

Week 11 – Web Usage Data

Yasmin AlNoamany

L&S 88-2

University of California, Berkeley

Announcements

- Project Milestone 3: Status Update
 - Due Nov. 13th
- Assignment 2
- Thanksgiving week?

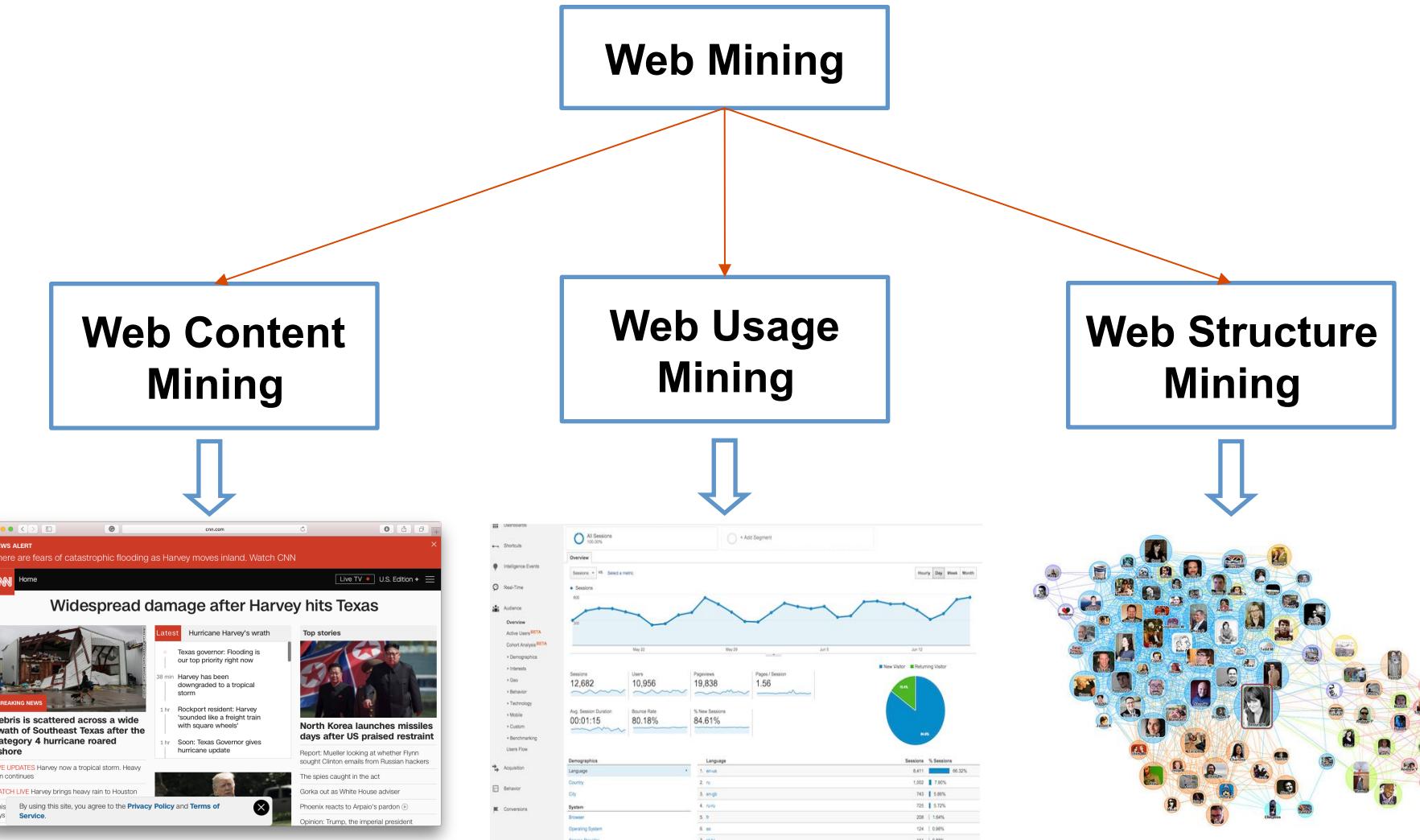
Previous lecture

- Lecture by Alberto Cairo
- Lecture by Ian and Zan from Google
- Guests' lectures evaluation

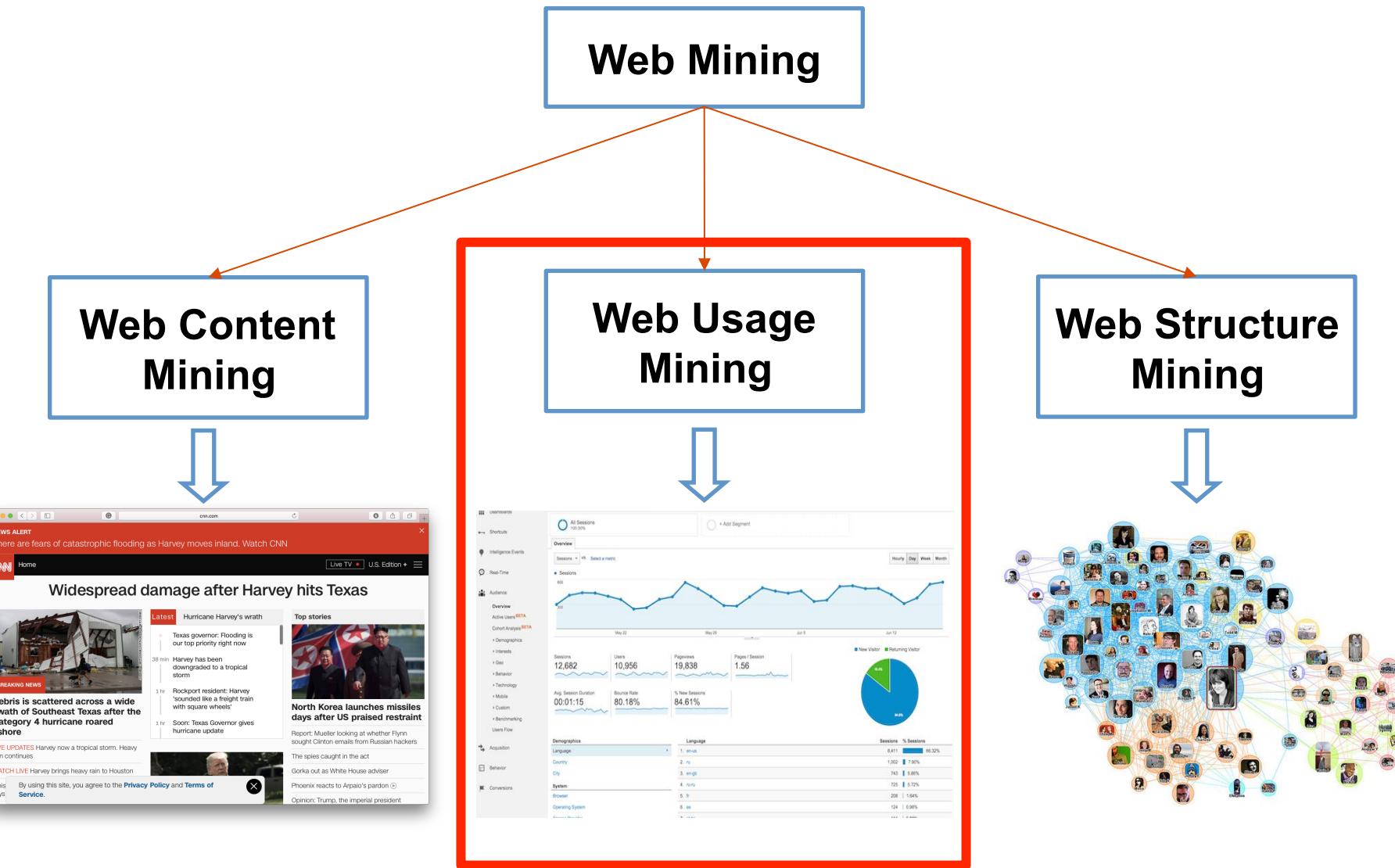
Today's lecture

- Intro to Web usage mining
- Applications of Web usage mining
- Web analytics versus Web usage mining
- Web usage data
- HTTP response codes
- Analyzing usage data

Types of Web Mining



Types of Web Mining



Web usage mining

- Extracting interesting patterns from user interactions with resources on one or more Web sites
- Data
 - clickstream data from Web/application server access logs
 - e-commerce and product-oriented user events (e.g., shopping cart changes, product click-throughs, purchases, etc.)
 - user profiles data, user ratings, user contributed data (tags, comments, reviews)

Applications

- User and customer behavior modeling
- Web site optimization
- E-customer relationship management
- Web marketing
- Targeted advertising
- Recommender systems

Applications

The screenshot shows a product listing for an "AmazonBasics Waterproof Hammock Seat Cover for Pets". The main image displays a brown dog lying comfortably in the hammock seat cover installed in the back seat of a car. Below the main image, there are several smaller thumbnail images showing different angles of the product and its installation.

