Lecture 16
World Wide Web

The World Wide Web

- what it is
- a brief history
- how it works
- cookies, Javascript, other tracking
- how advertising works
- technical issues
- political / legal / social / economic / jurisdictional issues

(World Wide) Web

- a way to connect computers that provide information (servers) with computers that ask for it (clients like you and me)
  - uses the Internet, but it's not the same as the Internet
- URL (uniform resource locator, e.g., http://www.amazon.com)
  - a way to specify what information to find, and where
- HTTP (hypertext transfer protocol)
  - a way to request specific information from a server and get it back
- HTML (hyptertext markup language)
  - a language for describing information for display
- browser (Firefox, Safari, Internet Explorer, Opera, Chrome, …)
  - a program for making requests, and displaying results
- embellishments
  - pictures, sounds, movies, ...
  - loadable software
- the set of everything this provides

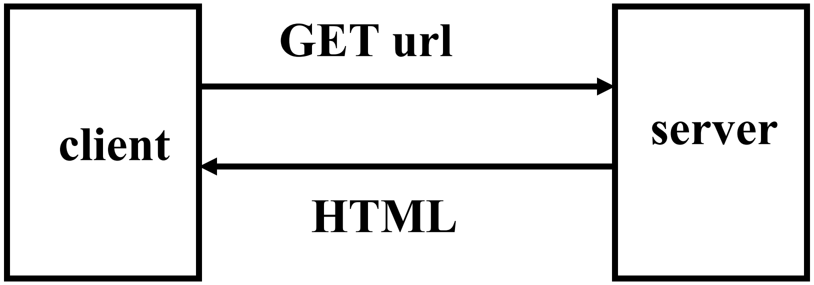
Web history

- 1989: Tim Berners-Lee at CERN
  - a way to make physics literature and research results accessible on the Internet
- 1991: first software distributions
- Feb 1993: Mosaic browser
  - Marc Andreessen at NCSA (Univ of Illinois)
- Mar 1994: Netscape
  - first commercial browser
- technical evolution managed by World Wide Web Consortium
  - non-profit organization at MIT, Berners-Lee is director
  - official definition of HTML and other web specifications
  - see **`www.w3.org`**

HTTP: Hypertext transfer protocol

- What happens when you click on a URL?
- client opens TCP/IP connection to host, sends request

```
GET  /filename  HTTP/1.0
```



- server returns
  - header info
  - HTML
- since server returns the text, it can be created as needed
  - can contain encoded material of many different types (MIME)
- URL format

```
service://hostname/filename?other_stuff
```

- *filename?other_stuff* part can encode
  - data values from client (forms)
  - request to run a program on server (cgi-bin)
  - anything else

Embellishments

- original design of HTTP just returns text to be displayed
- now includes pictures, sound, video, ...
  - need helpers or plug-ins to display non-text content

    e.g., GIF, JPEG graphics; sound; movies

- forms filled in by user
  - need a program on the server to interpret the information (cgi-bin)
- cookies to remember information on client
  - HTTP is stateless: server doesn't saveanything from one request to next
  - cookies are a way to remember information at the client
- active content: download code to run on the client
  - Javascript
  - plug-ins

Forms and CGI programs

- "common gateway interface"
  - standard way to request the server to run a program
  - using information provided by the client via a form

- if the target file on server is an executable program
- and it has the right properties and permissions
  - e.g., in /cgi-bin directory and executable
- then run it on server to produce HTML to send back to client
  - using the contents of the form as input
  - output depends on client request: created on the fly, not just a file

- CGI programs can be written in any programming language
  - Perl, Python, PHP, Java, Ruby, …

Cookies

- HTTP is stateless: it doesn't remember from one request to the next
- cookies are intended to deal with stateless nature of HTTP
  - remember preferences, manage "shopping cart", etc.
- cookie: one chunk of text sent by server to be stored on client
  - stored in browser while it is running (transient)
  - stored in client file system when browser terminates (persistent)
- when client reconnects to same domain,
  browser sends the cookie back to the server
  - sent back verbatim; nothing added
  - sent back only to the same domain that sent it originally
  - contains no information that didn't originate with the server

