://www.google.com/*
- Everything else is ***relative***
  - e.g., */not/an/absolute/path*
  - The scheme and host are determined by context

HTTP Request

HTTP Request Message Example: GET



HTTP Response



Static Web Page Example

<HTML>

<BODY>

<H1>Simple Web Page</H1>

<IMG SRC="/images/image1.jpg">

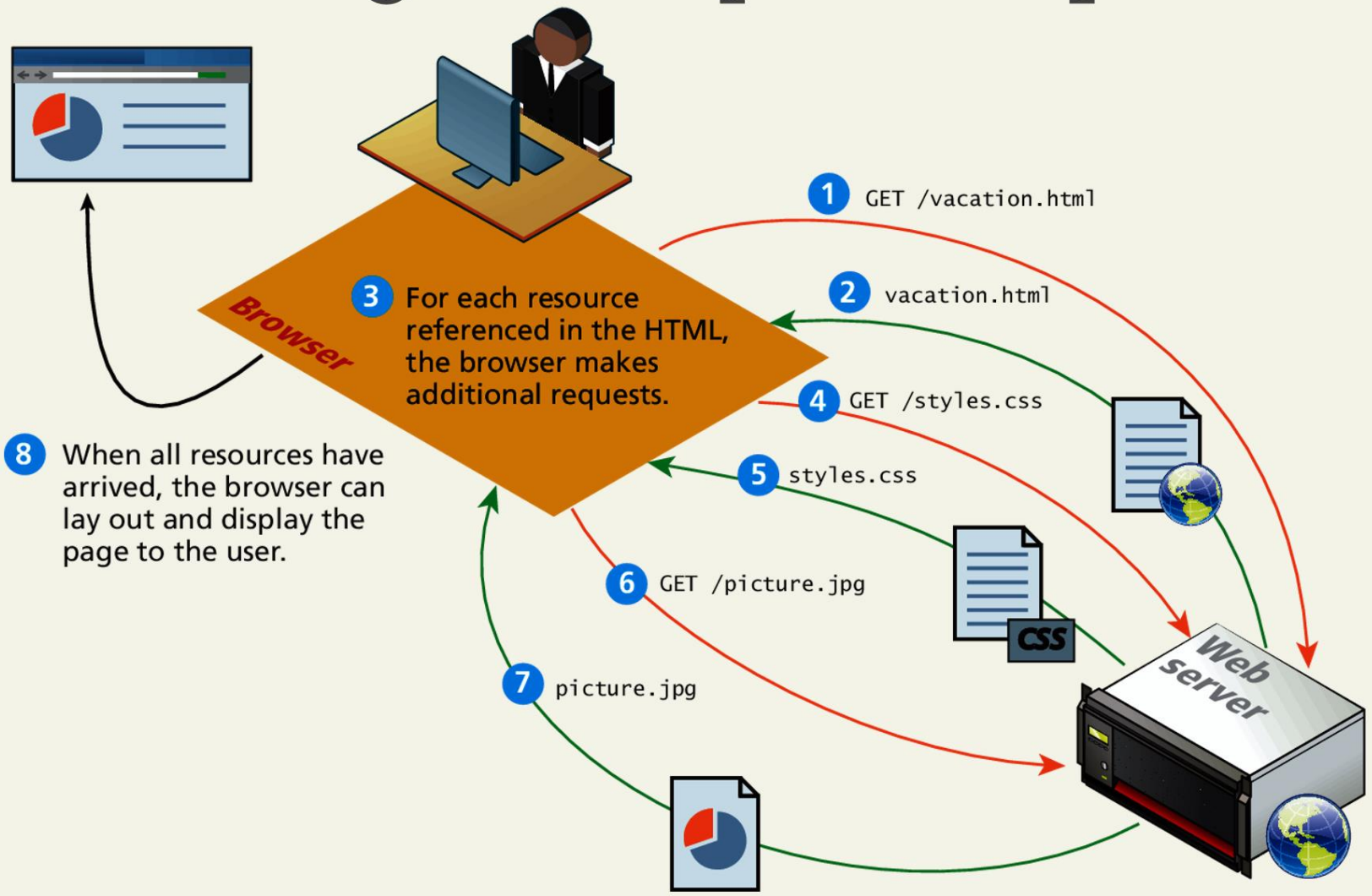
</BODY>

</HTML>

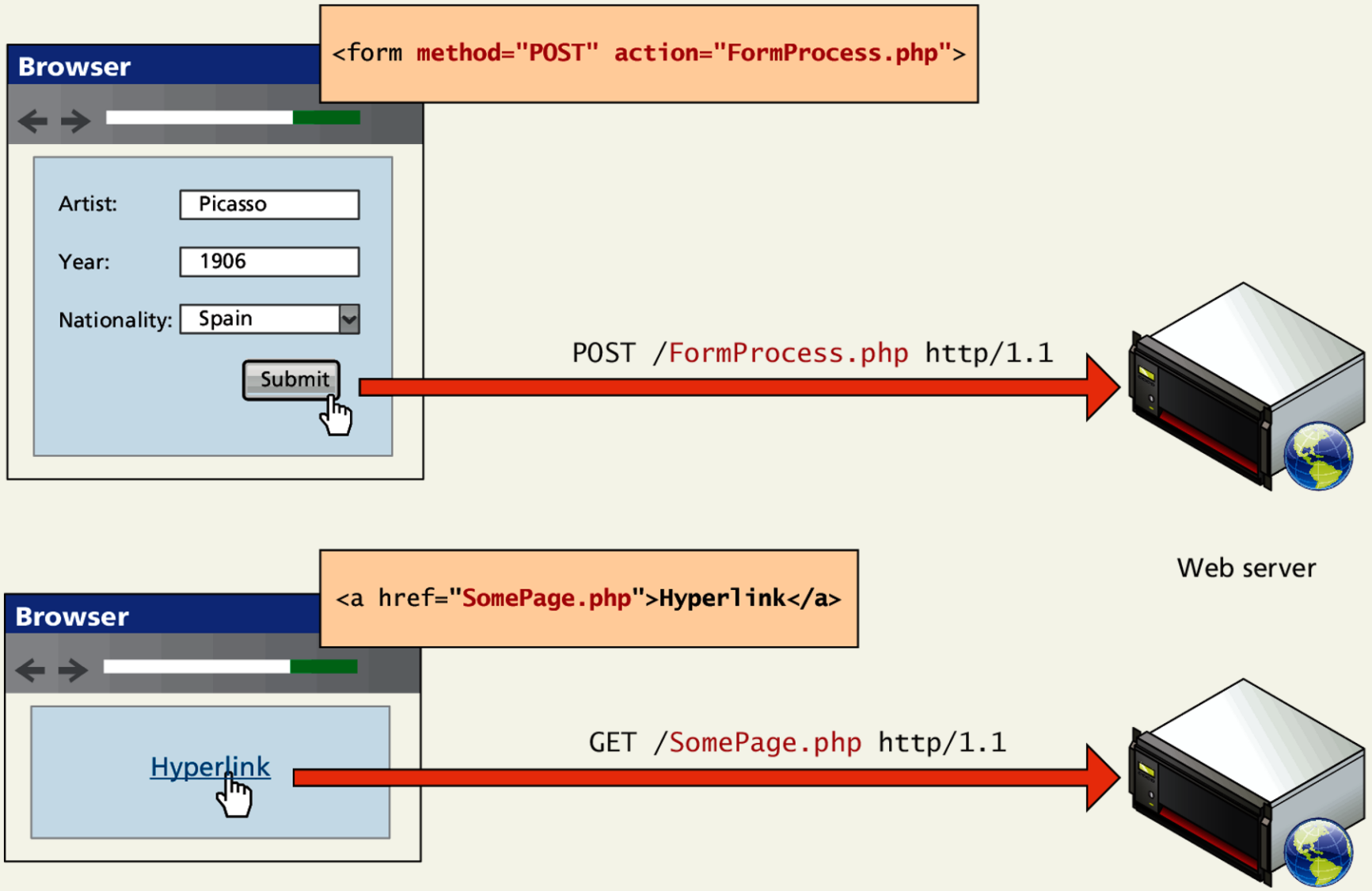
Rendering a Web Page

- Browser requests HTML “root” page
- Root page has links for images, etc
- Browser requests embedded objects
- Browser integrates and renders objects

Browser parsing HTML and making subsequent requests



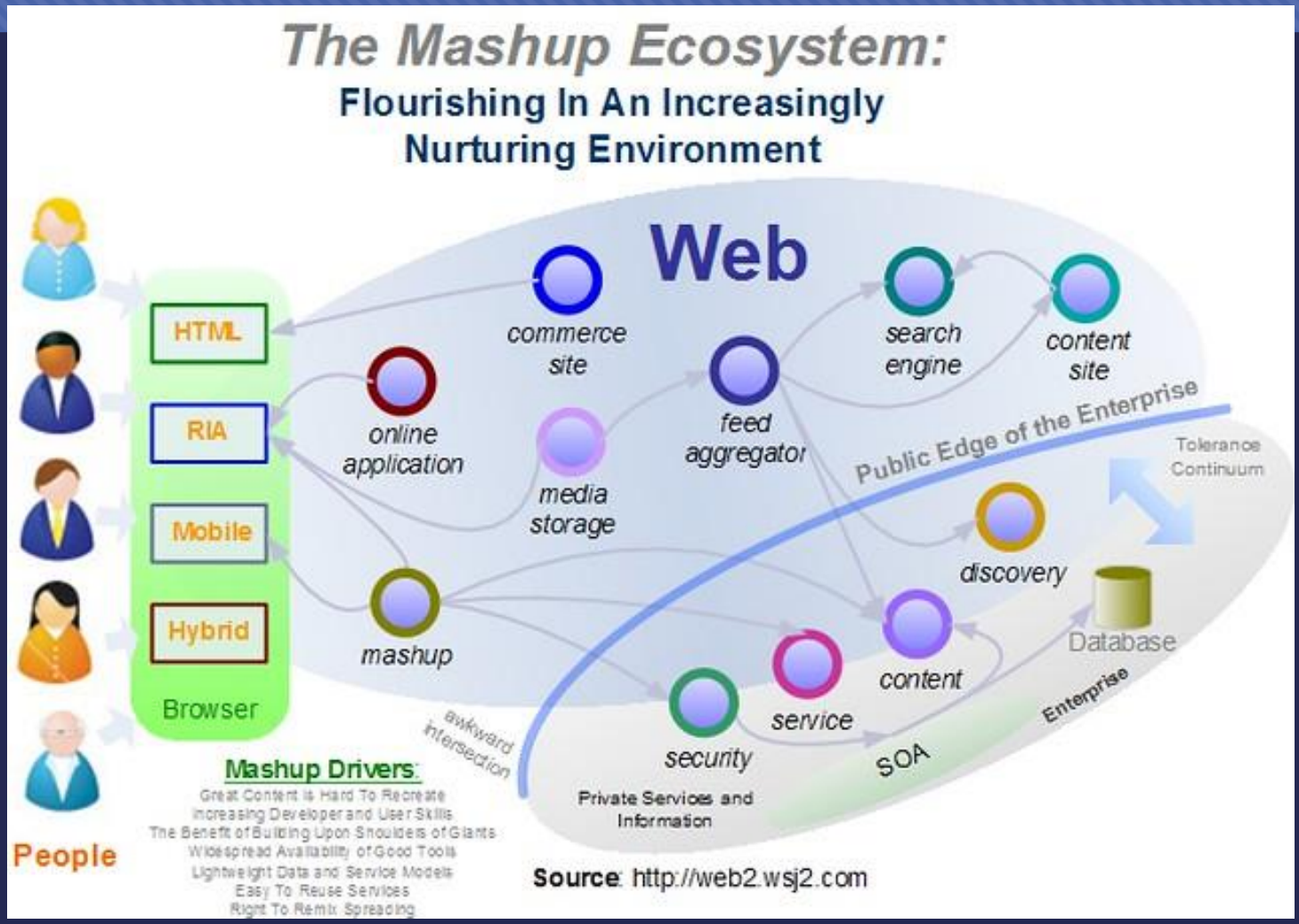
GET versus POST requests



Web Stack

- The “stack” of software needed to run a web server
- Typically: O/S, web server, database, scripting engines, etc
- Very Common: LAMP:
  - Linux
  - Apache
  - MySQL DB
  - PHP