AmazonBasics Waterproof Hammock Seat Cover for Pets

4.5 stars, 878 customer reviews | 68 answered questions

Price: \$19.99 **prime**
FREE Shipping on orders over \$25—or get FREE Two-Day Shipping with Amazon Prime

In Stock. Want it Monday, Aug. 28? Order within 8 hrs 2 mins and choose Two-Day Shipping at checkout. Details
Ships from and sold by Amazon.com in easy-to-open packaging. Gift-wrap available.

- Waterproof hammock seat cover to keep your car tidy while transporting your pets
- Designed with two pockets to conveniently store grooming supplies, leashes, toys, treats and more
- Protects the back seat from scratches, dirt, dander, and spills
- Secures in back seat using four headrest loops and two seat anchors
- Measures approximately 55 x 59 inches (WxL)
- Made of 100% polyester

Compare with similar items
Used & new (14) from \$12.99 & FREE shipping on orders over \$25.00. [Details](#)

This item's packaging will indicate what is inside. To

Share [Email](#) [Facebook](#)

Buy new:

Qty: 1

Yes, I want F
Shipping with Amazon Prime

Add to Cart

Turn on 1-Click
Buy One-Click
Buy One-Click

Ship to:
Select a shipping method

Buy used:

Add to List

Other Seller Options

Used & new (14) from \$12.99 & FREE shipping on orders over \$25.00. [Details](#)

Applications

AmazonBasics Waterproof Hammock Seat Cover for Pets
4.5 stars, 878 reviews | 68 answered questions

Price: \$19.99 prime
FREE Shipping on orders over \$25—or get FREE Two-Day Shipping with Amazon Prime

In Stock. Want it Monday, Aug. 28? Order within 8 hrs 2 mins and choose Two-Day Shipping at checkout. Details Ships from and sold by Amazon.com in easy-to-open packaging. Gift-wrap available.

- Waterproof hammock seat cover to keep your car tidy while transporting your pets
- Designed with two pockets to conveniently store grooming supplies, leashes, toys, treats and more
- Protects the back seat from scratches, dirt, dander, and spills
- Secures in back seat using four headrest loops and two seat anchors
- Measures approximately 55 x 59 inches (WxL)
- Made of 100% polyester

Compare with similar items
Used & new (14) from \$12.99 & FREE shipping on orders over \$25.00. [Details](#)

This item's packaging will indicate what is inside. To...

Share

Buy new:
Qty: 1
 Yes, I want F...
Shipping with

Turn on 1-Click buying for this item

Ship to: Select a shipping method
 Buy used:

 Other Seller
 Used & new (14) shipping on orders over \$25.00. [Details](#)

Frequently bought together

Total price: \$27.98
 +

i These items are shipped from and sold by different sellers. [Show details](#)

This item: AmazonBasics Waterproof Hammock Seat Cover for Pets \$19.99
 Comsun Collapsible Dog Bowl, Food Grade Silicone BPA Free FDA Approved, Foldable Expandable Cup Dish... \$6.99
 Vastar 2 Packs Adjustable Pet Dog Cat Car Seat Belt Safety Leads Vehicle Seatbelt Harness, Made from... \$7.99

Sponsored products related to this item (What's this?)

Dog Car Seat Covers | Waterproof Cargo Liner Cover For Suv | Car Bench Seat Cover for...
4.5 stars, 12 reviews | \$15.95 prime

Dog Seat Cover - Pet Car Seat Covers - Waterproof Dog Hammock for Cars and SUVs - with Two...
4.5 stars, 33 reviews | \$32.99 prime

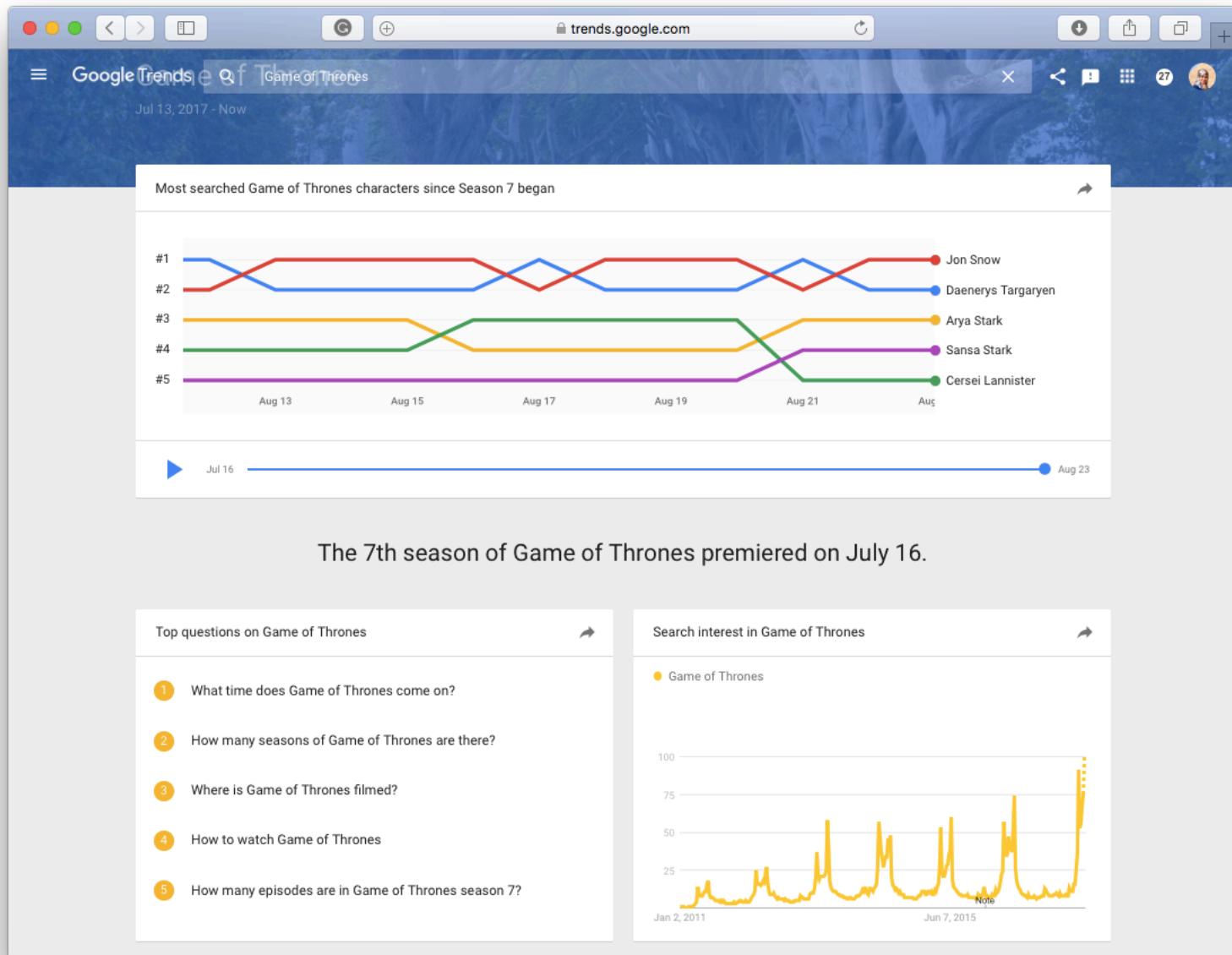
WENFENG Pet Seat Cover, Waterproof & Scratch Proof Dog Car Seat Covers, Hammock...
4.5 stars, 40 reviews | \$20.99 prime

Auelife Pet Seat Cover for Cars, Truck and SUV,Waterproof and Non-Slip Dog Car Seat...
4.5 stars, 21 reviews | \$36.99 prime

Bulldogology Premium Dog Car Seat Covers - Heavy Duty Durable Quality for Cars,...
4.5 stars, 60 reviews | \$64.97 prime

Customers who bought this item also bought

Using the Web to understand the Web



Web analytics versus Web usage mining

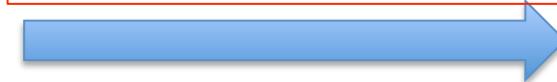
- Web Analytics
 - refers to the measurement, analysis, and reporting of user behavior on the Web
 - usually involves descriptive statistics from clickstream and other user behavior data at different levels of aggregations across predetermined dimensions such as time, content/product categories, referring sites, etc.
 - e.g., Google Analytics
- Web Usage Mining
 - goes beyond basic analytics to
 - discover patterns in usage data
 - identify and characterize important customer segments
 - find affinities across pages or products
 - build models to predict future behavior

How the Web works



request

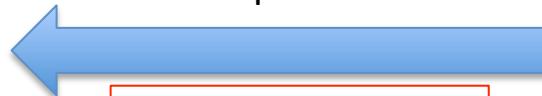
Get http://pbskids.org



Web Server

Response

HTTP/1.1 200 OK



HTTP responses

- 1xx: Informational - Request received, continuing process
 - 100 Continue
- 2xx: Success - The action was successfully received, understood, and accepted
 - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
 - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
 - 404 Page Not found, 403 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
 - 501 Internal Server Error

HTTP responses

Why this is important?