- in principle, pretty benign
- but heavily used to monitor browsing habits, for commercial purposes

Cookie crumbs

- fetch a page from xyz.com
  - it contains <img src=http://doubleclick.com/advt.gif>
  - this causes a page to be fetched from DoubleClick.com
  - which now knows your IP address and what page you were looking at
- DoubleClick sends back (or arranges for) a suitable advertisement
  - with a cookie that identifies "you" at DoubleClick
- next time you fetch any page that contains a DoubleClick.com image
  - the most recent DoubleClick cookie is sent back to DoubleClick
  - DoubleClick now knows even more about you
  - the set of sites and images that you are viewing is used to
    - update the record of where you have been and what you have looked at
    - send further targeted advertising
    - allow more cookies from advertisers

Advertising

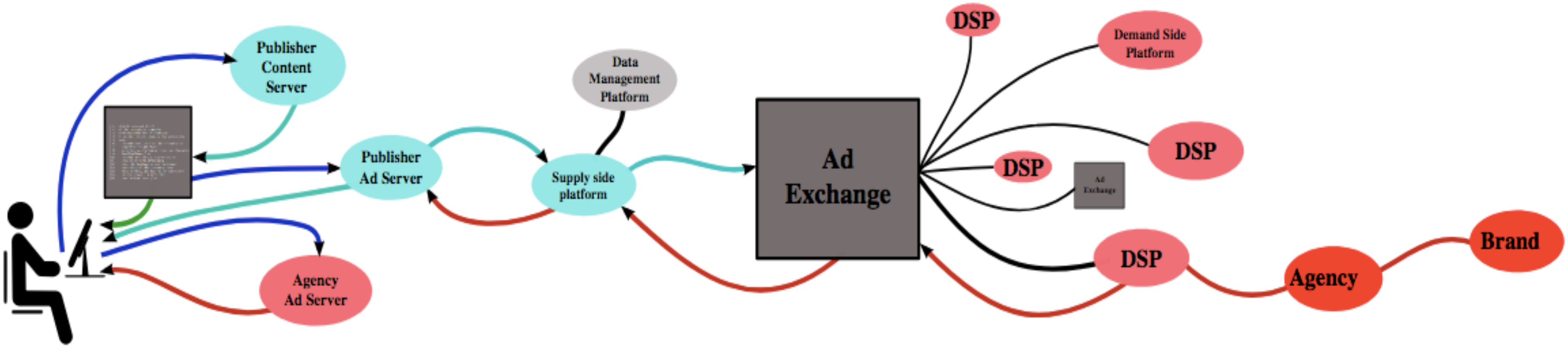
- advertising exchanges
  - Doubleclick Ad Exchange, ...
- a person uses a browser to request a web page
- the web page "publisher" notifies exchanges that advertising space on that page is available
  - publishers are typically portals or entertainment and news sites
  - publisher provides information about the person: past online activity, viewing and shopping habits, geographic location, demographics
    - (probably) not actual identity
- advertisers bid on the ad space
  - amount depends on person's attributes and location, advertiser's budget, etc.
- winner's advertisement is inserted into the page
- elapsed time: 10-100 milliseconds

- this happens for multiple advertisements on one page

Who's involved?

- publisher: integrates advertisements into its online content
- advertiser: provides the advertisements to be displayed on the publisher's content
- advertising agencies: generate and place the ad copy
- ad server: delivers the ad and tracks statistics

- geo-targeting
- behavioral targeting



Cookies are not the only tracking mechanism

- JavaScript
  - potentially continuous reporting of activity on page

- web bugs, web beacons, single-pixel gifs
  - tiny images that report the use of a particular page
  - these can be used in mail messages, not just browsers

- "super cookies"
  - e.g., Verizon's X-UIDH HTTP header on cellphones
  - cookie respawning

- HTML canvas fingerprinting
  - uses subtle differences in browser behavior to distinguish users

- defenses: addons like

  AdBlock, uMatrix, Ghostery, Privacy Badger, NoScript, ...