Web 2.0 and Beyond



Cookies

- HTTP is ***STATELESS***
- A webserver doesn't "connect" requests
- To simulate a "session", use cookies
- Put "cookie: <session id>" in request/response header

Basic Idea

GET \ HTTP/1.1
Cookie: ac39f210ef120

Page 1

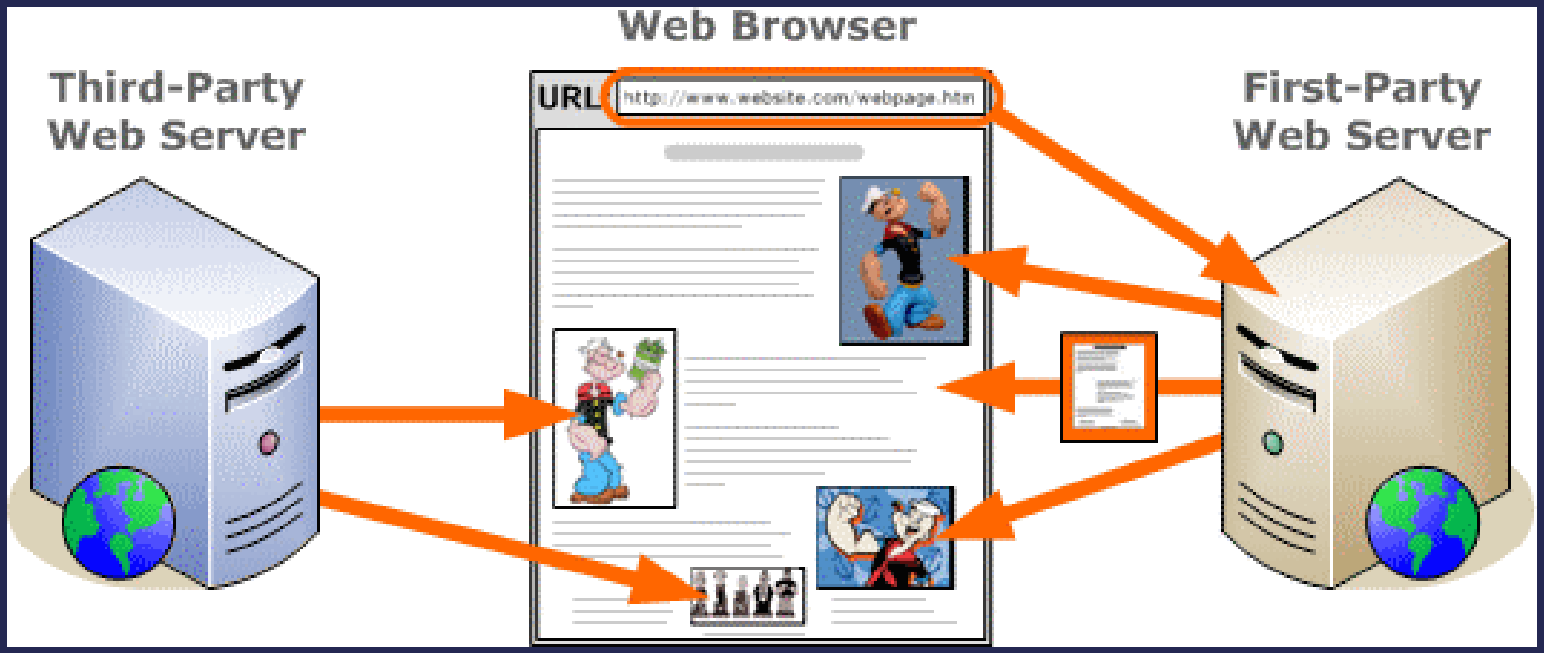
Page 2

GET \ HTTP/1.1
Cookie: 9b8dde1783ff3e

Cookies and Domains

- Cookies are most assigned by domain
- For example, “google.com” cookies
- This is important for security and privacy

First-Party, Third Party



How do companies track?

- First-party façade: advertising_company.amazon.com
- Collusion: first-party, third-party share data
  - First-party can send data to third-party in URL
  - <IMG src="http://third-party?cookie=stolen">