- Multiple errors from assignment 2:

TweepError: Twitter error response: status code = 401

HTTP responses

- 1xx: Informational - Request received, continuing process
 - 100 Continue
- 2xx: Success - The action was successfully received, understood, and accepted
 - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
 - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
 - 404 Page Not found, 403 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
 - 500 Internal Server Error



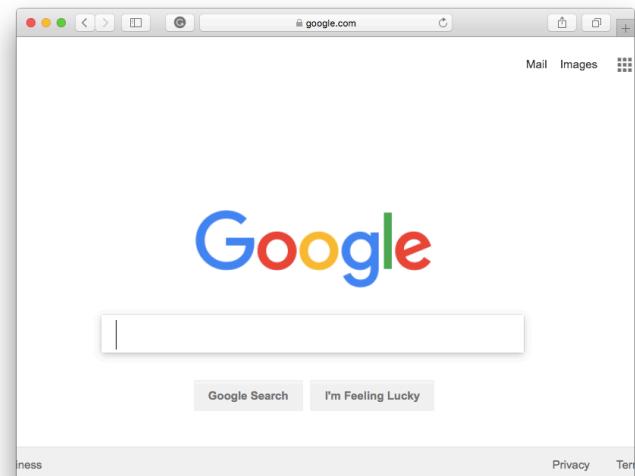
HTTP responses

- 1xx: Informational - Request received, continuing process
 - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
 - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
 - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
 - 404 Page Not found, 403 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
 - 501 Internal Server Error



HTTP responses

- 1xx: Informational - Request received, continuing process
 - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
 - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
 - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
 - 404 Page Not found, 403 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
 - 501 Internal Server Error



HTTP responses

- 1xx: Informational - Request received, continuing process
 - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
 - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
 - 302 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
 - 404 Page Not found, 403 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
 - 501 Internal Server Error



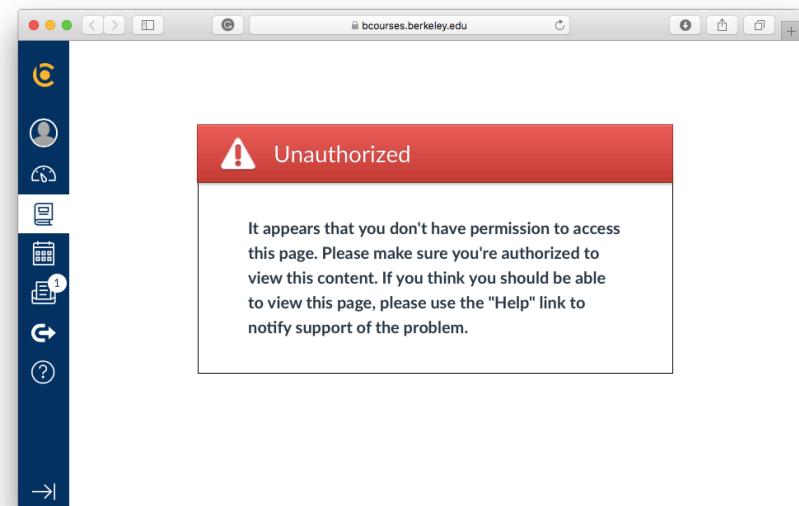
HTTP responses

- 1xx: Informational - Request received, continuing process
 - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
 - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
 - 302 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
 - 404 Not found, 401 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
 - 501 Internal Server Error



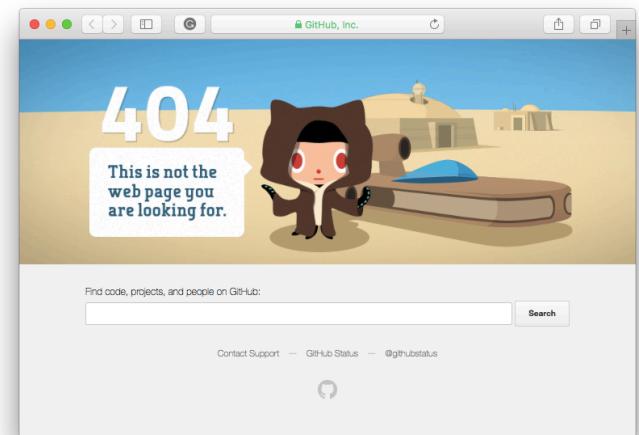
HTTP responses

- 1xx: Informational - Request received, continuing process
 - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
 - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
 - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
 - 404 Not found, 401 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
 - 501 Internal Server Error



HTTP responses

- 1xx: Informational - Request received, continuing process
 - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
 - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
 - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
 - 404 Not found, 401 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
 - 501 Internal Server Error



HTTP responses

- 1xx: Informational - Request received, continuing process
 - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
 - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
 - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
 - 404 Not found, 401 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
 - 501 Internal Server Error



<https://i0.wp.com/s3.amazonaws.com/production-wordpress-assets/blog/wp-content/uploads/2016/11/29074529/500-internal-server-error.png?fit=604%2C237&ssl=1>

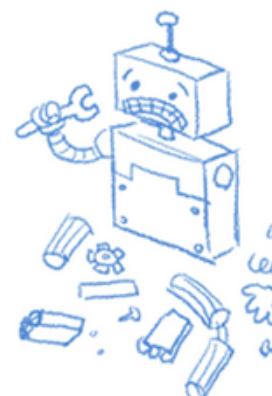
HTTP responses

- 1xx: Informational - Request received, continuing process
 - 180 Ringing
- 2xx: Success - The action was successfully received, understood, and accepted
 - 200 OK
- 3xx: Redirection - Further action must be taken in order to complete the request
 - 303 Moved Temporarily
- 4xx: Client Error - The request contains bad syntax or cannot be fulfilled
 - 404 Not found, 401 Unauthorized
- 5xx: Server Error - The server failed to fulfill an apparently valid request
 - 501 Internal Server Error



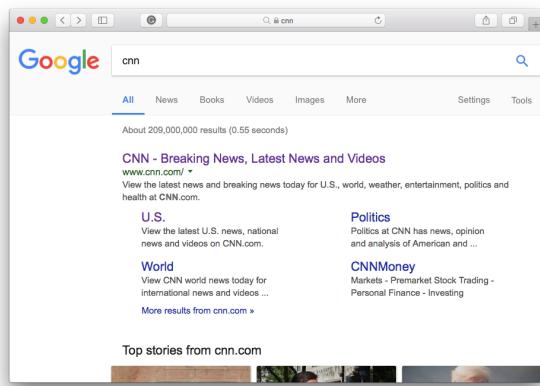
500. That's an error.

There was an error. Please try again later. That's all we know.



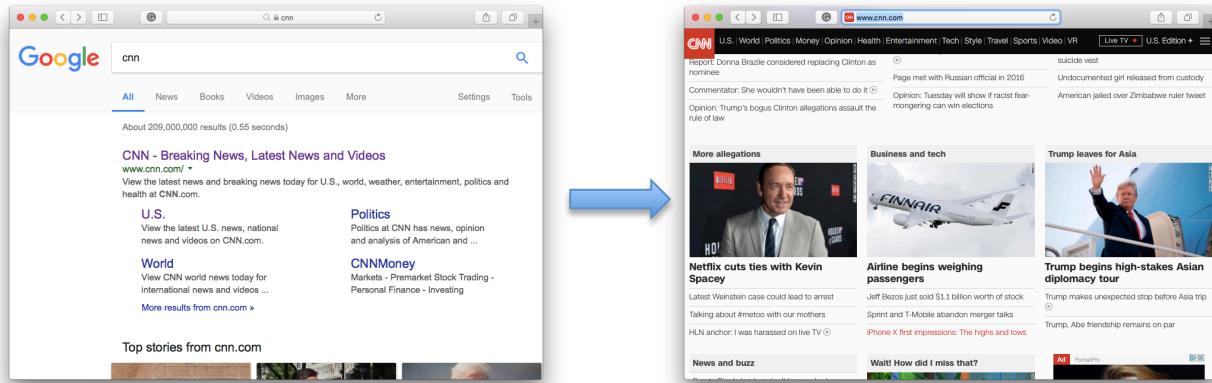
<https://i0.wp.com/s3.amazonaws.com/production-wordpress-assets/blog/wp-content/uploads/2016/11/29074529/500-internal-server-error.png?fit=604%2C237&ssl=1>