Javascript, plug-ins, add-ons, extensions, etc.

- **scripting languages interpret downloaded programs**
  - Javascript

    compiled into instructions for a virtual machine

    (like the Toy machine on steroids)

    instructions are interpreted by virtual machine in browser

- **programs that extend capabilities of browser (and other programs)**
  - browser provides an API and a protocol for data exchange
  - extension focuses on specific application area

    e.g., documents, pictures, sound, movies, scripting language, ...
  - may exist standalone as well as in plug-in form
  - e.g., Acrobat Reader, Flash, Quicktime, Windows Media Player, ...

Javascript tracking

- most web pages include some Javascript
- some is used for interactive features, validation, etc.
- much is used for tracking:

  "Google Analytics offers a great breadth of functionality - you can use it to track visitor flow through your site, to view the source of referrals to your site, and to see how well visitors make it through a conversion process such as purchasing an item or signing up for a newsletter."

```
<script>
function move(event) {
  document.getElementById("body").innerHTML =
    "position: " + event.clientX + " " + event.clientY;
}
</script>
<body>
  <div id="body" style="width:100%; height: 500px;"
    onmousemove="move(event)">
  </div>
</body>
```

Targeted advertising at Target

Whenever possible, Target assigns each shopper a unique code — known internally as the Guest ID number — that keeps tabs on everything they buy. “If you use a credit card or a coupon, or fill out a survey, or mail in a refund, or call the customer help line, or open an e-mail we’ve sent you or visit our Web site, we’ll record it and link it to your Guest ID … We want to know everything we can.”