What happen when I surf the internet?



```
0.247.222.86 -- [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433
"https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7
(KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

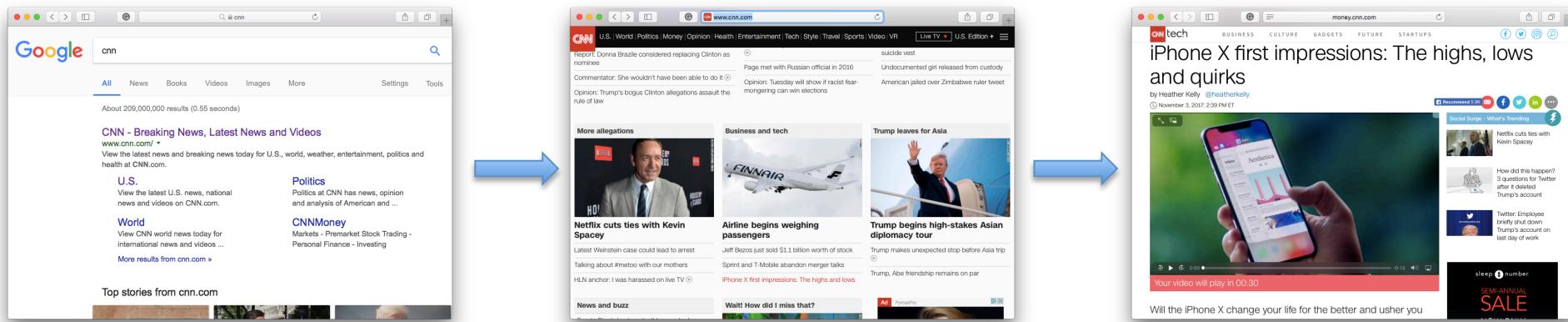
What happen when I surf the internet?



```
0.247.222.86 -- [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 "
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7"
```

```
0.247.222.86 -- [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433
"https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7
(KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

What happen when I surf the internet?



```
0.247.222.86 -- [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 " "
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7"

0.247.222.86 -- [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433
"https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7
(KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"

0.247.222.86 -- [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/
gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 " http://www.cnn.com"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7"
```

What else do you think is recorded
in log files?

Web usage data

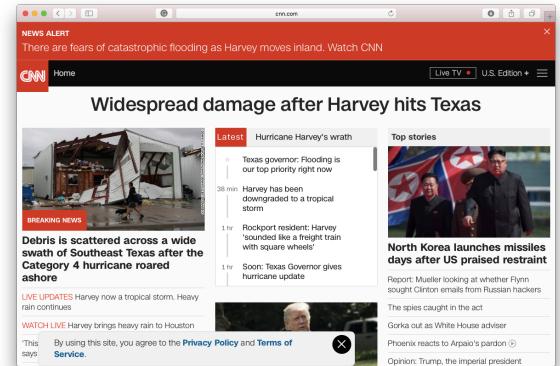
Server log files

- Each time a client requests a resource the server of that resource may record the following in its log files:
 - Client IP
 - Access time
 - HTTP request method
 - URI
 - Protocol
 - HTTP status code
 - Bytes sent
 - Referring URI
 - User-Agent

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86

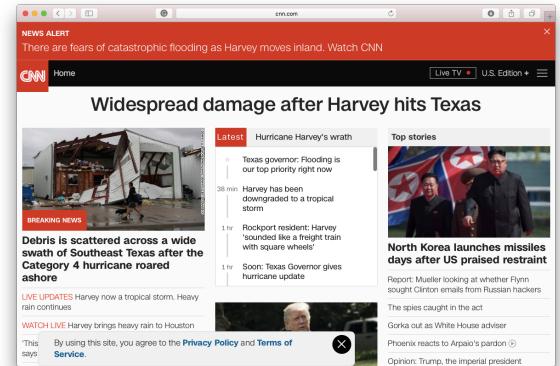


<http://www.cnn.com>

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000

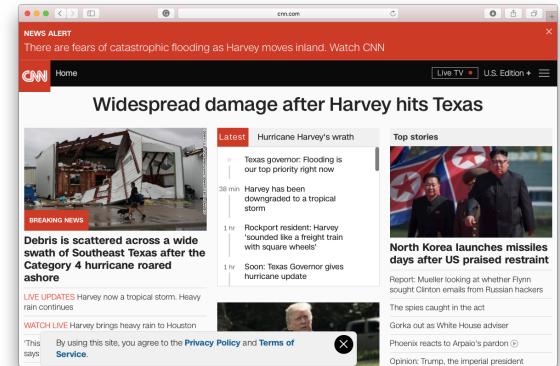


<http://www.cnn.com>

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- **HTTP request method:** GET

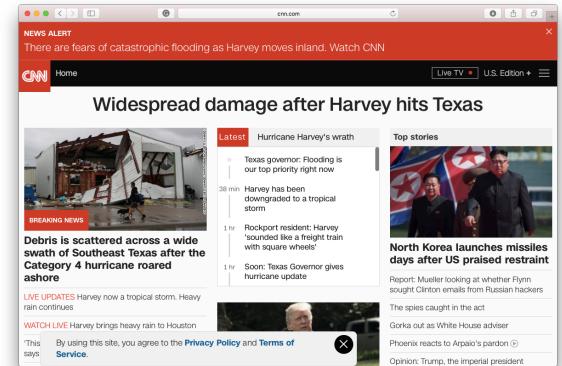


<http://www.cnn.com>

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- **URI:** <http://www.cnn.com>

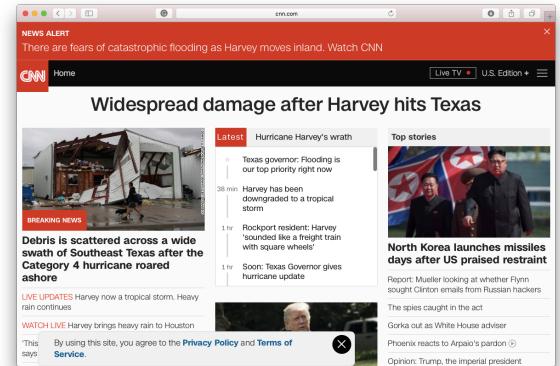


<http://www.cnn.com>

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: http://www.cnn.com
- **Protocol:** HTTP/1.1

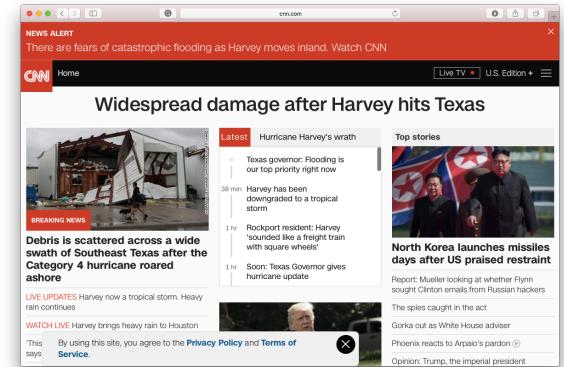


http://www.cnn.com

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: http://www.cnn.com
- Protocol: HTTP/1.1
- **HTTP status code:** 200

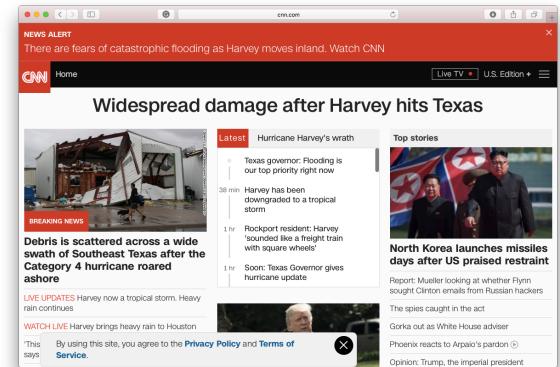


http://www.cnn.com

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: http://www.cnn.com
- Protocol: HTTP/1.1
- HTTP status code: 200
- **Bytes sent:** 96433

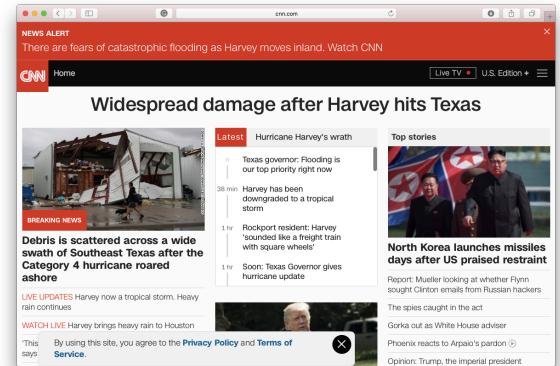


<http://www.cnn.com>

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: <http://www.cnn.com>
- Protocol: HTTP/1.1
- HTTP status code: 200
- Bytes sent: 96433
- Referring URI: <https://www.google.com>

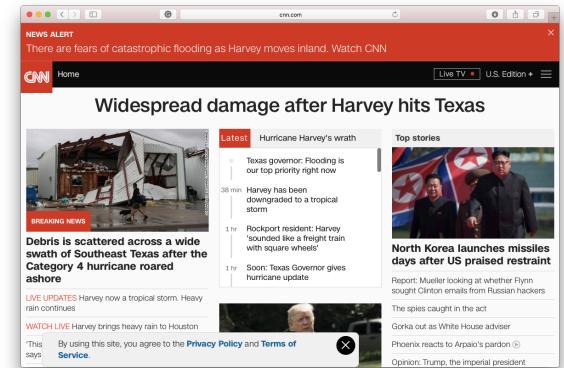


<http://www.cnn.com>

Server log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

- Client IP: 0.247.222.86
- Access time: 0/Sep/2017:07+0000
- HTTP request method: GET
- URI: http://www.cnn.com
- Protocol: HTTP/1.1
- HTTP status code: 200
- Bytes sent: 96433
- Referring URI: https://www.google.com
- **User-Agent:** Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7



http://www.cnn.com

What we get from the log

Web usage data processing

The image displays two screenshots of the Blogger platform's audience statistics interface.

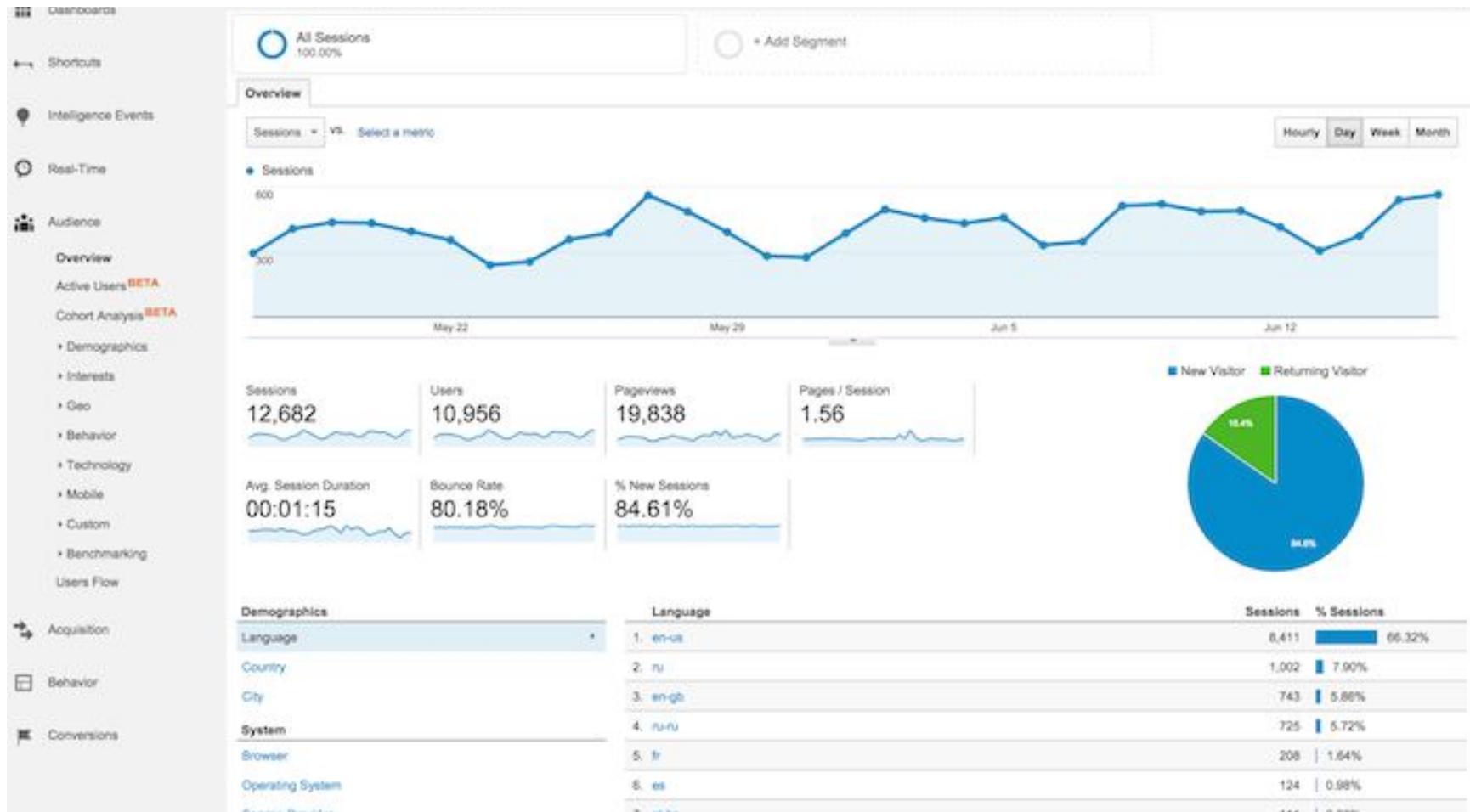
Top Screenshot (Audience Stats):

- Header:** Blogger: ComputerAngel - Audience stats | https://www.blogger.com/blogger.g?blogID=16158115#audiencestats
- Left Sidebar:** ComputerAngel | View blog
Posts, Stats (selected), Overview, Posts, Traffic sources, Audience, Comments
- Right Content Area:**
 - Cookie Consent Message:** European Union laws require you to give European Union visitors information about cookies used on your blog. In many cases, these laws also require you to obtain consent.
As a courtesy, we have added a notice on your blog to explain Google's use of certain Blogger and Google cookies, including use of Google Analytics and AdSense cookies.
 - Pageviews by Countries:** A world map showing pageviews. The United States is highlighted in green.
 - Pageviews by Browsers:** Entry: Chrome, Pageviews: 3 (100%).
 - Pageviews by Operating Systems:** Entry: Macintosh, Pageviews: 3 (100%).

Bottom Screenshot (Overview Stats):

- Header:** Blogger: ComputerAngel - Overview stats | https://www.blogger.com/blogger.g?blogID=16158115#overviewstats
- Left Sidebar:** ComputerAngel | View blog
Posts, Stats (selected), Overview, Posts, Traffic sources, Audience, Comments, Earnings, Campaigns, Pages, Layout, Theme, Settings, Reading List, Help
- Right Content Area:**
 - Pageviews Chart:** A line chart showing pageviews over time from Oct 28, 2017 to Nov 4, 2017. The chart shows a single spike of 1 pageview on October 30, 2017.
 - Traffic Sources:** No stats yet, check back later.
 - Audience:** A world map showing audience distribution. The United States is highlighted in green.

Google analytics



Analyzing Web usage data

- Data preprocessing
- Data cleaning
- User identification
- Session identification
- Pattern detection

Data preprocessing

- Parse log files

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```



Client IP	Access time	HTTP request method	URI	Protocol	HTTP status code	Bytes sent	Referring URI	User-Agent
0.247.222.86	0/Sep/2017:07+0000	GET	http://www.cnn.com	HTTP/1.1	200	96433	https://www.google.com	Mozilla/5.0 (Macintosh; ...

Python library for parsing logs: <https://github.com/rory/apache-log-parser>

Data cleaning

Filter our irrelevant requests

```
0.247,.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"

0.247,.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"

0.247,.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 "http://www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

But I viewed only two cnn
pages = 2 requests

~400 requests

Data cleaning

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308/  
http://www.jcdl.org/ HTTP/1.1" 200 2137 "--"  
"Mozilla/5.0"
```

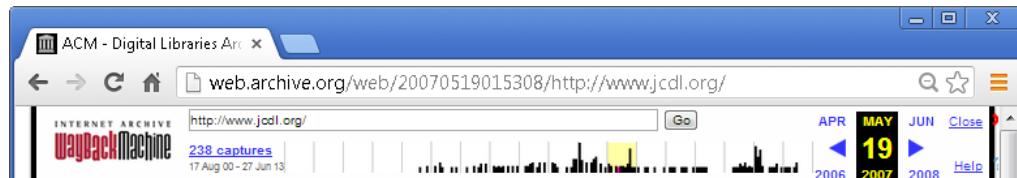
```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308im/  
http://www.jcdl.org/images/jcdl2007-edie.jpg  
HTTP/1.1" 200 2137 "--" "Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://staticweb.archive.org/images/toolbar/  
wayback-toolbar-logo.png HTTP/1.1" 200 3700  
"--" "Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20100102003557/  
about:blank HTTP/1.1" 302 0 "www.xx.com"  
"Mozilla/4.0"
```

```
0.26.129.146 - - [02/Feb/2012:00:01:54] "GET  
http://web.archive.org/web/20140004100000/  
http://www.jcdl.org/ HTTP/1.1" 302 0 "--"  
"Mozilla/5.0"
```

<http://web.archive.org/web/20070519015308/>
<http://www.jcdl.org/>



The screenshot shows a web browser window titled 'ACM - Digital Libraries Ar...'. The address bar contains 'web.archive.org/web/20070519015308/http://www.jcdl.org/'. The page itself is from the Internet Archive's Wayback Machine, showing a screenshot of the JCDL 2007 website. The main heading is 'Joint Conference on Digital Libraries'. Below it, a banner for 'Upcoming Conference: JCDL 2007 (Vancouver, BC, June 17-23, 2007)' is visible. To the left, there's a sidebar with links like 'Home', 'About JCDL', 'JCDL Sponsors', etc. On the right, there's information about the conference dates and a photo of a speaker at a podium.

This site is an archive of Digital Library conferences and resources. We welcome your comments and suggestions for developing and enriching this site.

JCDL 2007

JCDL 2007 will be held in Vancouver, BC from June 17-23, 2007.

May 21st - Advance registration ends
May 17th - Last day to obtain conference rate at the Westin Bayshore

Embedded resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308/  
http://www.jcdl.org/ HTTP/1.1" 200 2137 "--"  
"Mozilla/5.0"
```

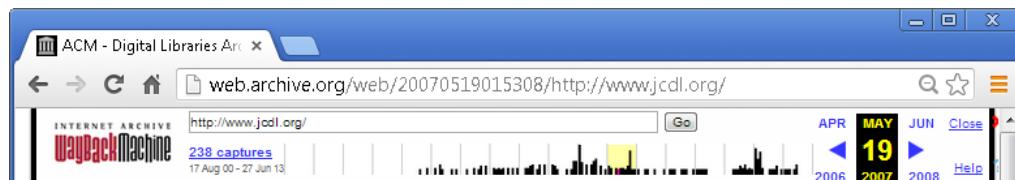
```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308im\_  
http://www.jcdl.org/images/jcdl2007-edie.jpg  
HTTP/1.1" 200 2137 "--" "Mozilla/5.0"
```

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://staticweb.archive.org/images/toolbar/  
wayback-toolbar-logo.png HTTP/1.1" 200 3700  
"--" "Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20100102003557/  
about:blank HTTP/1.1" 302 0 "www.xx.com"  
"Mozilla/4.0"
```

```
0.26.129.146 - - [02/Feb/2012:00:01:54] "GET  
http://web.archive.org/web/20140004100000/  
http://www.jcdl.org/ HTTP/1.1" 302 0 "--"  
"Mozilla/5.0"
```

<http://web.archive.org/web/20070519015308/>
<http://www.jcdl.org/>



The screenshot shows a web browser window titled 'ACM - Digital Libraries Ar...'. The address bar contains 'web.archive.org/web/20070519015308/http://www.jcdl.org/'. The page itself is from the Internet Archive's Wayback Machine, dated 17 Aug 00 - 27 Jun 13. It features a banner for the 'Joint Conference on Digital Libraries' and information about the 'JCDL 2007' conference in Vancouver, BC, June 17-23, 2007. A sidebar on the left lists links such as Home, About JCDL, JCDL Sponsors, JCDL Steering Committee, Upcoming Events and Conferences, Past events and conferences, and Awards. A red box highlights a thumbnail image of a presentation slide for 'JCDL 2007: Vancouver'.

Joint Conference on Digital Libraries

Upcoming Conference: [JCDL 2007](#) (Vancouver, BC, June 17-23, 2007).

This site is an archive of Digital Library conferences and resources. We welcome your comments and suggestions for developing and enriching this site.

JCDL 2007

 JCDL 2007 will be held in Vancouver, BC from June 17-23, 2007.

May 21st - Advance registration ends
May 17th - Last day to obtain conference rate at the Westin Bayshore

Embedded resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308/  
http://www.jcdl.org/ HTTP/1.1" 200 2137 "--"  
"Mozilla/5.0"
```

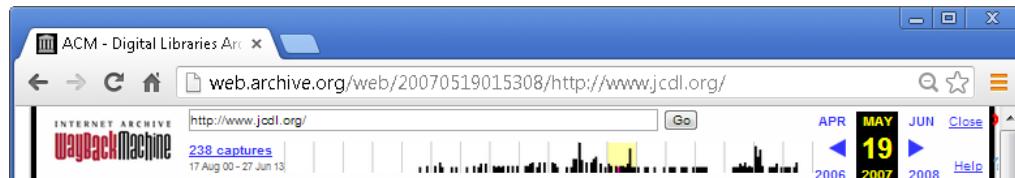
0.11.160.135 [02/Feb/2012:00:01:03] "GET
http://web.archive.org/web/20070519015308im_
<http://www.jcdl.org/images/jcdl2007-edie.jpg>
HTTP/1.1" 200 2137 "--" "Mozilla/5.0"

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://staticweb.archive.org/images/toolbar/  
wayback-toolbar-logo.png HTTP/1.1" 200 3700  
"--" "Mozilla/5.0"
```

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20100102003557/  
about:blank HTTP/1.1" 302 0 "www.xx.com"  
"Mozilla/4.0"
```

```
0.26.129.146 -- [02/Feb/2012:00:01:54] "GET  
http://web.archive.org/web/20140004100000/  
http://www.jcdl.org/ HTTP/1.1" 302 0 "--"  
"Mozilla/5.0"
```

<http://web.archive.org/web/20070519015308/>
<http://www.jcdl.org/>



The screenshot shows a web browser window titled "ACM - Digital Libraries Archiving". The address bar contains the URL <http://web.archive.org/web/20070519015308/http://www.jcdl.org/>. The page content is from a Wayback Machine capture dated 17 Aug 00 - 27 Jun 13. The main heading is "Joint Conference on Digital Libraries". Below it, a section for the "Upcoming Conference: [JCDL 2007](#) (Vancouver, BC, June 17-23, 2007)." A sidebar on the left lists links such as Home, About JCDL, JCDL Sponsors, JCDL Steering Committee, Upcoming Events and Conferences, Past events and conferences, and Awards. A large image in the center shows a person speaking at a podium with a screen behind them displaying the conference logo. To the right of the image, text indicates the conference will be held in Vancouver, BC from June 17-23, 2007, and provides registration information.

This site is an archive of Digital Library conferences and resources. We welcome your comments and suggestions for developing and enriching this site.

JCDL 2007

JCDL 2007 will be held in Vancouver, BC from June 17-23, 2007.

May 21st - Advance registration ends
May 17th - Last day to obtain conference rate at the Westin Bayshore

Static resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308/  
http://www.jcdl.org/ HTTP/1.1" 200 2137 "--  
"Mozilla/5.0"
```

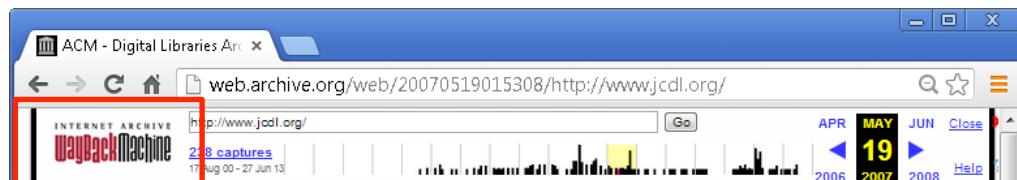
0.11.160.135 [02/Feb/2012:00:01:03] "GET
<http://web.archive.org/web/20070519015308im/>
<http://www.jcdl.org/images/jcdl2007-edie.jpg>
HTTP/1.1" 200 2137 "-- Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
<http://staticweb.archive.org/images/toolbar/>
<wayback-toolbar-logo.png> HTTP/1.1" 200 3700
"-- Mozilla/5.0"

0.151.147.108 [02/Feb/2012:00:01:03] "GET
<http://web.archive.org/web/20100102003557/>
<about:blank> HTTP/1.1" 302 0 "www.xx.com"
"Mozilla/4.0"

0.26.129.146 -- [02/Feb/2012:00:01:54] "GET
<http://web.archive.org/web/20140004100000/>
<http://www.jcdl.org/> HTTP/1.1" 302 0 "--
"Mozilla/5.0"

<http://web.archive.org/web/20070519015308/>
<http://www.jcdl.org/>



The screenshot shows a web browser window titled "ACM - Digital Libraries Archiving". The address bar contains the URL <http://web.archive.org/web/20070519015308/http://www.jcdl.org/>. The page content is from a Wayback Machine capture dated May 19, 2007. The main heading is "Joint Conference on Digital Libraries". Below it, a message says "Upcoming Conference: [JCDL 2007](#) (Vancouver, BC, June 17-23, 2007)". A sidebar on the left lists links such as Home, About JCDL, JCDL Sponsors, JCDL Steering Committee, Upcoming Events and Conferences, Past events and conferences, and Awards. A large red box highlights the Wayback Machine logo and the capture date. Another red box highlights the "Upcoming Conference" section.

This site is an archive of Digital Library conferences and resources. We welcome your comments and suggestions for developing and enriching this site.

JCDL 2007



JCDL 2007 will be held in Vancouver, BC from June 17-23, 2007.

May 17th - Advance registration ends
May 17th - Last day to obtain conference rate at the Westin Bayshore

Static resources

```
0.11.160.135 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20070519015308/  
http://www.jcdl.org/ HTTP/1.1" 200 2137 "--  
"Mozilla/5.0"
```

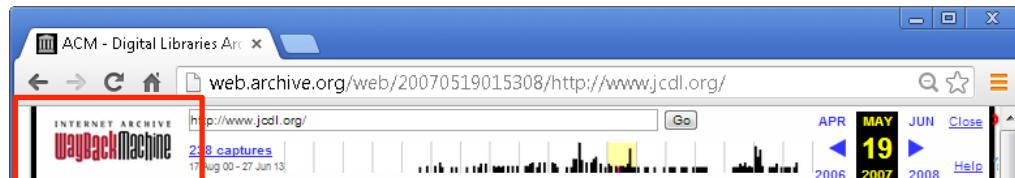
0.11.160.135 [02/Feb/2012:00:01:03] "GET
<http://web.archive.org/web/20070519015308im/>
<http://www.jcdl.org/images/jcdl2007-edie.jpg>
HTTP/1.1" 200 2137 "--" "Mozilla/5.0"

0.11.160.135 [02/Feb/2012:00:01:03] "GET
<http://staticweb.archive.org/images/toolbar/>
<http://wayback-toolbar-1> HTTP/1.1" 200 3700
"--" "Mozilla/5.0"

```
0.151.147.108 [02/Feb/2012:00:01:03] "GET  
http://web.archive.org/web/20100102003557/  
about:blank HTTP/1.1" 302 0 "www.xx.com"  
"Mozilla/4.0"
```

```
0.26.129.146 -- [02/Feb/2012:00:01:54] "GET  
http://web.archive.org/web/20140004100000/  
http://www.jcdl.org/ HTTP/1.1" 302 0 "--  
"Mozilla/5.0"
```

<http://web.archive.org/web/20070519015308/>
<http://www.jcdl.org/>



The screenshot shows a web browser window titled "ACM - Digital Libraries Ar". The address bar contains the URL <http://web.archive.org/web/20070519015308/http://www.jcdl.org/>. The page content is from the Wayback Machine, specifically a capture from May 19, 2007. The main heading is "Joint Conference on Digital Libraries". Below it, a message says "Upcoming Conference: [JCDL 2007](#) (Vancouver, BC, June 17-23, 2007)". On the left, there's a sidebar with links like "Home", "About JCDL", "JCDL Sponsors", etc. On the right, there's information about the conference dates and a photo of a speaker at a podium.

This site is an archive of Digital Library conferences and resources. We welcome your comments and suggestions for developing and enriching this site.

JCDL 2007

JCDL 2007 will be held in Vancouver, BC from June 17-23, 2007.

May 17th - Advance registration ends
May 17th - Last day to obtain conference rate at the Westin Bayshore

User identification

- Grouping: based on the IP and User-Agent

```
0.247.222.86 -- [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 " Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"

0.247.222.86 -- [0/Sep/2017:07:05:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433 "https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"

0.247.222.86 -- [0/Sep/2017:07:12:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 "http://www.cnn.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

Same IP

Session identification

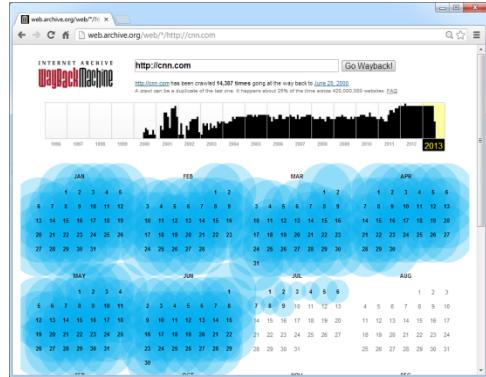
- Segment the user activity record of each user into sessions, each representing a single visit to the site
- Threshold timeout: 10 minutes Liu et al. 2007, Spiliopoulou et al. 2003

```
0.247.222.86 -- [0/Sep/2017:07:03:46 +0000] "GET https://www.google.com HTTP/1.1" 200 96433 "
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7"

0.247.222.86 -- [0/Sep/2017:07:05:46 +0000] "GET http://www.cnn.com HTTP/1.1" 200 96433
"https://www.google.com" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7
(KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"

0.247.222.86 -- [0/Sep/2017:07:12:46 +0000] "GET http://money.cnn.com/2017/10/31/technology/
gadgets/iphone-x-first-impressions/index.html HTTP/1.1" 200 96433 " http://www.cnn.com"
"Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8) AppleWebKit/535.7 (KHTML, like Gecko) Chrome/
16.0.912.77 Safari/535.7"
```

Session: set of Web pages requested by a particular user in a single visit



p1



1 mins



4 mins

p3



3 mins

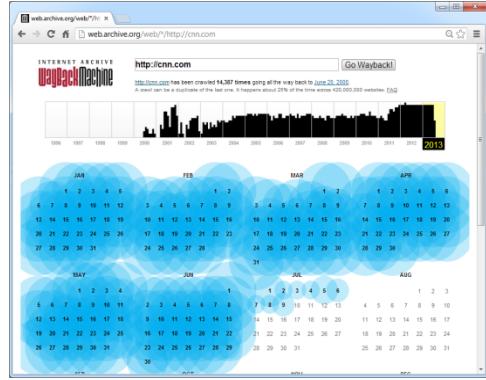
p4



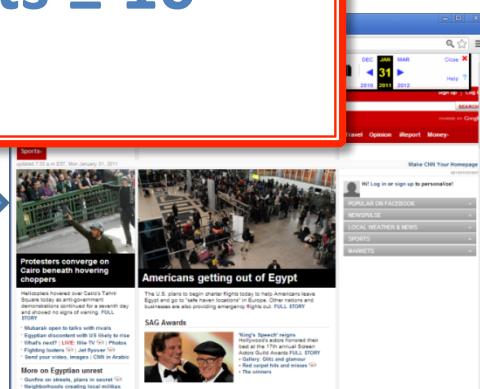
9 mins

p5

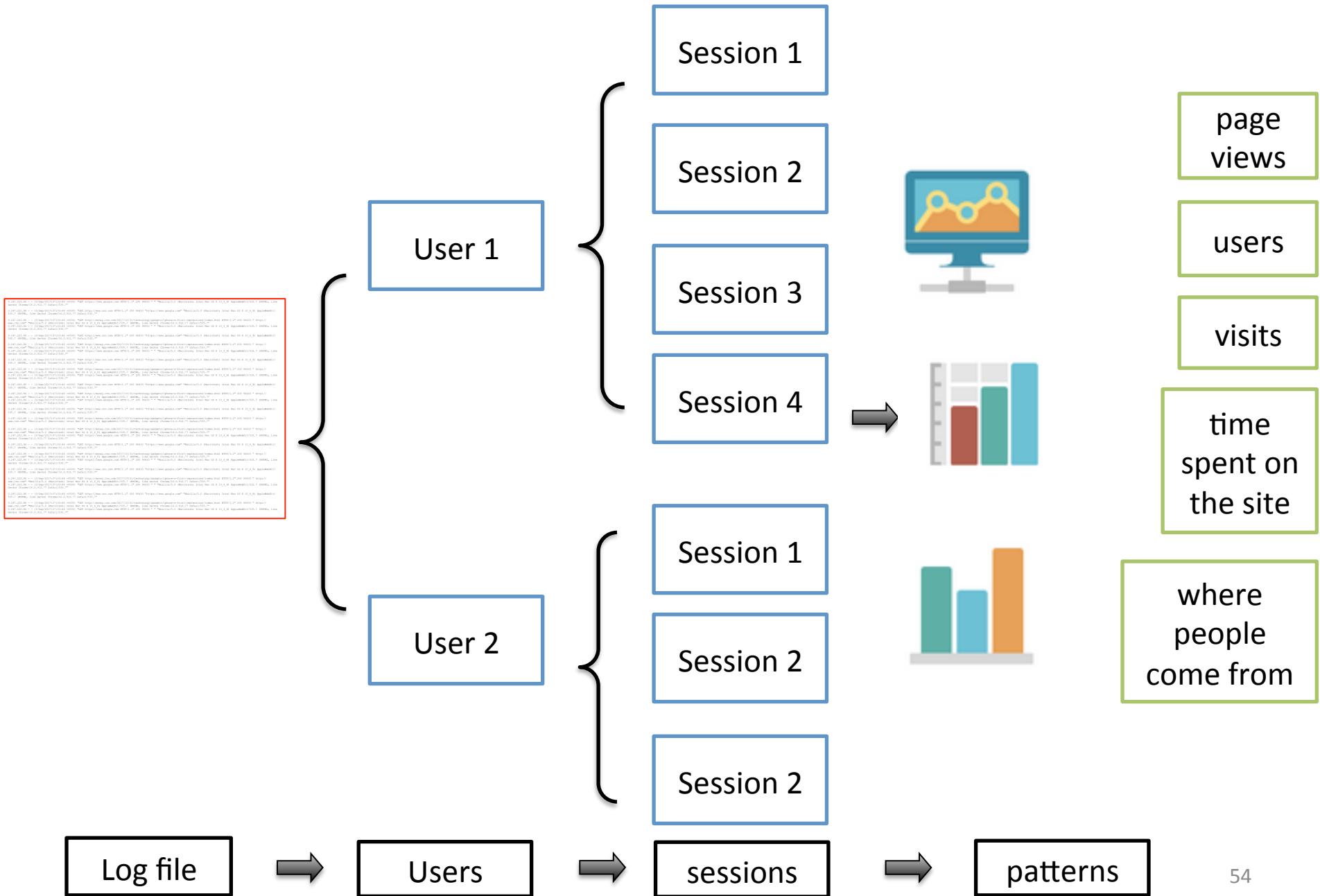
Session: set of Web pages requested by a particular user in a single visit



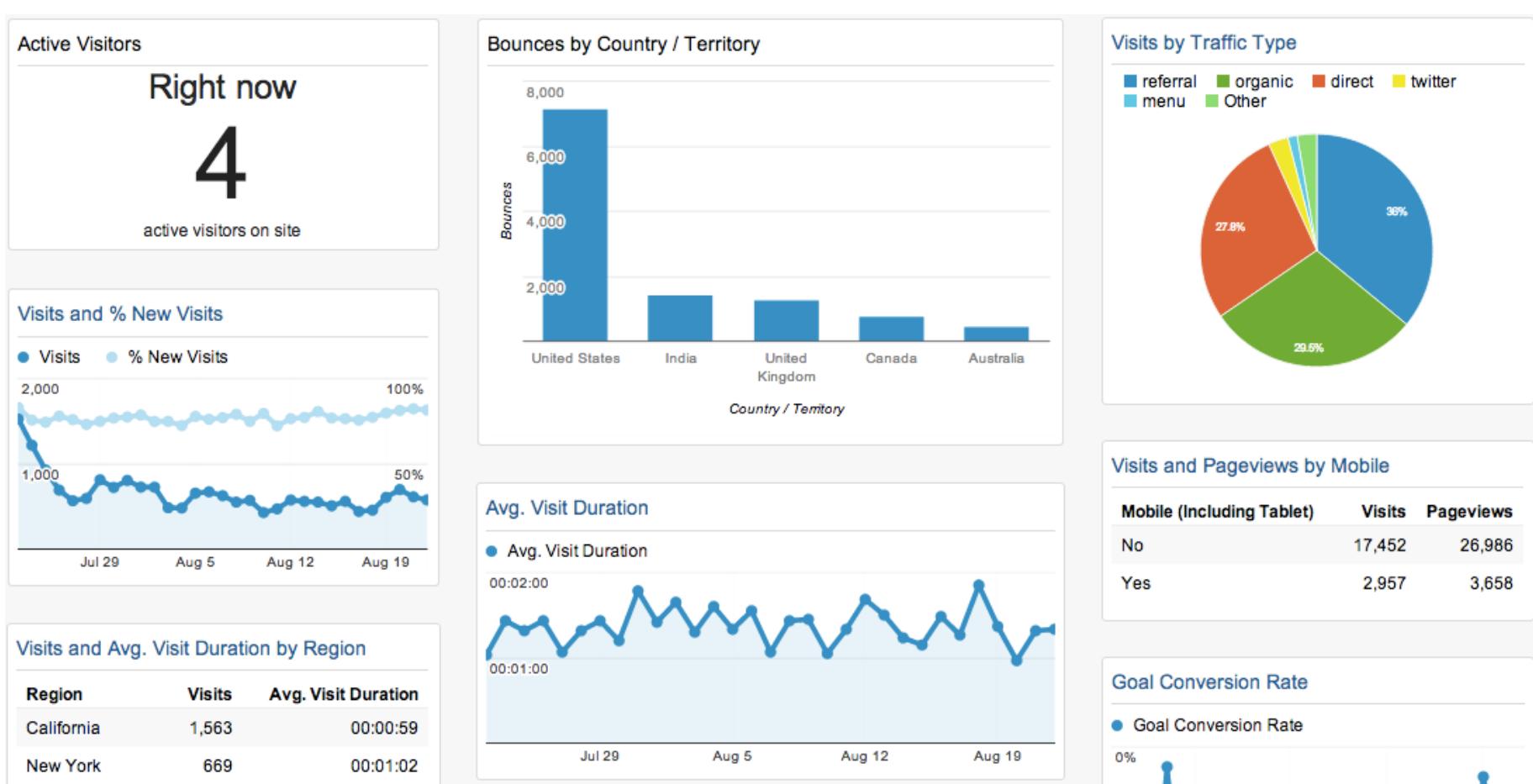
Time between two requests ≤ 10



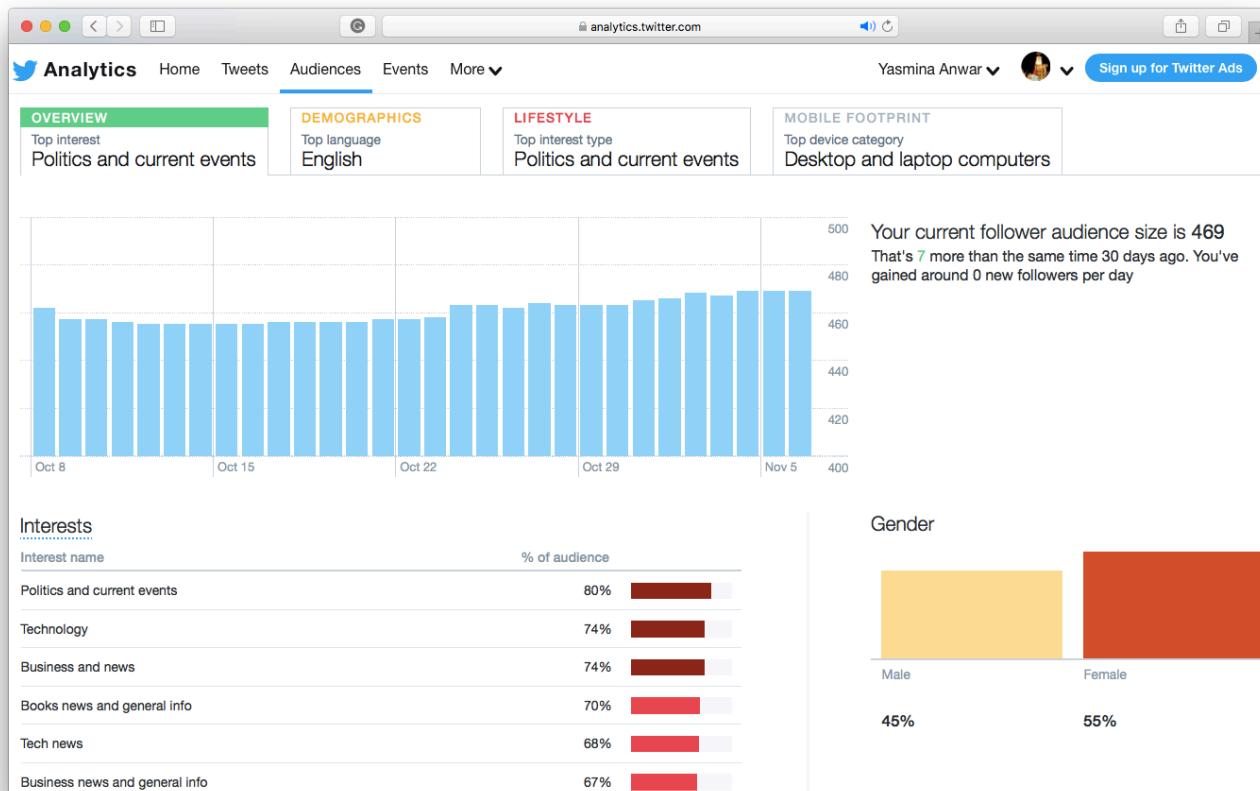
p5



Google analytics

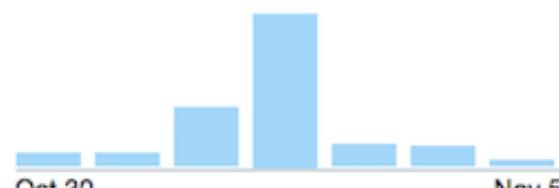


Twitter activity



Your Tweet activity

Your Tweets earned **6,463 impressions**
over the last **week**



Resources

- Chapter 12: Web Usage Mining by B. Mobasher ([pdf](#))
- Access Patterns for Robots and Humans in Web Archives ([pdf](#))

Next

- How to generate good visualizations