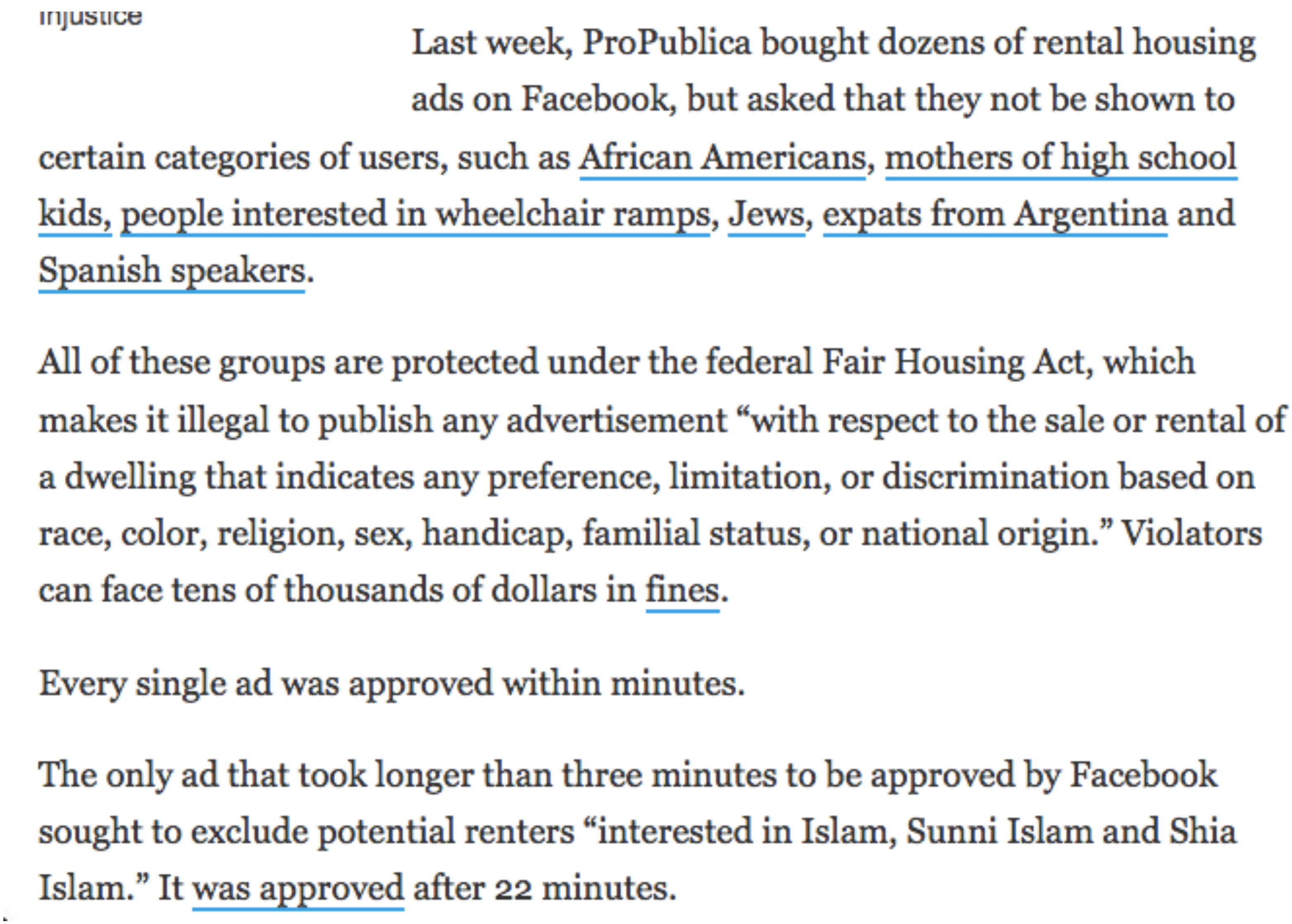
Also linked to your Guest ID is demographic information like your age, whether you are married and have kids, which part of town you live in, how long it takes you to drive to the store, your estimated salary, whether you’ve moved recently, what credit cards you carry in your wallet and what Web sites you visit. Target can buy data about your ethnicity, job history, the magazines you read, if you’ve ever declared bankruptcy or got divorced, the year you bought (or lost) your house, where you went to college, what kinds of topics you talk about online, whether you prefer certain brands of coffee, paper towels, cereal or applesauce, your political leanings, reading habits, charitable giving and the number of cars you own.

What does Facebook know about you?

- https://www.cnbc.com/2018/03/27/facebook-knows-a-lot-about-me.html

- It can recognize my face
- It knows every ad topic I’ve ever clicked
- It has a list of every company that has my contact information from the ads I’ve clicked
- It has a list of every contact in my phone book
- It knows every social event I was invited to and/or attended through Facebook
- It has a log of every friend I have on Facebook and when we became friends
- It knows every time I logged in
- It has a copy of my timeline going back to the time I joined
- It knows my major life events
- It knows every video I’ve watched on Facebook
- It knows exactly where I was
- It has old messages
- It has a copy of every photo I’ve ever uploaded

Pro Publica / Facebook (11/21/17)



injustice

Last week, ProPublica bought dozens of rental housing ads on Facebook, but asked that they not be shown to certain categories of users, such as African Americans, mothers of high school kids, people interested in wheelchair ramps, Jews, expats from Argentina and Spanish speakers.

All of these groups are protected under the federal Fair Housing Act, which makes it illegal to publish any advertisement "with respect to the sale or rental of a dwelling that indicates any preference, limitation, or discrimination based on race, color, religion, sex, handicap, familial status, or national origin." Violators can face tens of thousands of dollars in fines.

Every single ad was approved within minutes.

The only ad that took longer than three minutes to be approved by Facebook sought to exclude potential renters "interested in Islam, Sunni Islam and Shia Islam." It was approved after 22 minutes.

https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin