Week 2 - Web Data

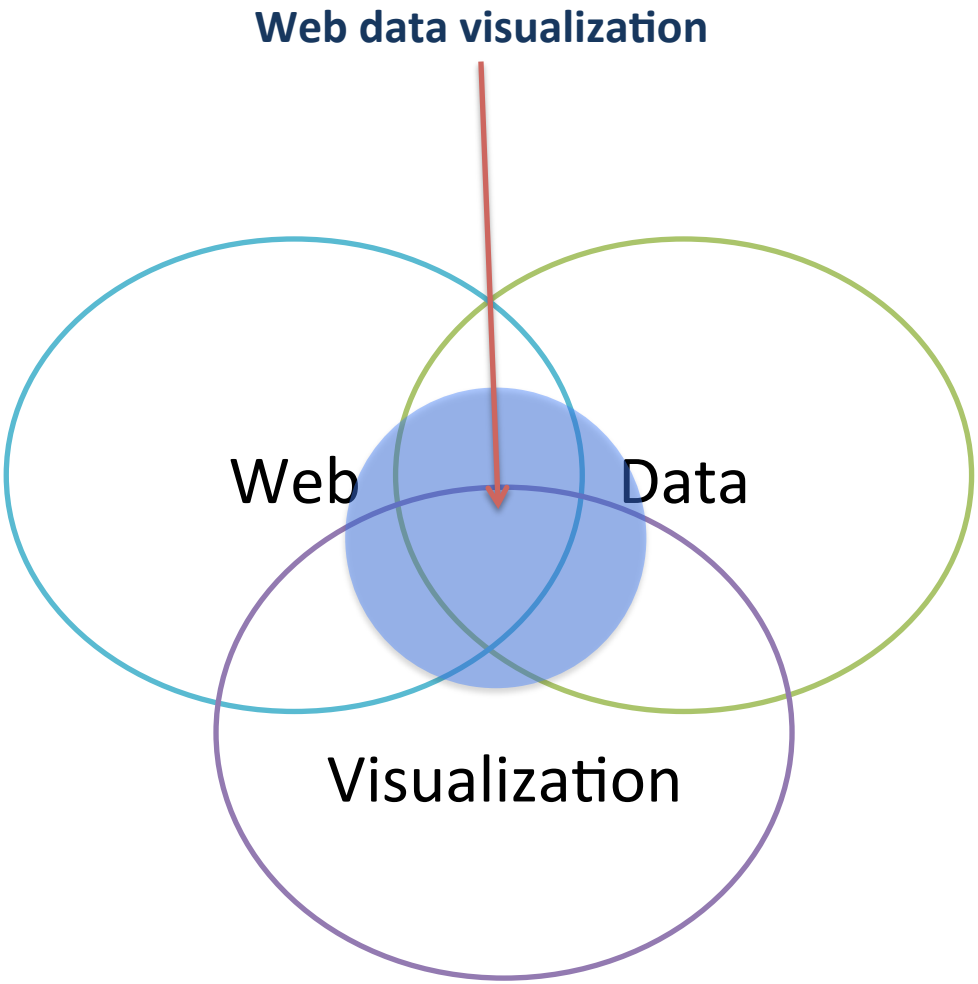
Yasmin AlNoamany

L&S 88-2

University of California, Berkeley

Previous lecture

- The Web
  - Why studying the Web is important
  - How the Web works
  - Web data formats
- Visualization
  - Intro
  - Why it is important
  - History of Visualization
- Web Data
  - importance of studying Web data
  - Types of Web data
  - Data formats



Today's lecture

- Types of Web data
- Introduction to Web Mining
- Getting data from the Web
  - Working with APIs
  - Web scraping (moved to the next lecture)

Resources

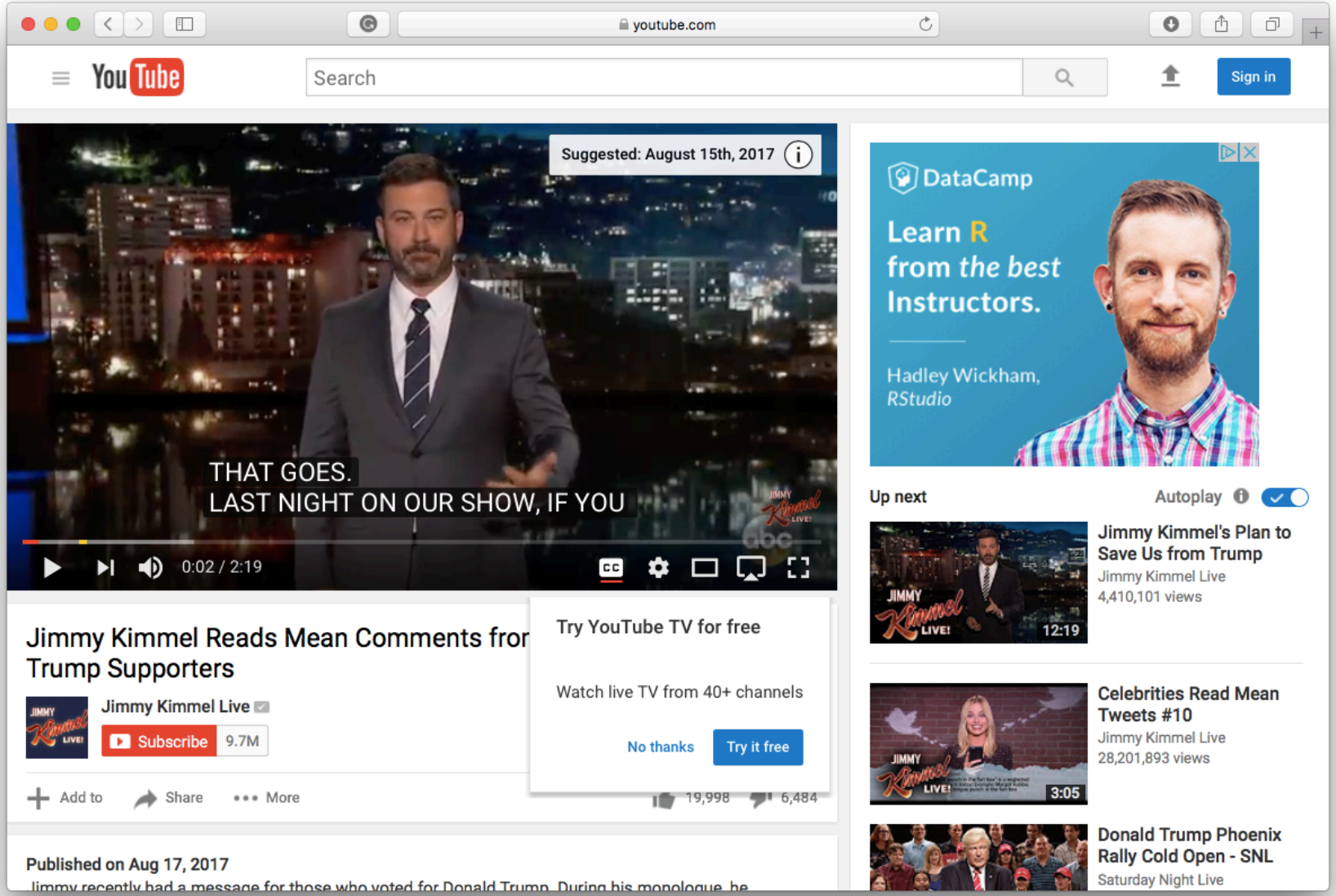
- Web Mining: Accomplishments & Future Directions by Jaideep Srivastava
- Working with data on the Web
- Intro to Web Scrapping
- Working with APIs:
  - Data and twitter
  - Accessing Databases via Web APIs

Objectives

- Distinguish the Web data types
- Identify Web mining
- Explain the difference between Web scraping and working with APIs
- Identify how to get data from Web APIs
- Understand how HTML works with your browser to display a website
- Identify HTML tags and attributes

Web data

Web content


youtube.com
YouTube
Search
Sign in
Suggested: August 15th, 2017
THAT GOES.
LAST NIGHT ON OUR SHOW, IF YOU
0:02 / 2:19
Jimmy Kimmel Reads Mean Comments fror
Trump Supporters
Jimmy Kimmel Live
Subscribe 9.7M
Add to
Share
More
19,998
6,484
Published on Aug 17, 2017
Try YouTube TV for free
Watch live TV from 40+ channels
No thanks
Try it free
DataCamp
Learn R
from the best
Instructors.
Hadley Wickham,
RStudio
Up next
Autoplay
Jimmy Kimmel's Plan to Save Us from Trump
Jimmy Kimmel Live
4,410,101 views
12:19
Celebrities Read Mean Tweets #10
Jimmy Kimmel Live
28,201,893 views
3:05
Donald Trump Phoenix Rally Cold Open - SNL
Saturday Night Live

Web content

cnn.com

NEWS ALERT

There are fears of catastrophic flooding as Harvey moves inland. Watch CNN

CNN Home Live TV U.S. Edition +

Widespread damage after Harvey hits Texas

BREAKING NEWS

Debris is scattered across a wide swath of Southeast Texas after the Category 4 hurricane roared ashore

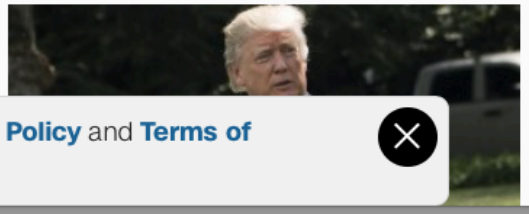
LIVE UPDATES Harvey now a tropical storm. Heavy rain continues

WATCH LIVE Harvey brings heavy rain to Houston

'This says

Latest Hurricane Harvey's wrath

- Texas governor: Flooding is our top priority right now
- 38 min Harvey has been downgraded to a tropical storm
- 1 hr Rockport resident: Harvey 'sounded like a freight train with square wheels'
- 1 hr Soon: Texas Governor gives hurricane update



Top stories

North Korea launches missiles days after US praised restraint

Report: Mueller looking at whether Flynn sought Clinton emails from Russian hackers

The spies caught in the act

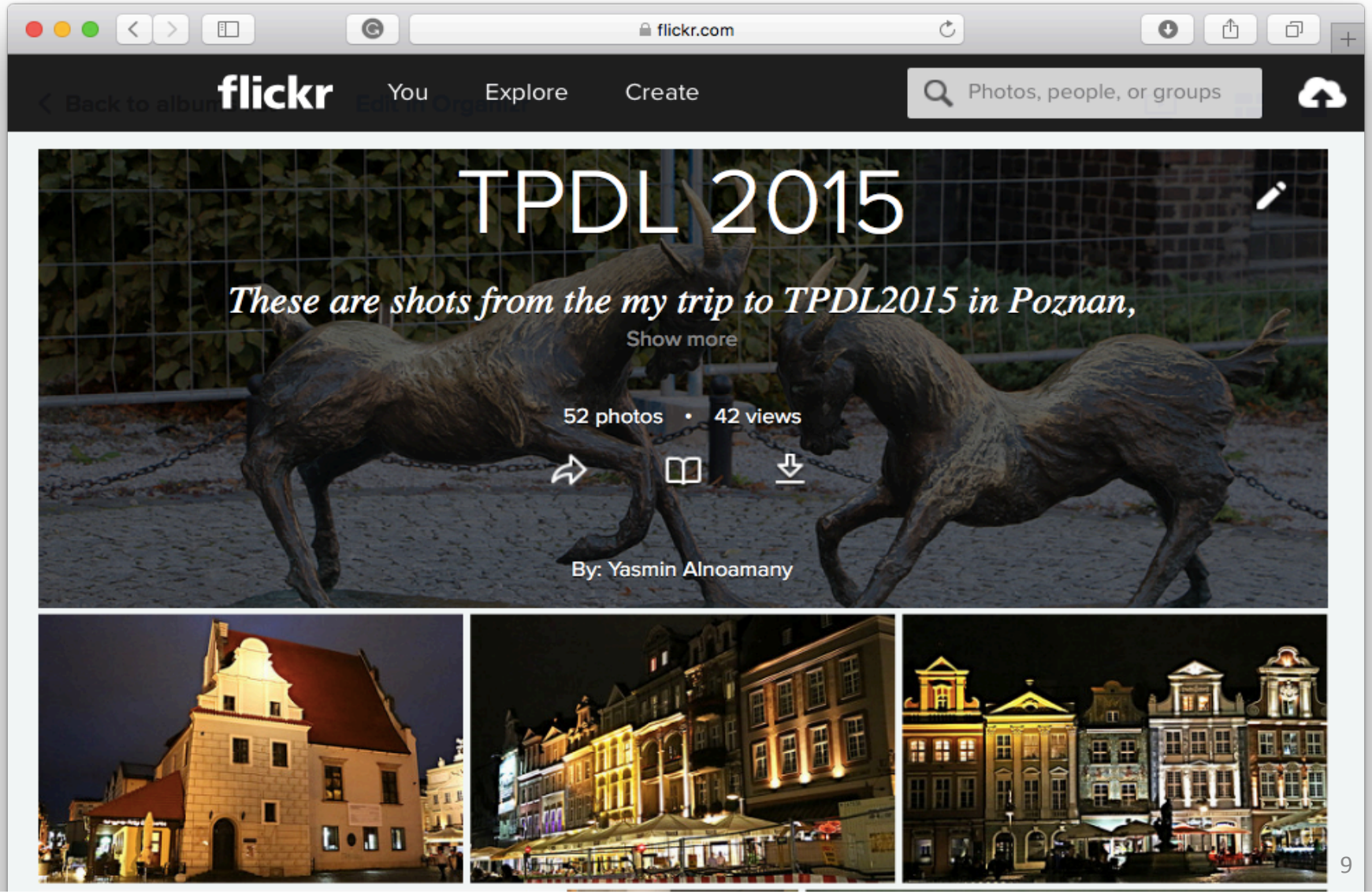
Gorka out as White House adviser

Phoenix reacts to Arpaio's pardon

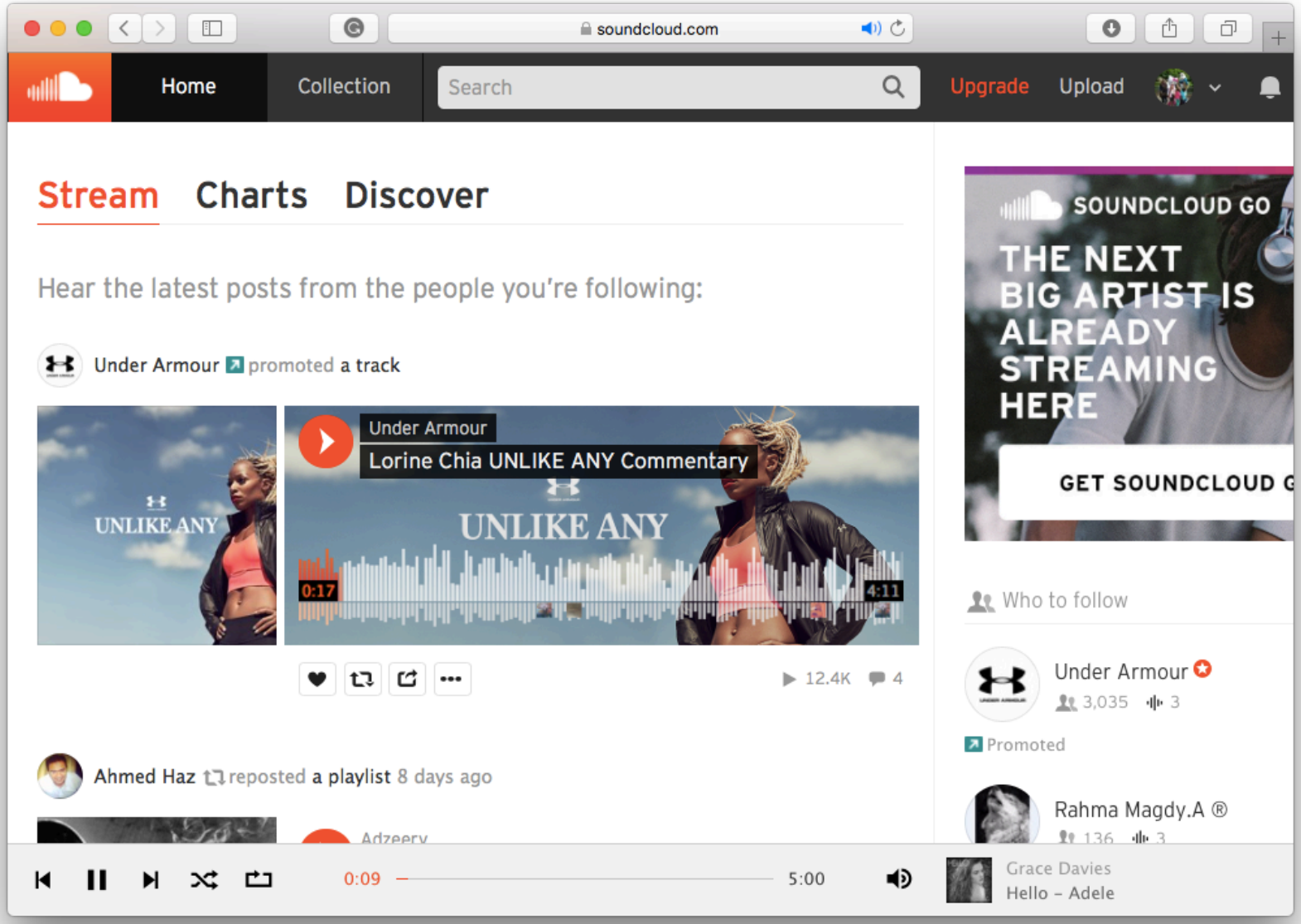
Opinion: Trump, the imperial president

By using this site, you agree to the **Privacy Policy** and **Terms of Service**.

Web content

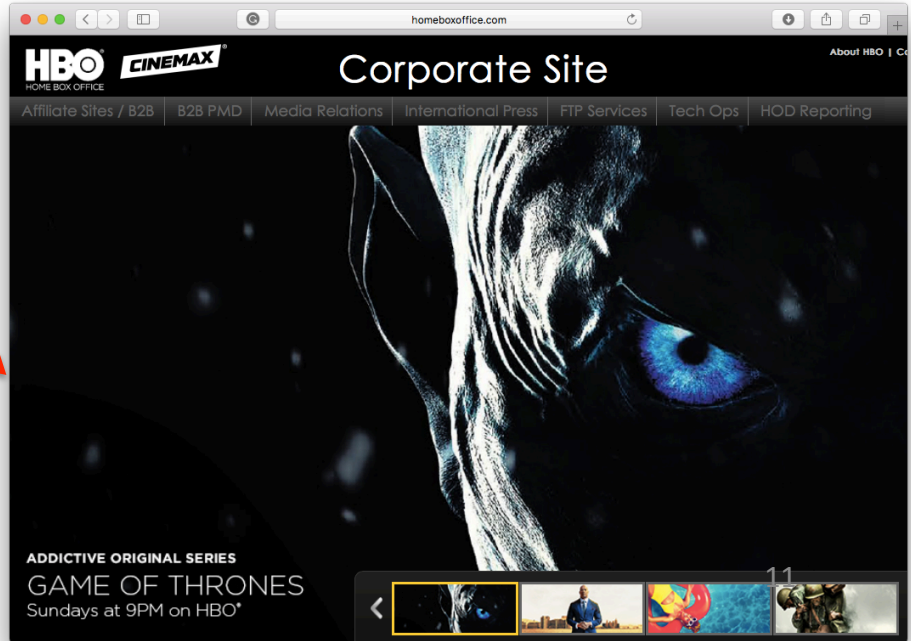

flickr.com
flickr
You
Explore
Create
Photos, people, or groups
TPDL 2015
These are shots from the my trip to TPDL2015 in Poznan,
Show more
52 photos • 42 views
By: Yasmin Alnoamany

Web content


soundcloud.com
Home
Collection
Search
Upgrade
Upload
Stream
Charts
Discover
Hear the latest posts from the people you're following:
Under Armour promoted a track
UNLIKE ANY
Under Armour
Lorine Chia UNLIKE ANY Commentary
UNLIKE ANY
0:17
4:11
12.4K
4
Ahmed Haz reposted a playlist 8 days ago
Adzeery
SOUNDCLOUD GO
THE NEXT BIG ARTIST IS ALREADY STREAMING HERE
GET SOUNDCLOUD G
Who to follow
Under Armour
3,035
3
Promoted
Rahma Magdy.A ®
136
3
0:09
5:00
Grace Davies
Hello – Adele

Web Structure

Graph Theory: Pages are nodes & links are directed edges

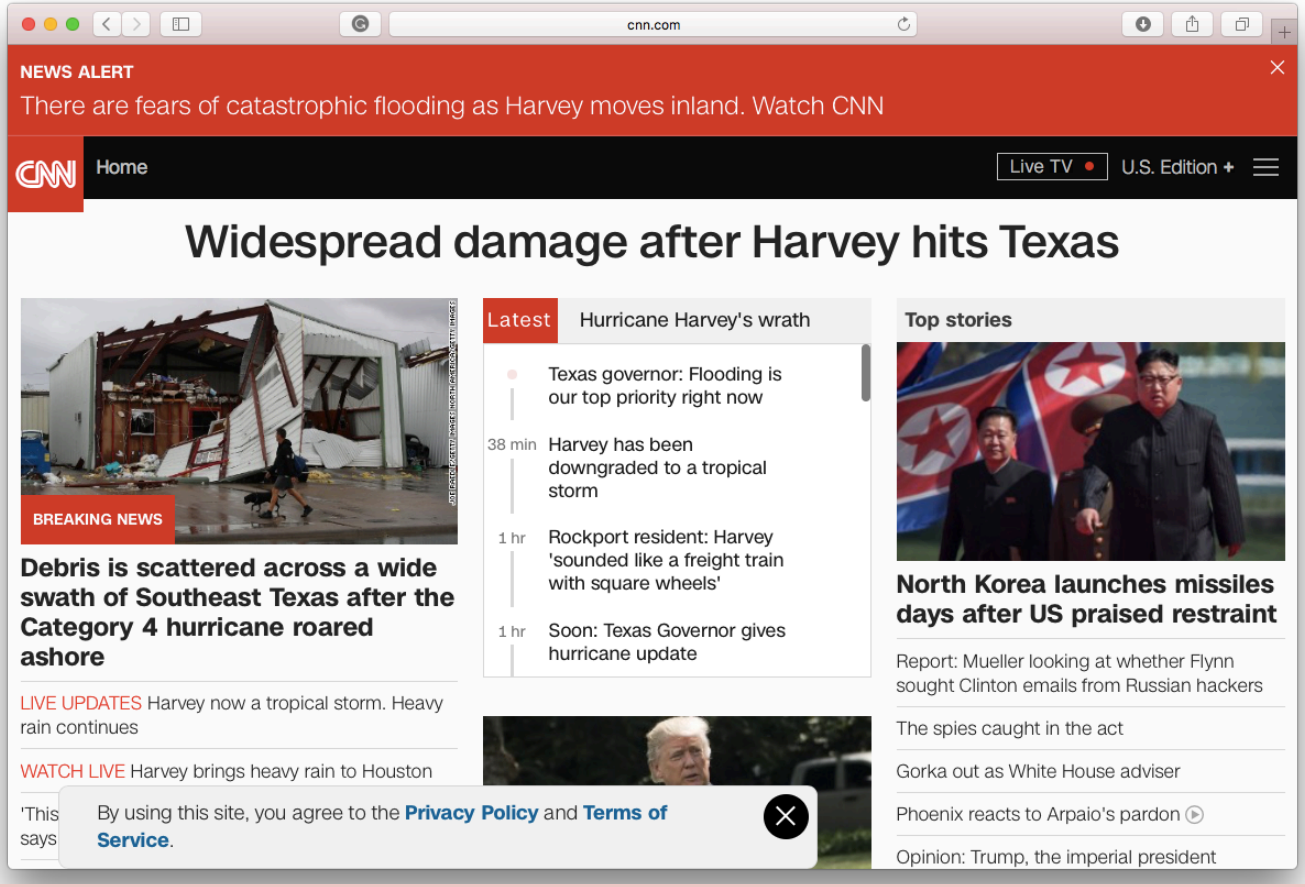


Web Structure

Source: https://motherboard.vice.com/en_us/article/ezz7be/search-is-the-future-of-facebook-but-only-if-you-like-it

Web usage

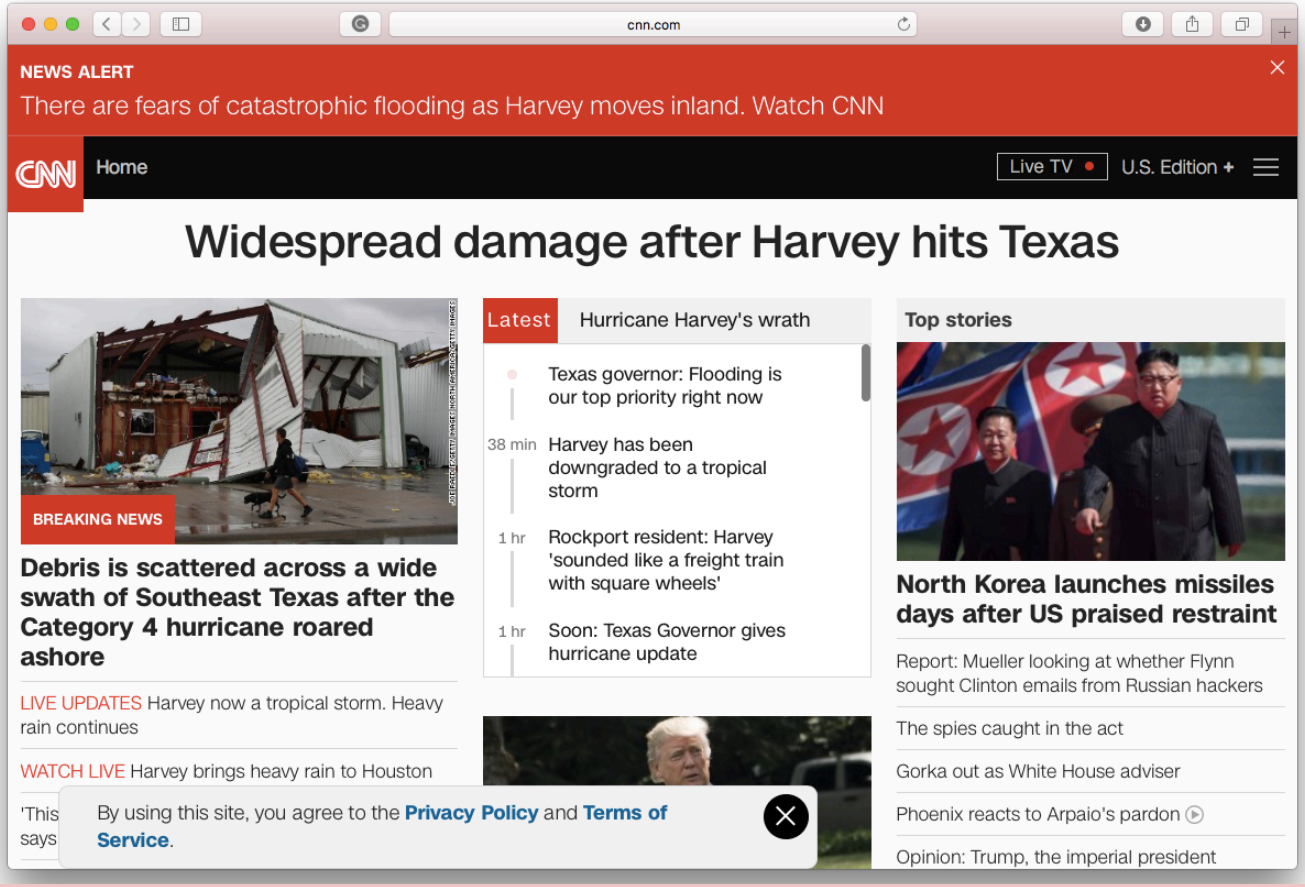
http://www.cnn.com




```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.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"
```

Web usage

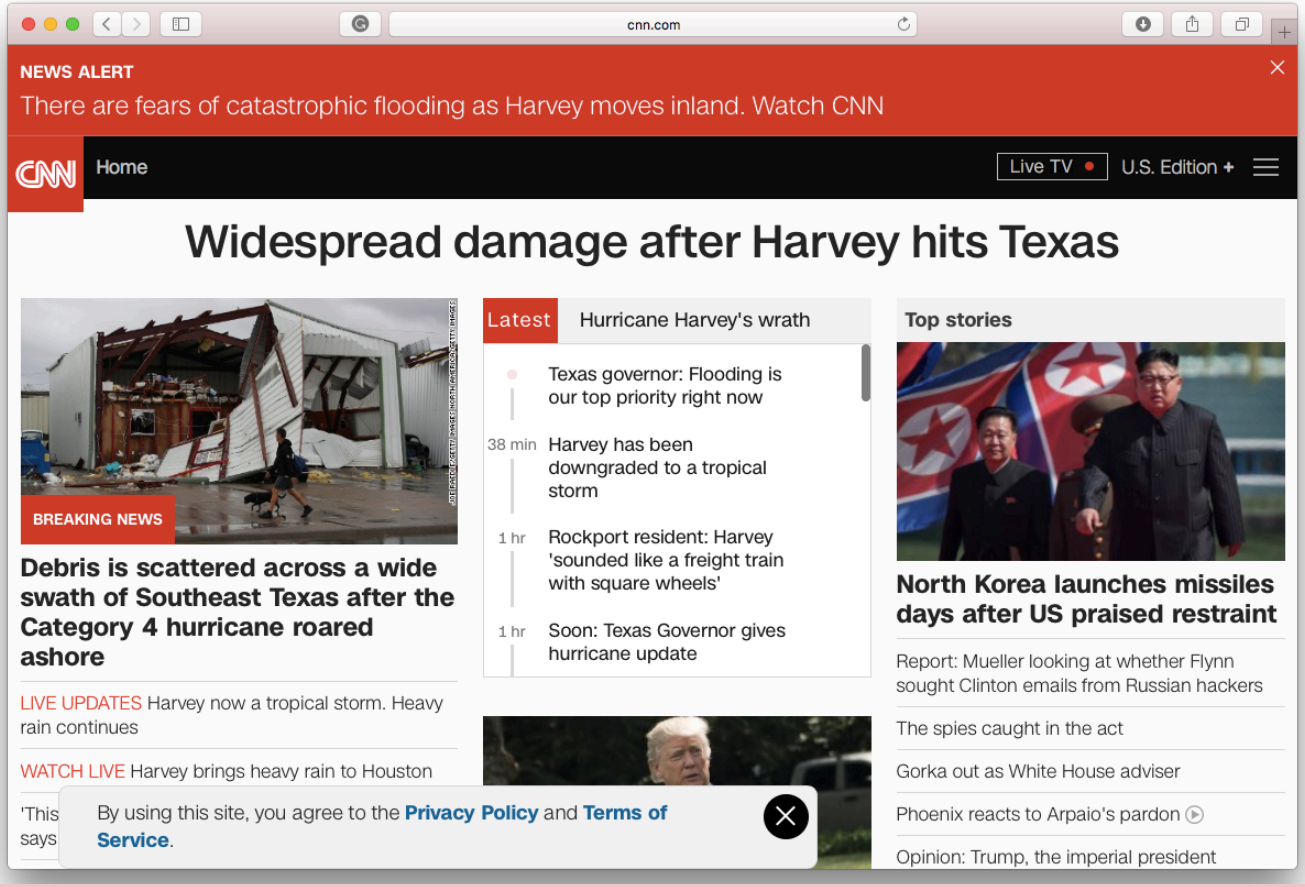
http://www.cnn.com




```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.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"
```

Web usage

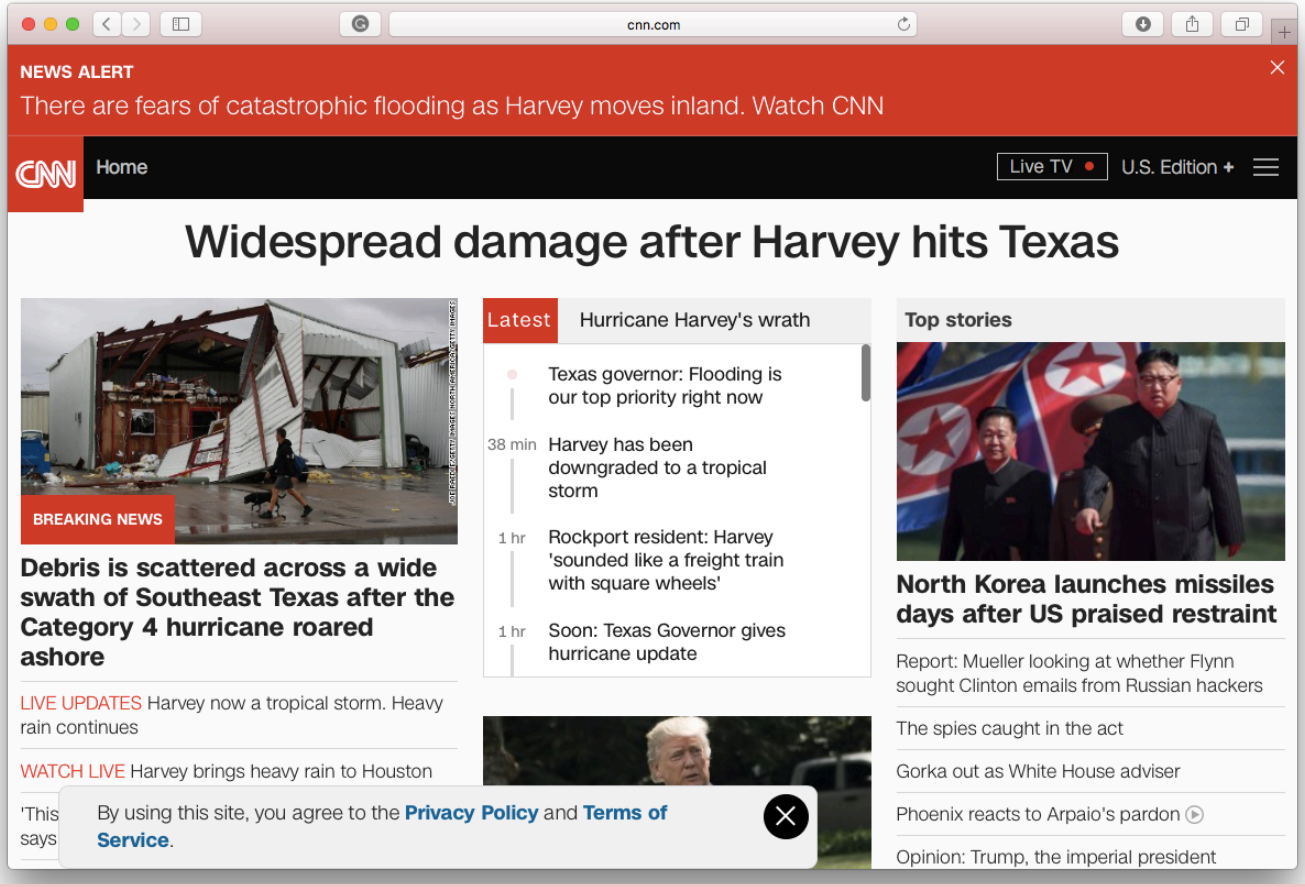
http://www.cnn.com




```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.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"
```

Web usage

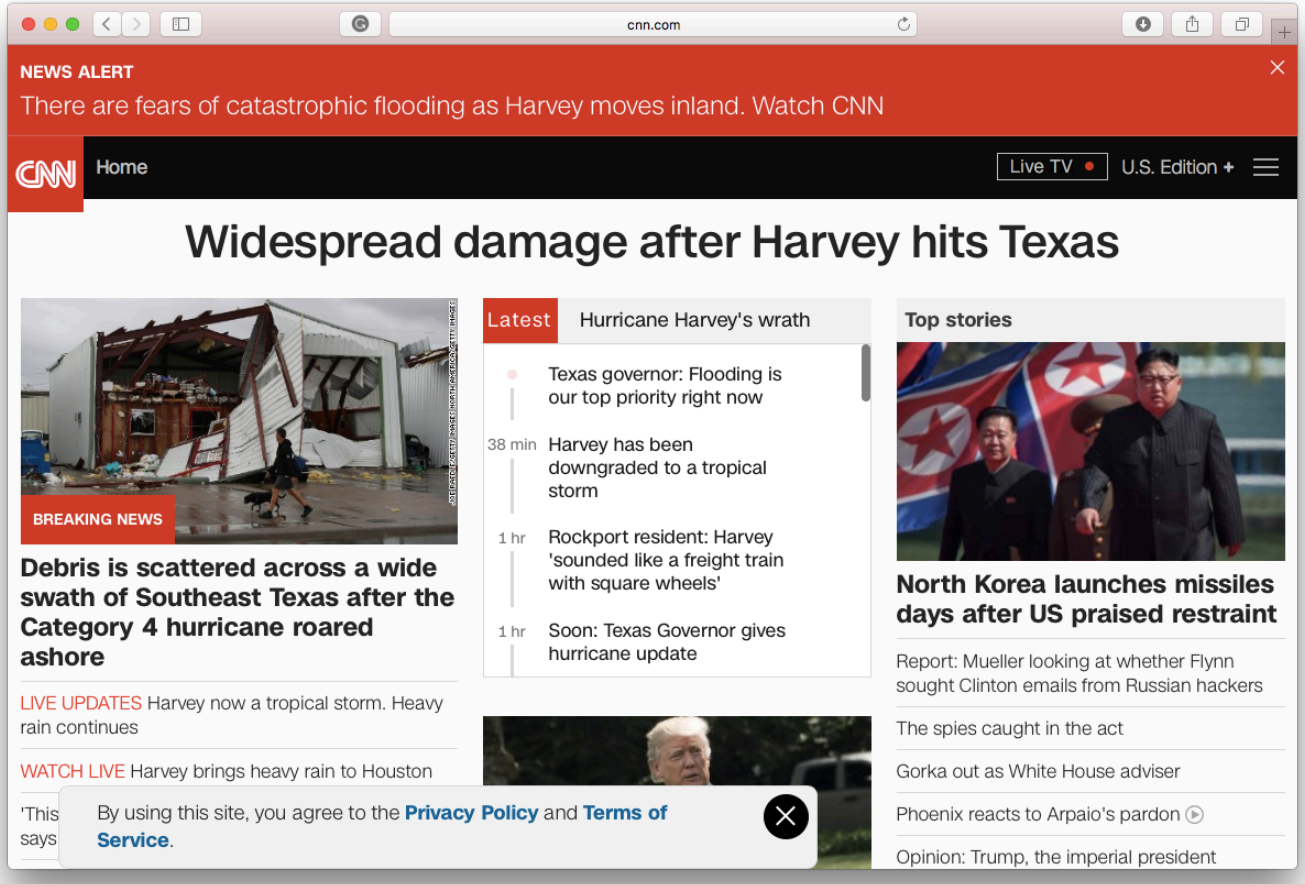
http://www.cnn.com




```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.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"
```

Web usage

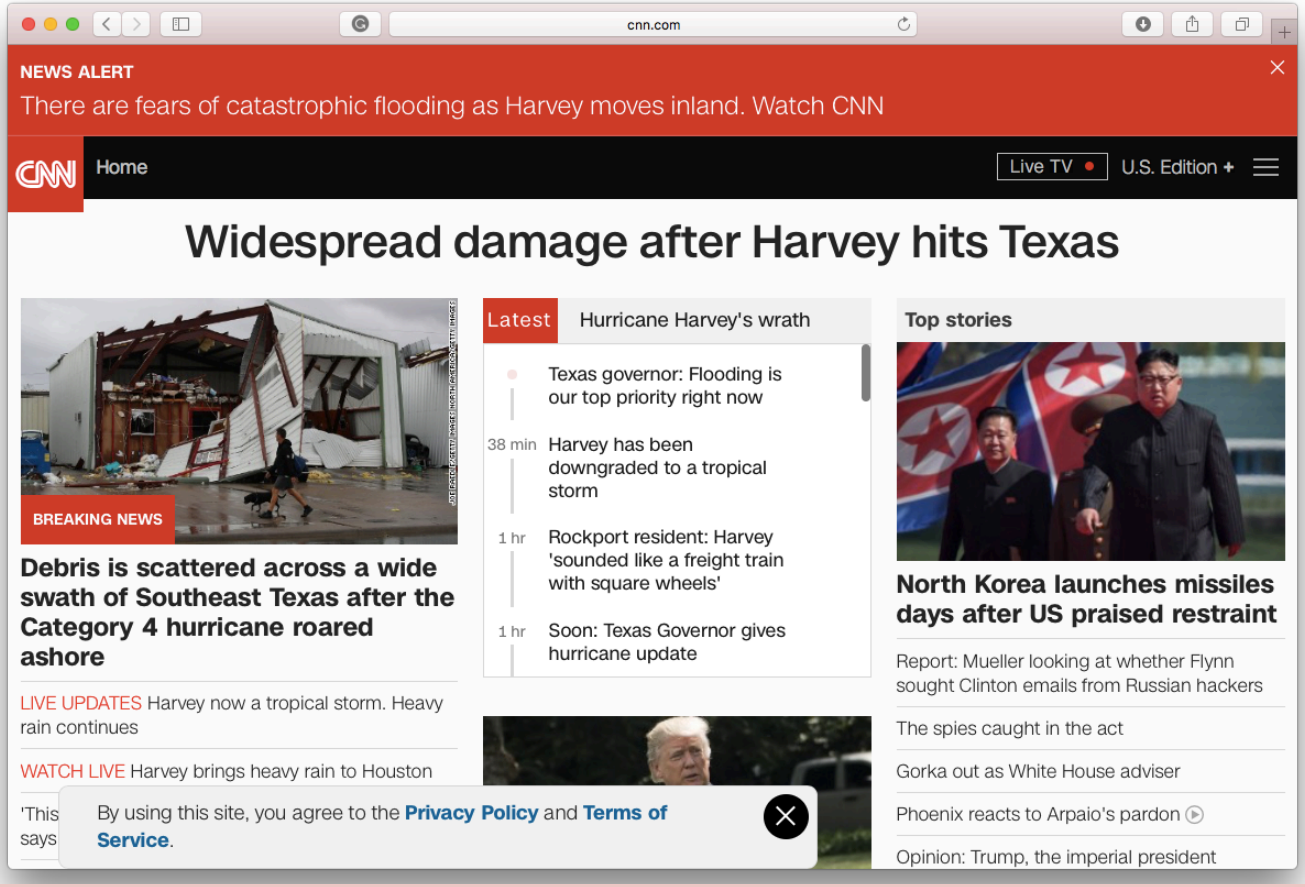
http://www.cnn.com




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

Web usage

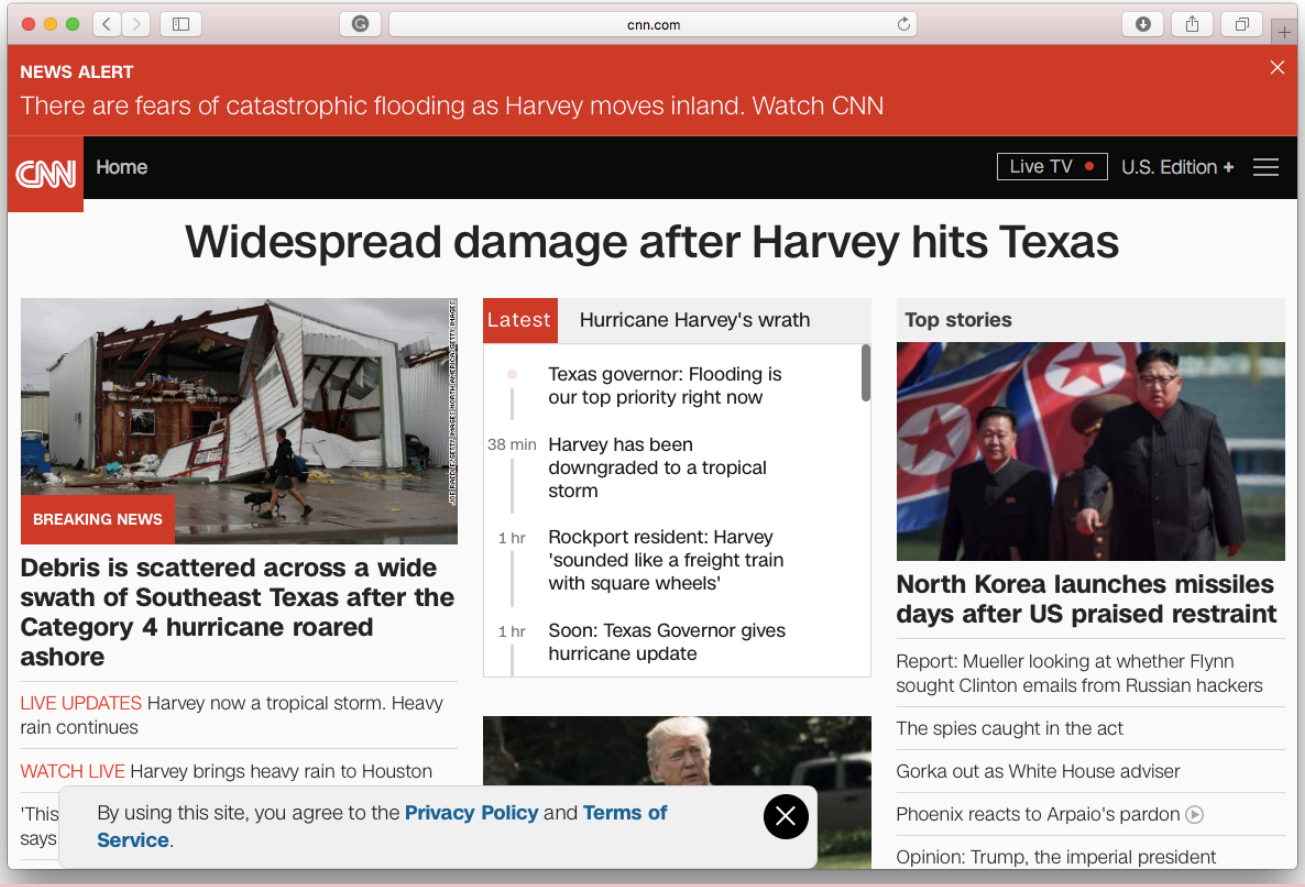
http://www.cnn.com




```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.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"
```

Web usage

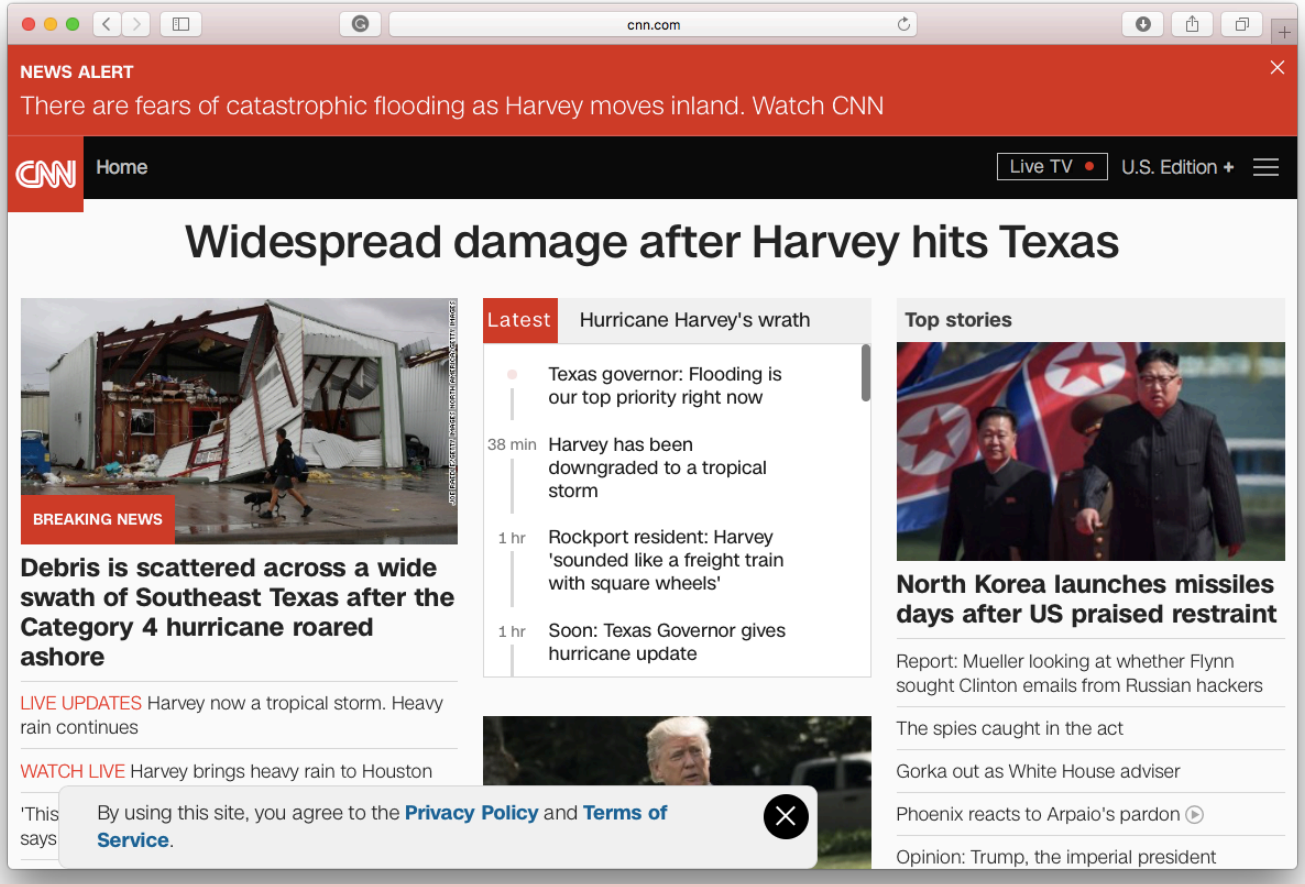
http://www.cnn.com




```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.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"
```

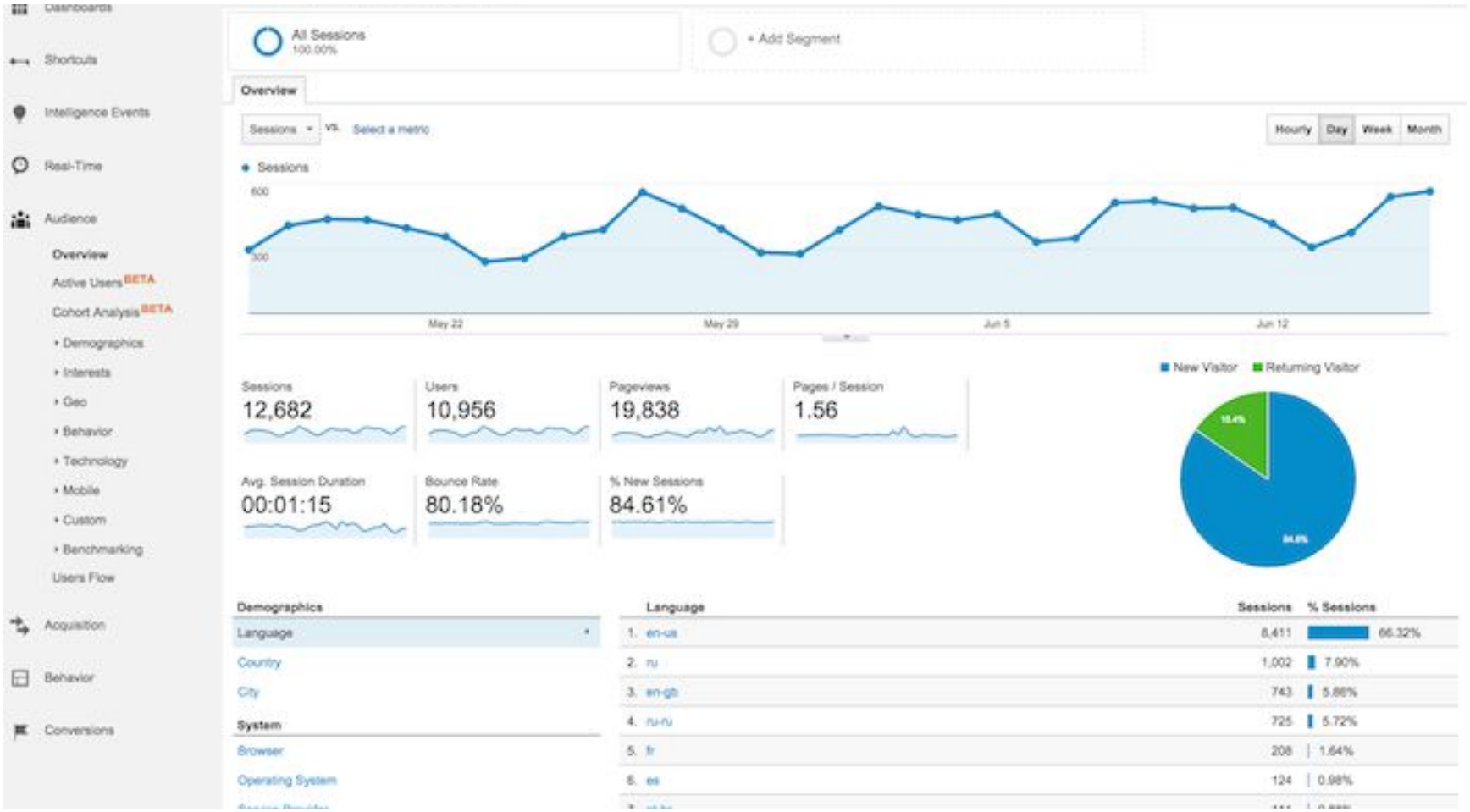
Web usage

http://www.cnn.com



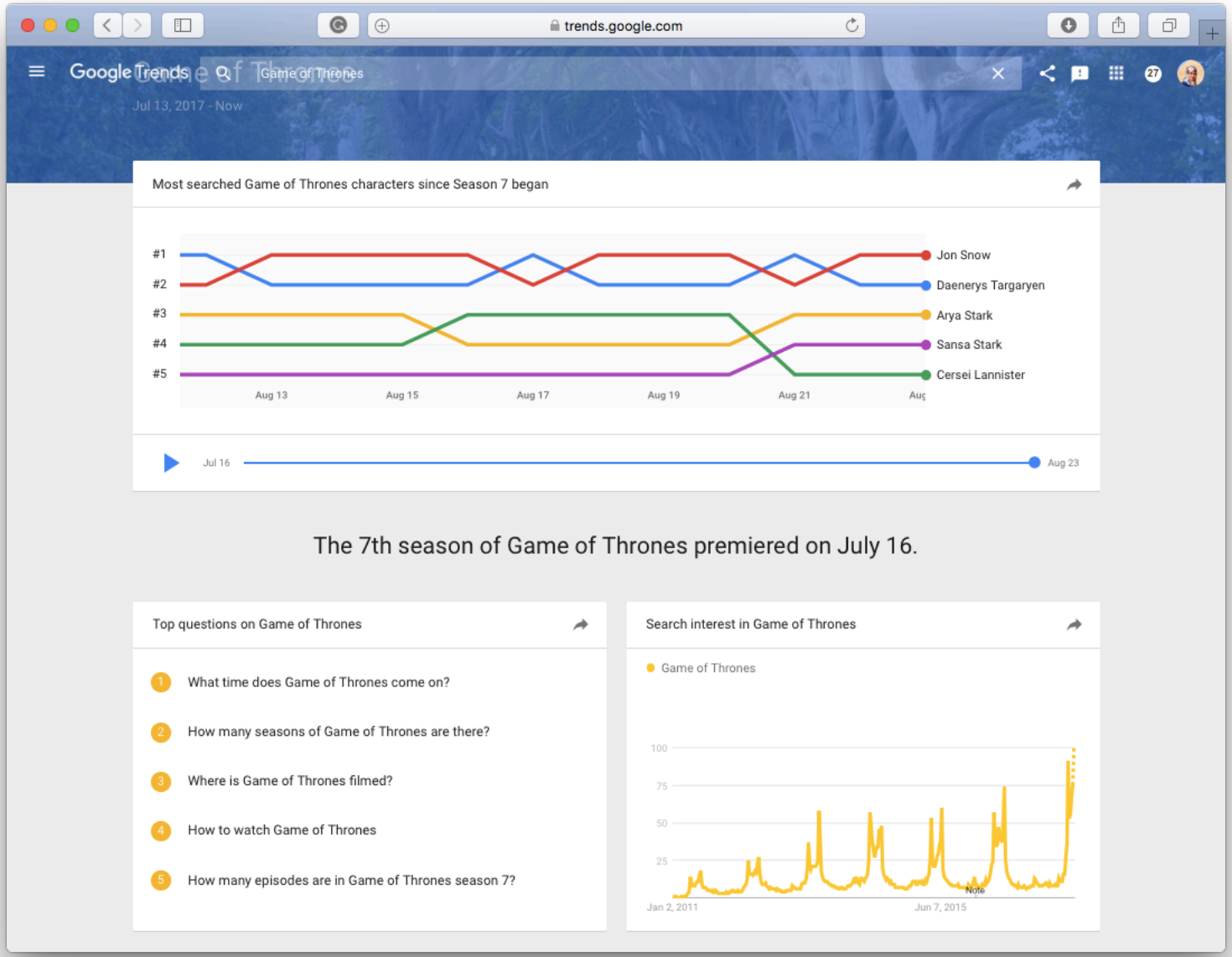

```
0.247.222.86 - - [0/Sep/2017:07:03:46 +0000] "GET http://www.cnn.com HTTP/
1.1" 200 96433 ”-" "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_6_8)
AppleWebKit/535.7 (KHTML, like Gecko) Chrome/16.0.912.77 Safari/535.7"
```

Web usage

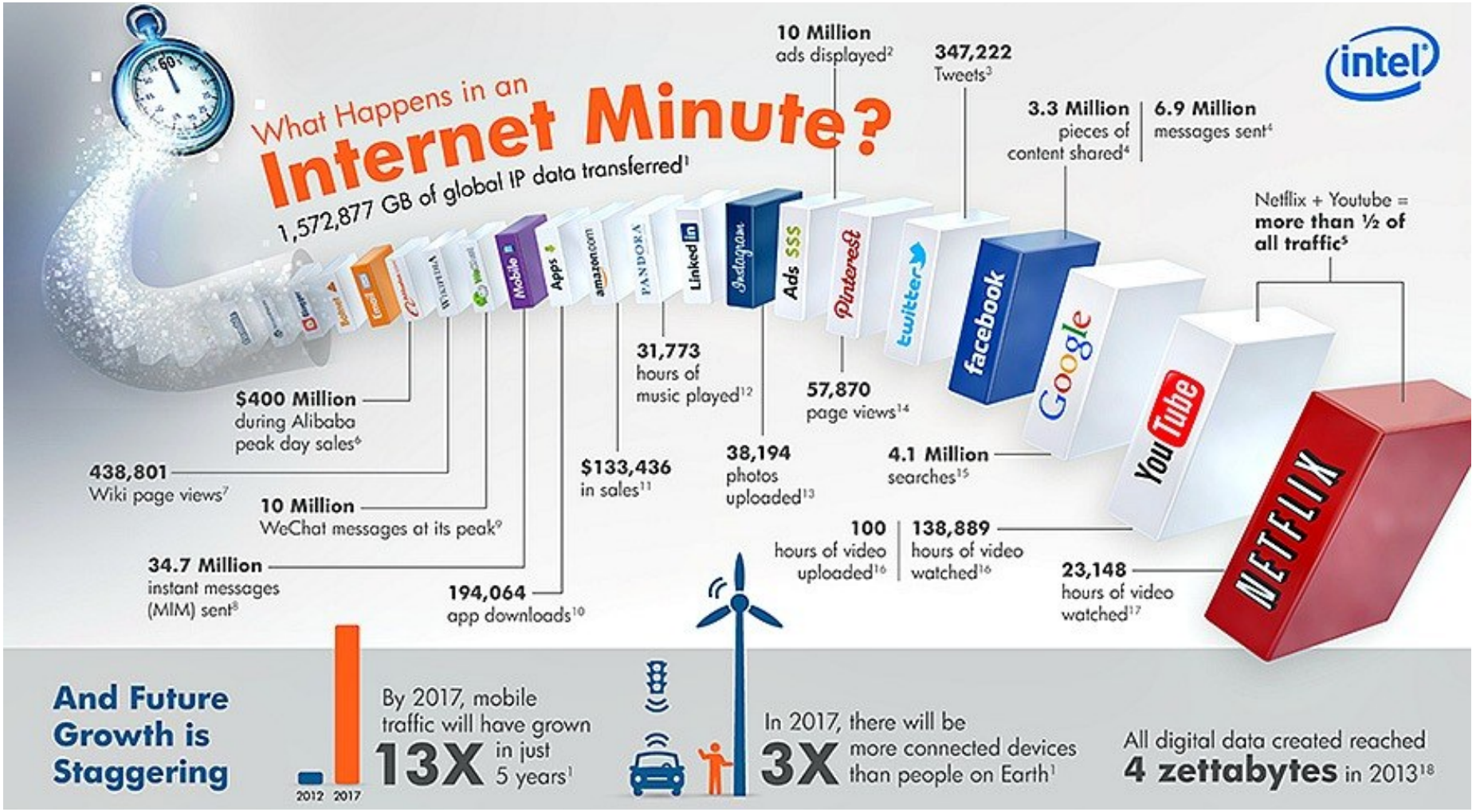


Source: https://moz.com/blog/absolute-beginners-guide-to-google-analytics

Using the Web to understand the Web



What happens in the Internet minute



Source: https://carestruck.org/happens-internet-minute/

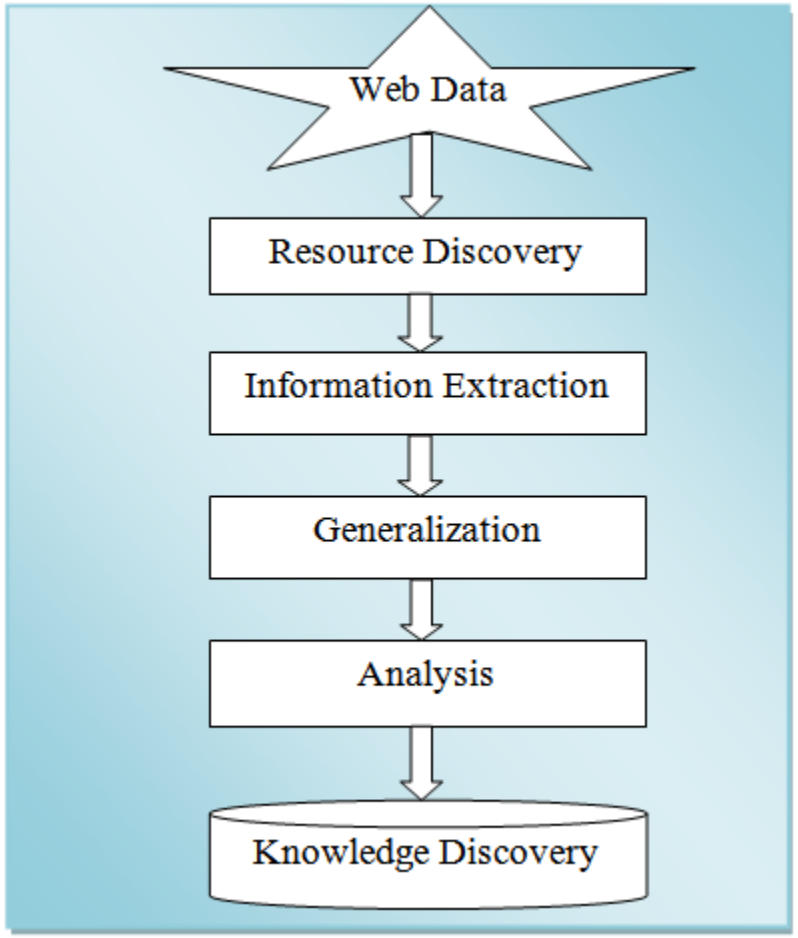
Web mining definition

- *Web* is a collection of inter-related files on one or more *Web servers.*

- Web mining is the application of data mining techniques to extract knowledge from Web data

The term created by Orem Etzioni (1996)

Decomposition of Web mining

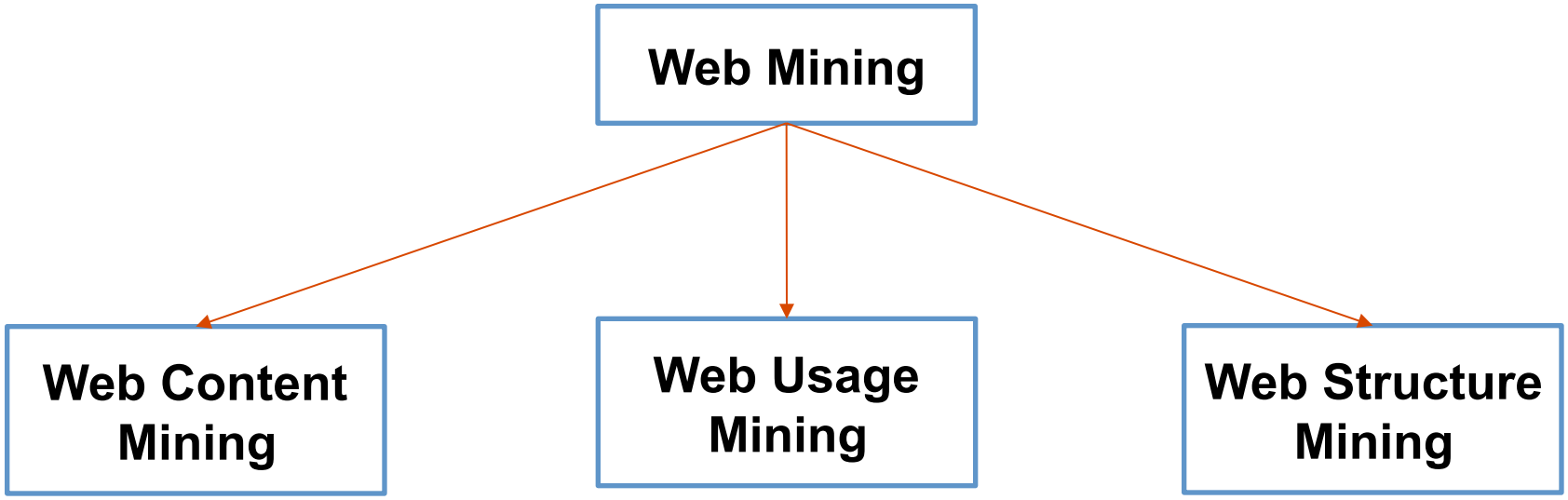


Shyam Nandan Kumar et al. World towards Advance Web Mining: A Review. American Journal of Systems and Software, 2015, Vol. 3, No. 2, 44-61. doi:10.12691/ajss-3-2-3

Web data vs data mining

- Traditional data mining
  - data is structured and relational
  - well-defined tables, columns, rows, keys, and constraints.
- Web data
  - Semi-structured and unstructured
  - readily available data
  - rich in features and patterns

Types of Web Mining



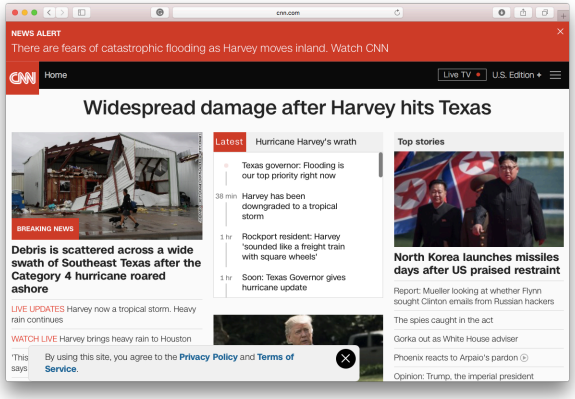
Types of Web Mining

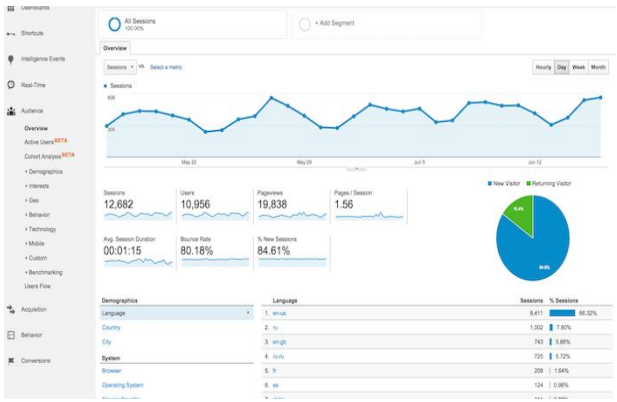
Web Mining

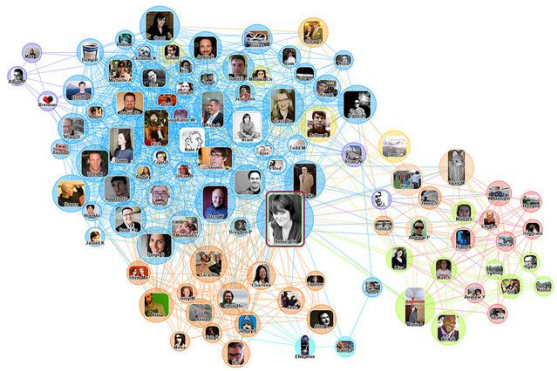
Web Content Mining

Web Usage Mining

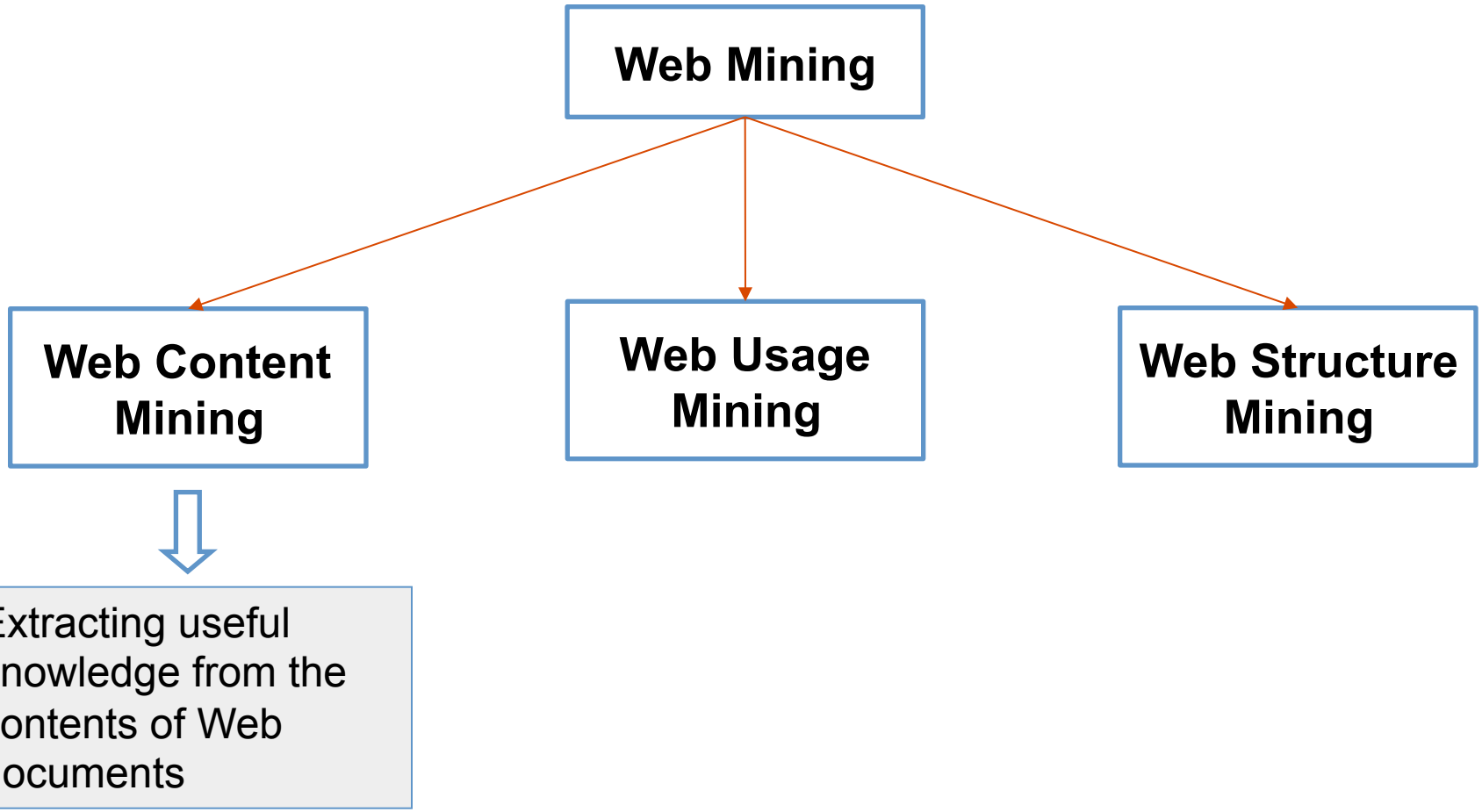
Web Structure Mining



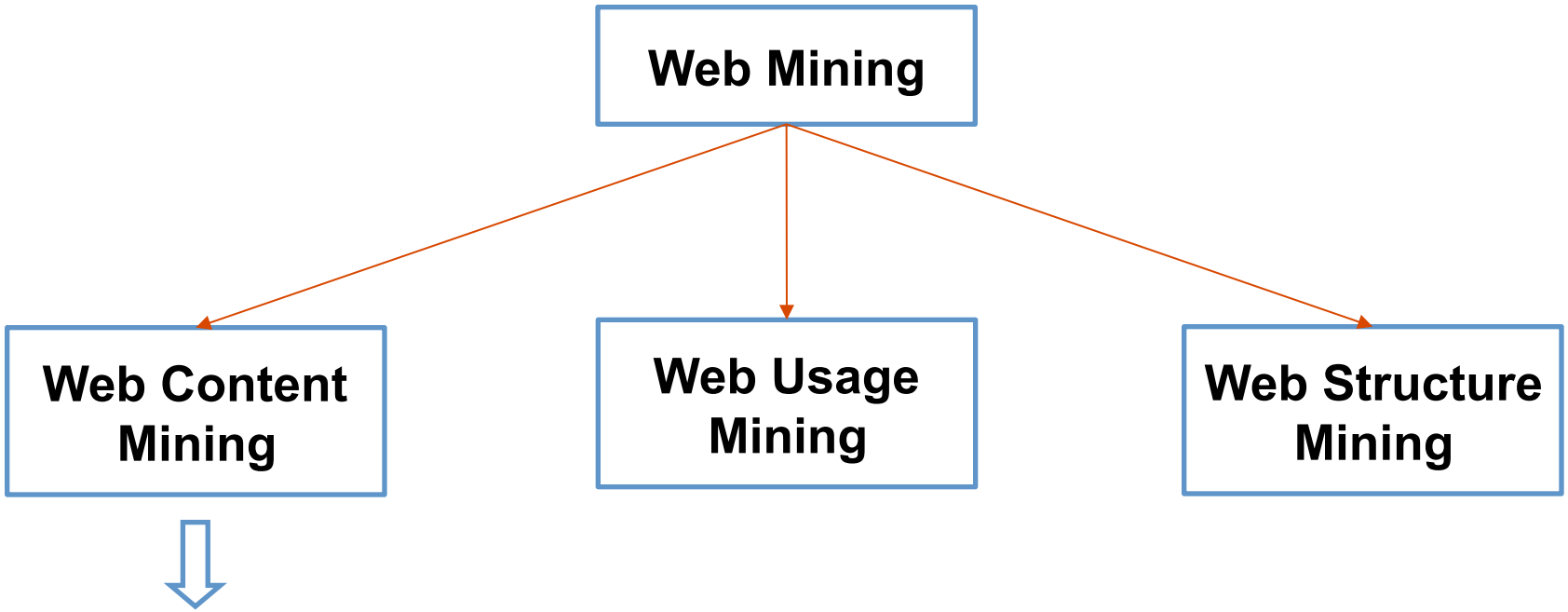




Types of Web Mining

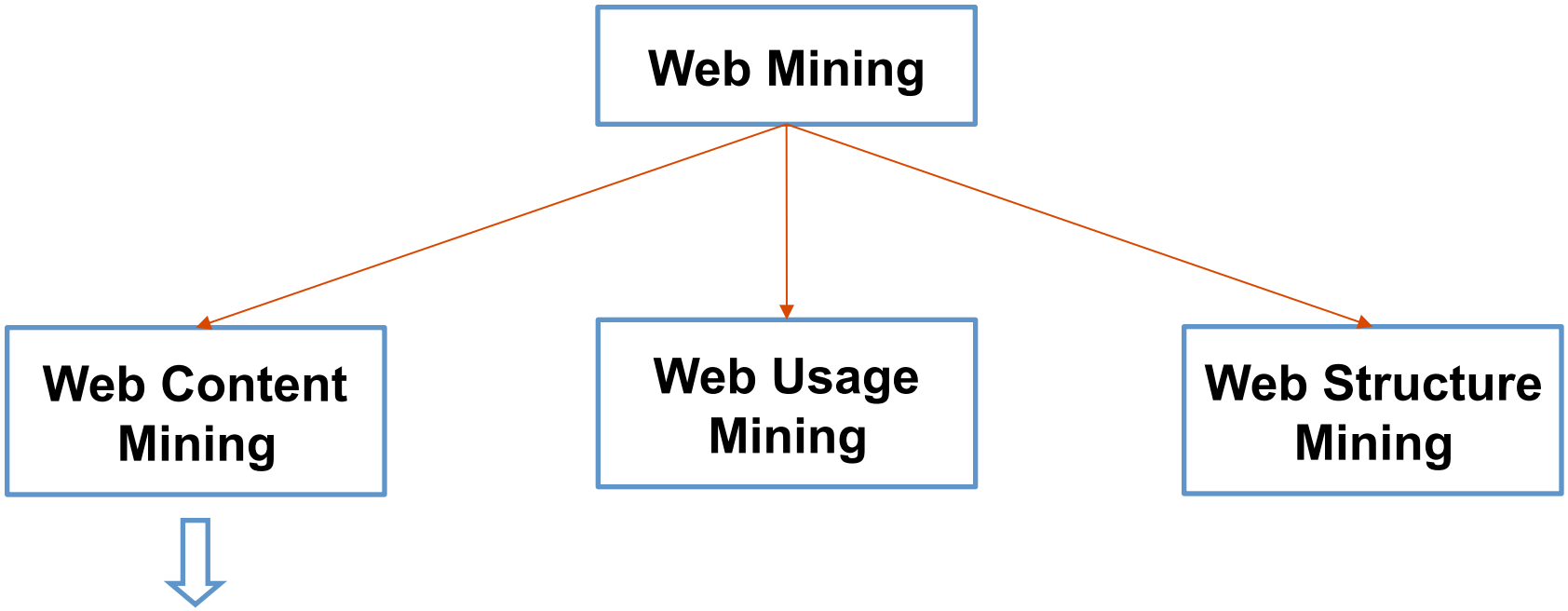


Types of Web Mining



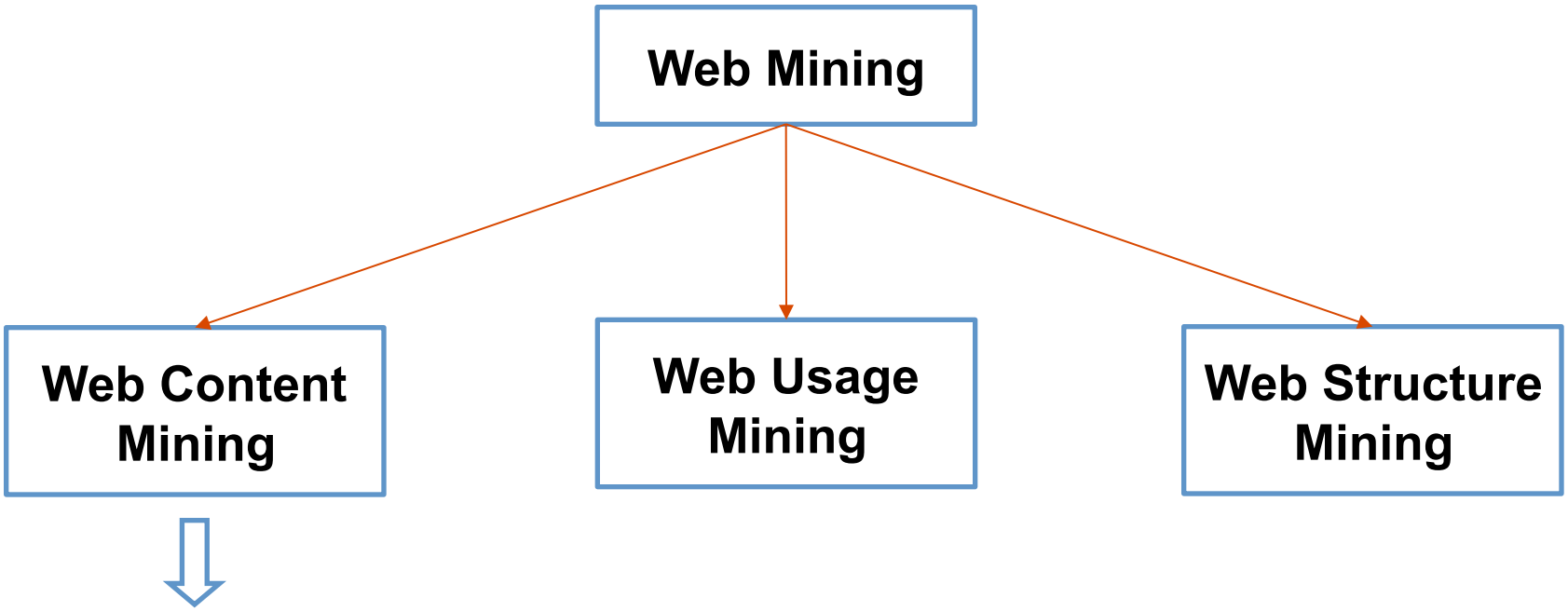
- text
- images
- audio
- video
- structured records from lists and tables
- item attributes from backend databases

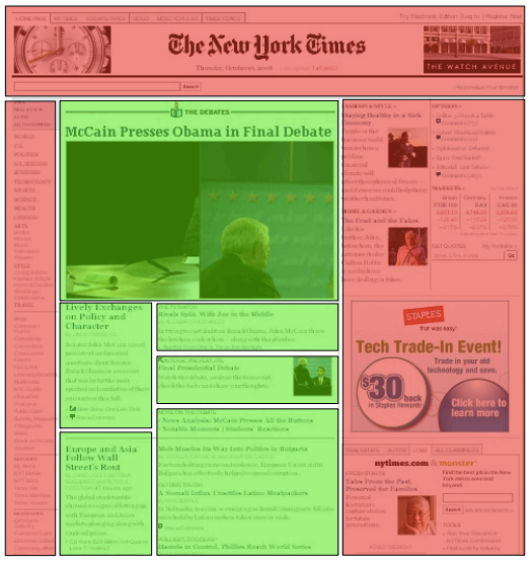
Types of Web Mining



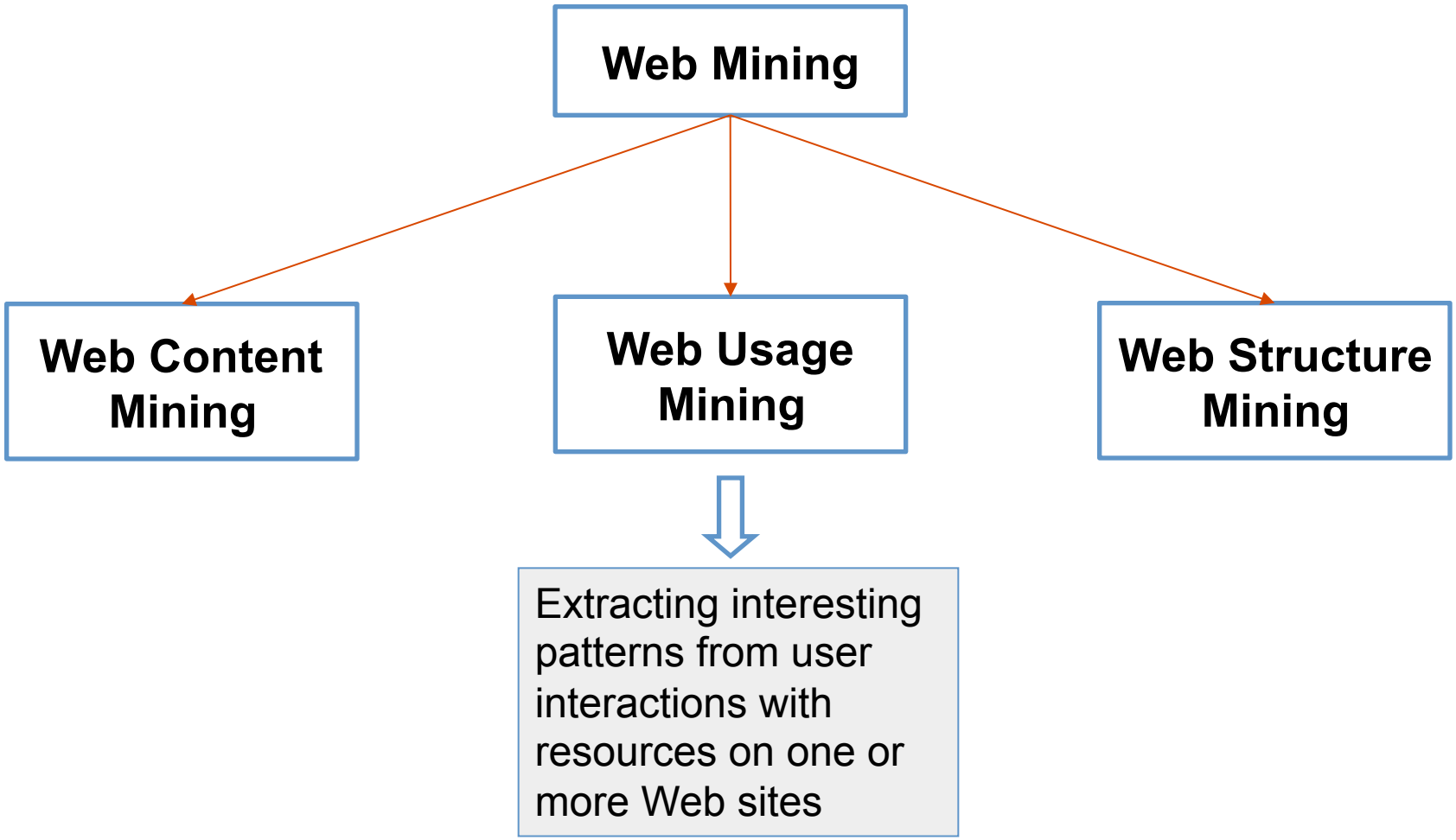
Applications:
- document clustering or categorization
- topic identification / tracking
- concept discovery
- focused crawling
- content-based personalization
- intelligent search tools

Types of Web Mining

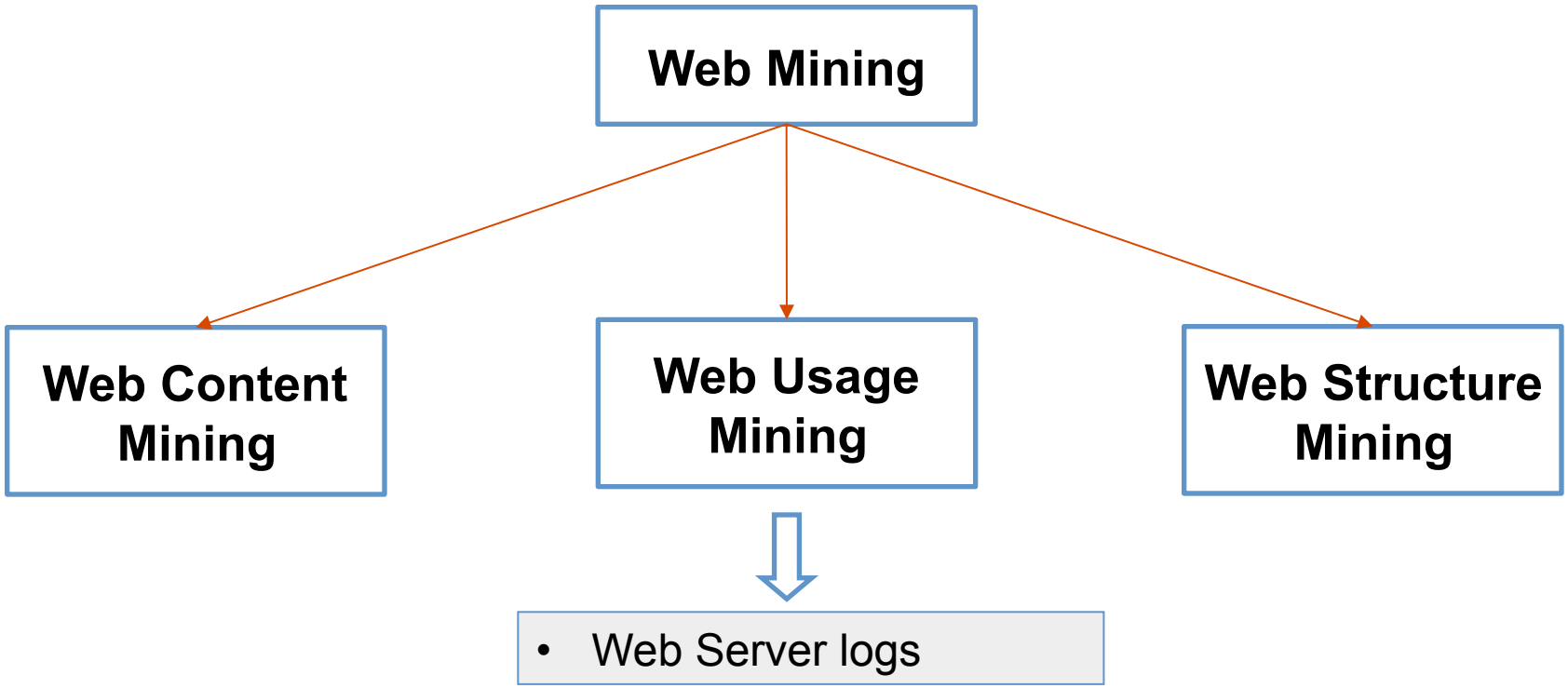




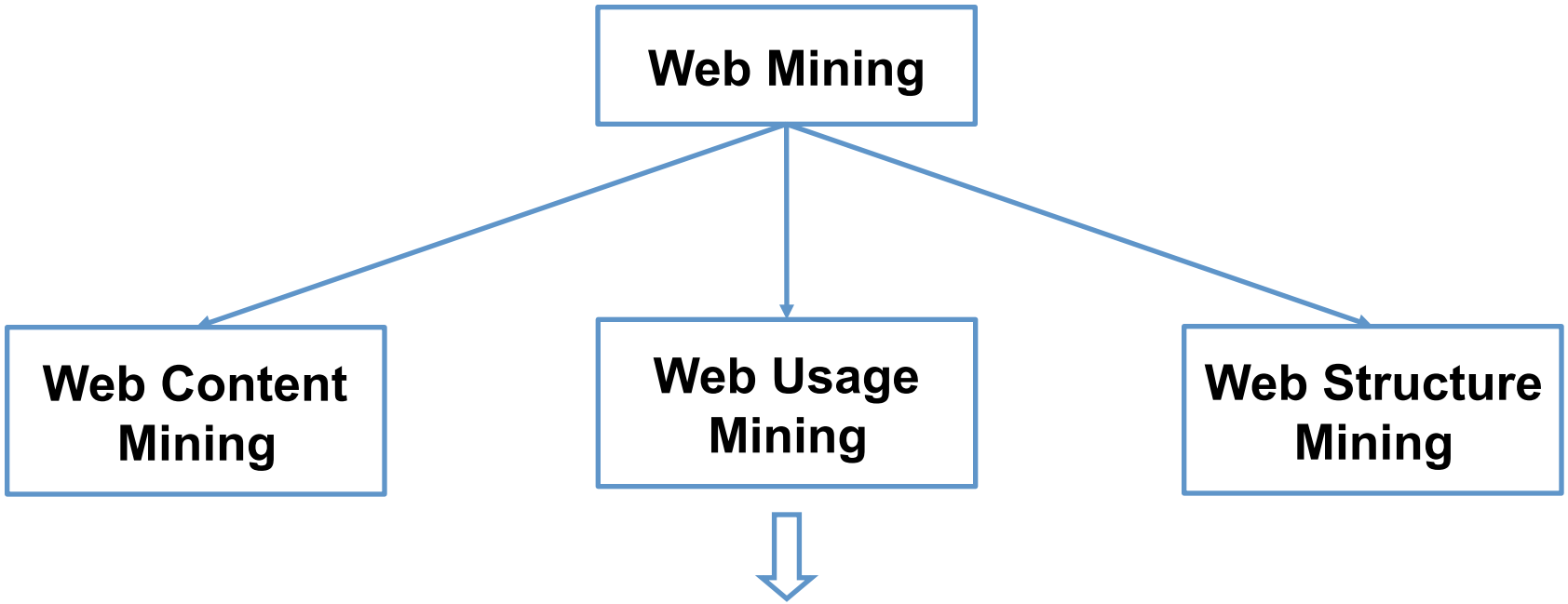
Types of Web Mining



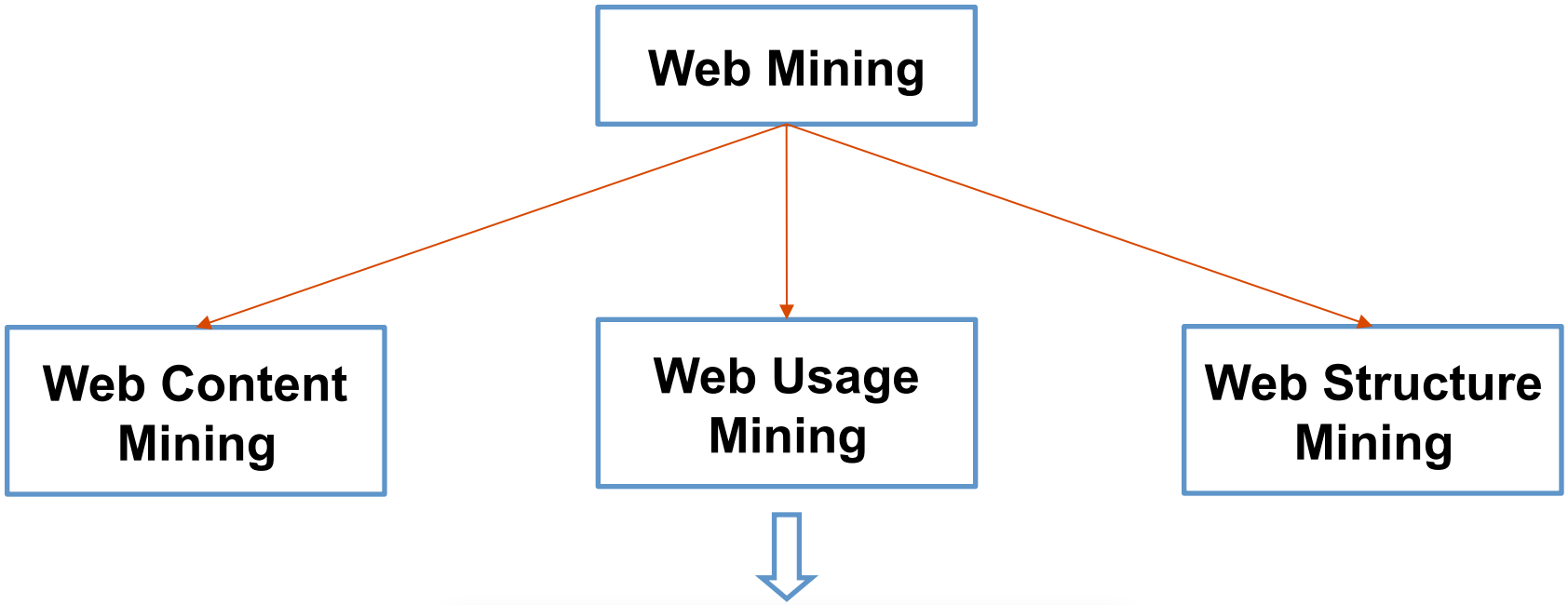
Types of Web Mining

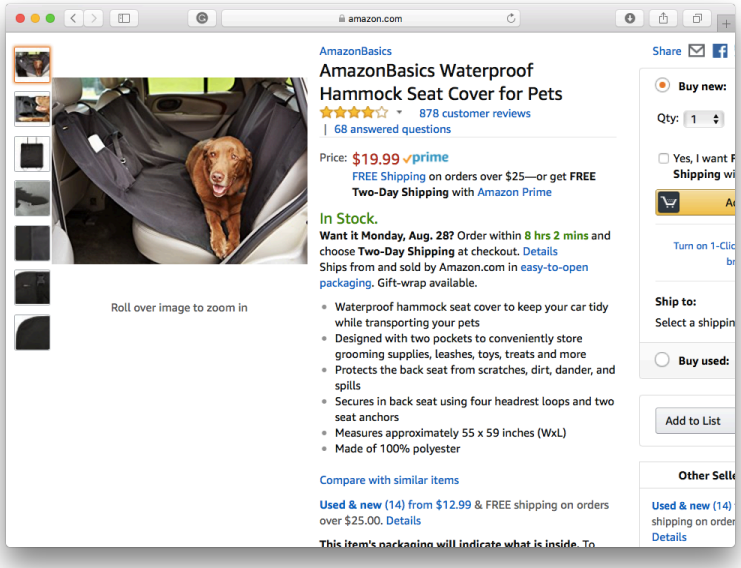


Types of Web Mining

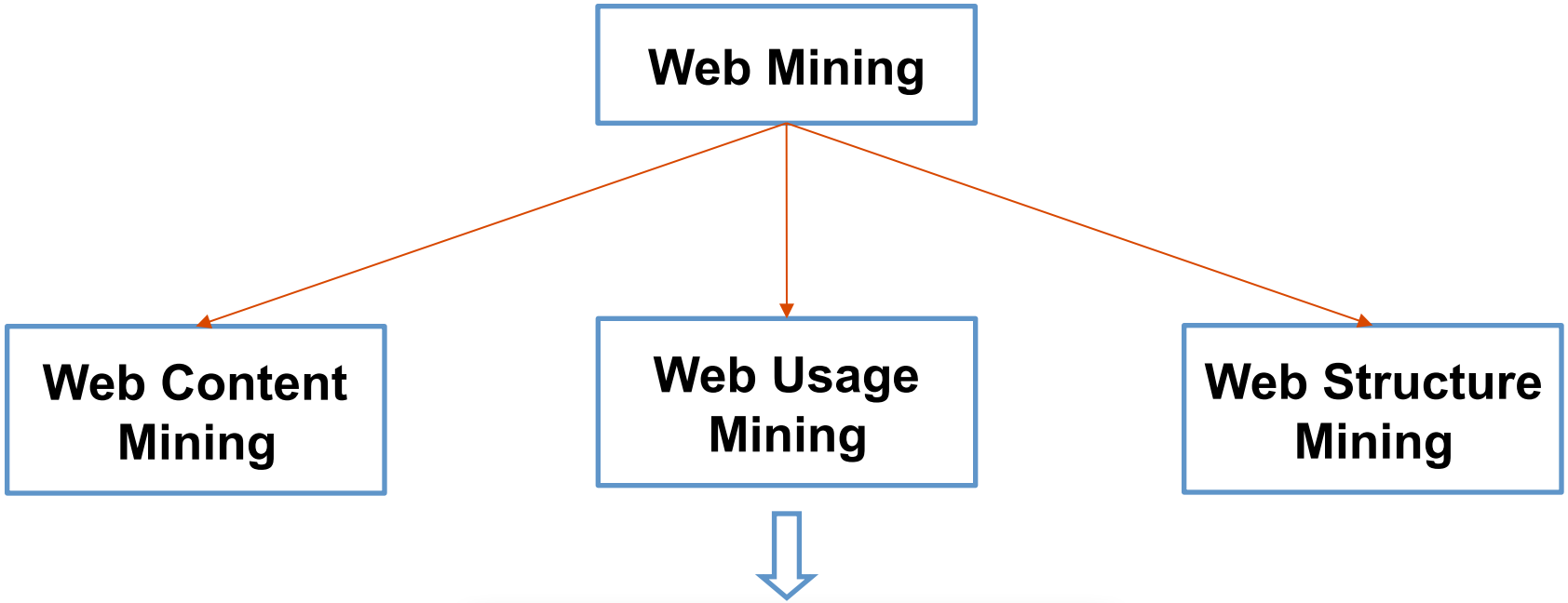


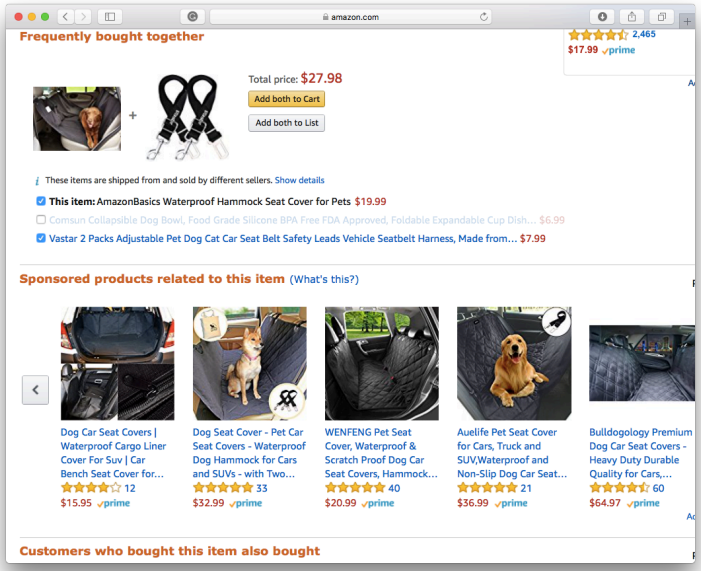
Types of Web Mining



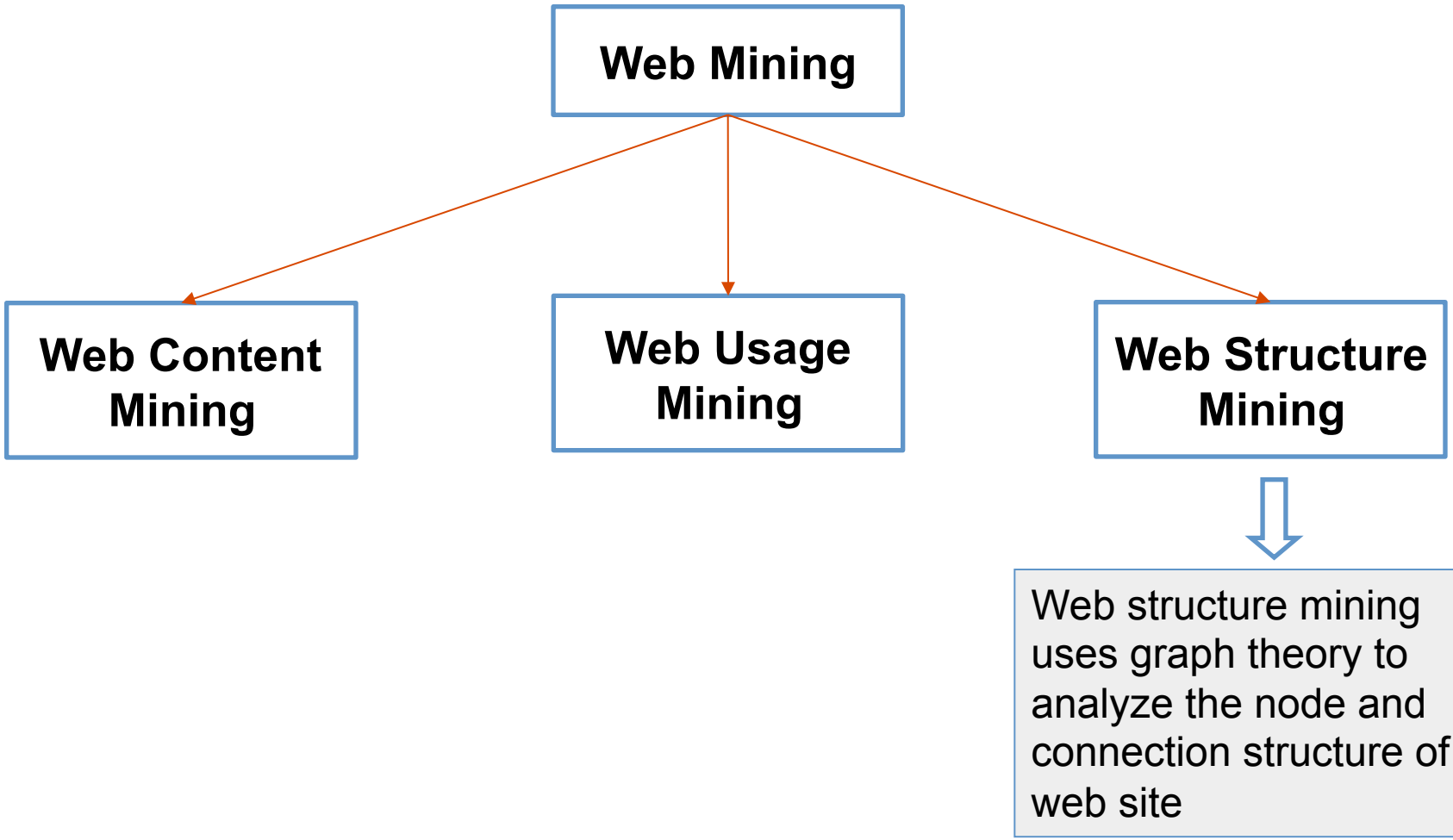


Types of Web Mining

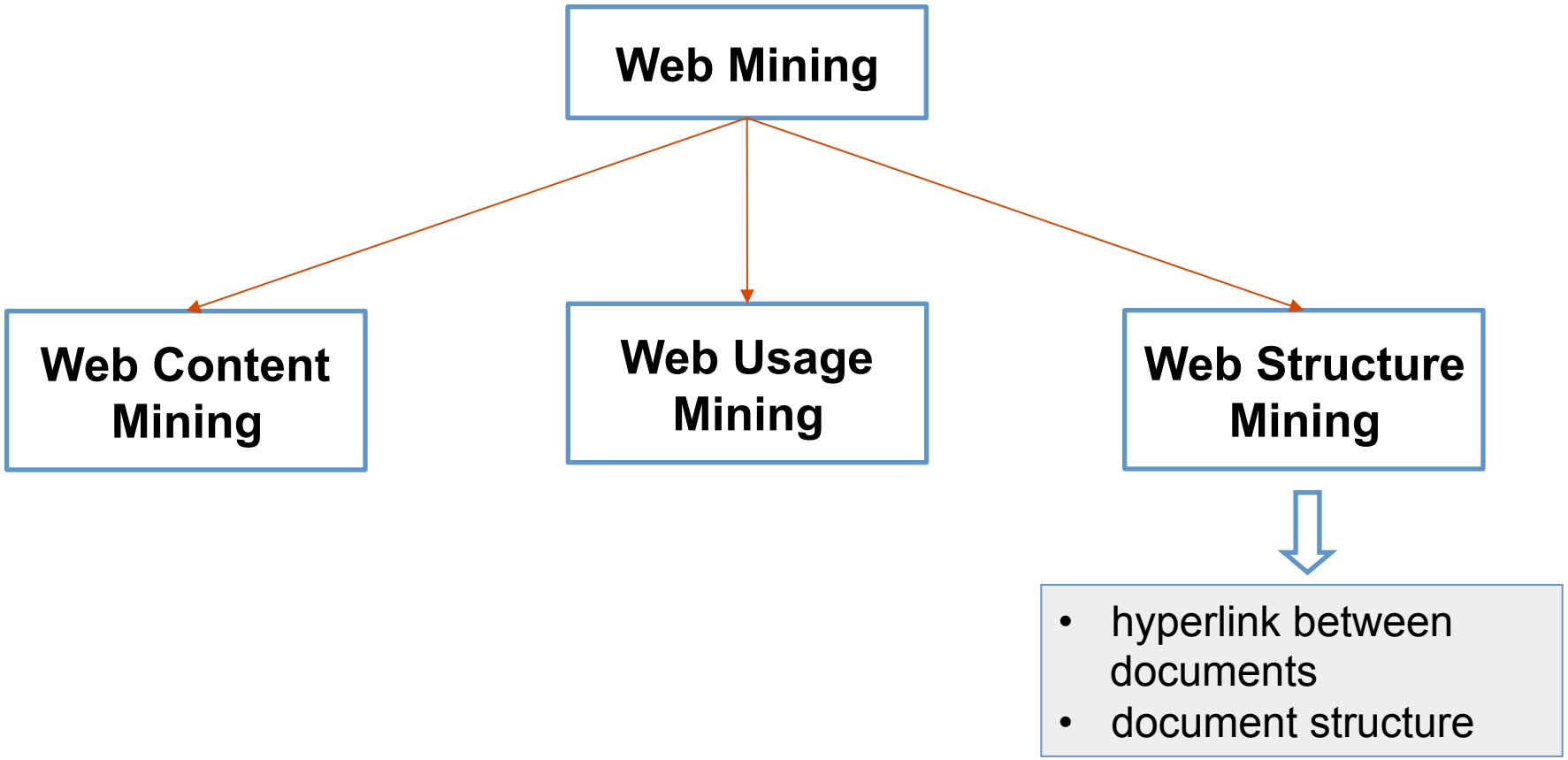




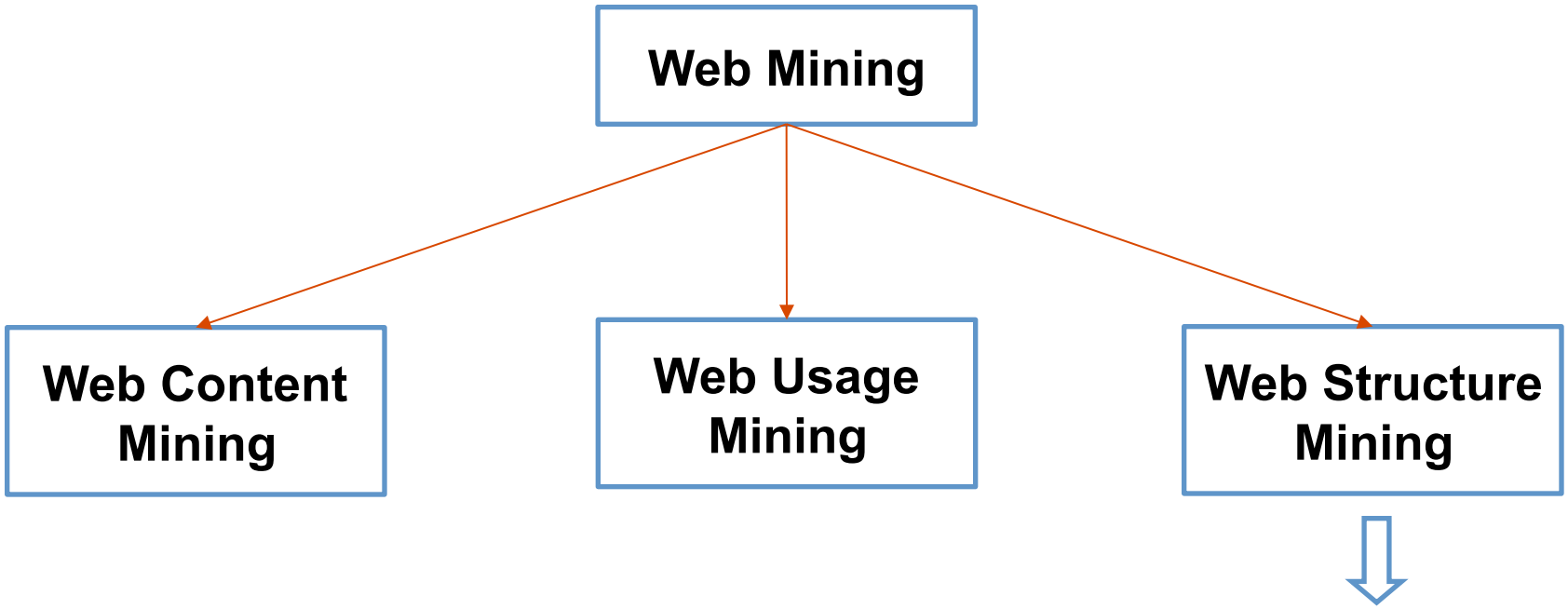
Types of Web Mining



Types of Web Mining



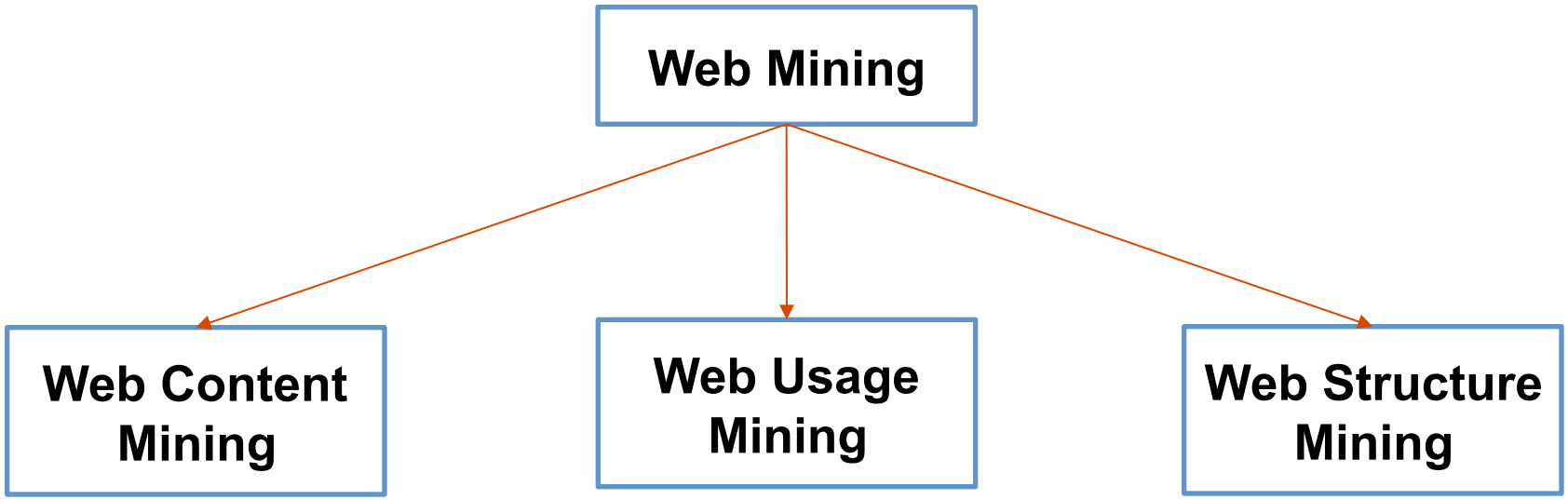
Types of Web Mining

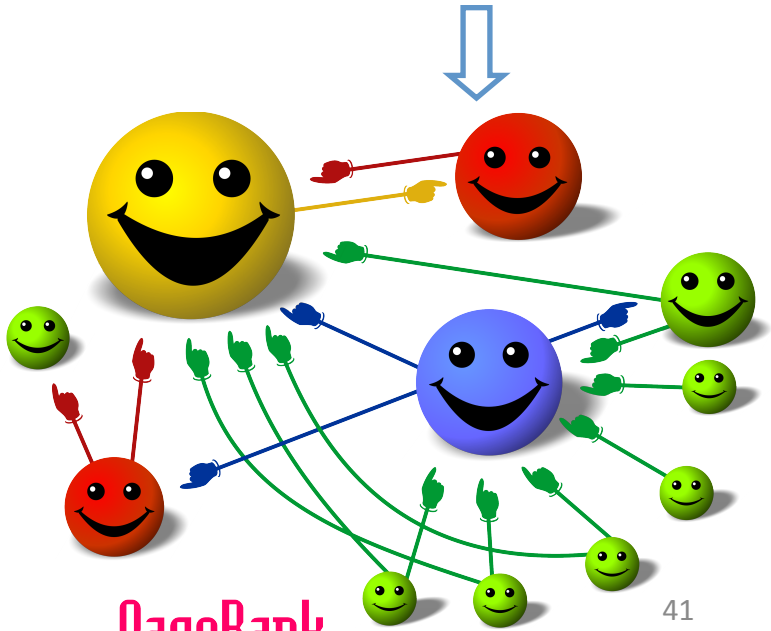


Applications:

- document retrieval and ranking (e.g., Google)
- discovery of "hubs" and "authorities"
- discovery of Web communities
- social network analysis

Types of Web Mining





Challenges

Web offers an unprecedented opportunity and challenge to data mining

- The amount of information on the Web is huge, and easily accessible.
- The coverage of Web information is very wide and diverse. One can find information about almost anything.
- Information/data of almost all types exist on the Web, e.g., structured tables, texts, multimedia data, etc.
- Much of the Web information is semi-structured due to the nested structure of HTML code.
- Much of the Web information is linked. There are hyperlinks among pages within a site, and across different sites.
- Much of the Web information is redundant. The same piece of information or its variants may appear in many pages.

Challenges

- The Web is noisy. A Web page typically contains a mixture of many kinds of information, e.g., main contents, advertisements, navigation panels, copyright notices, etc.
- The Web is also about services. Many Web sites and pages enable people to perform operations with input parameters, i.e., they provide services.
- The Web is dynamic. Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues.
- Above all, the Web is a virtual society. It is not only about data, information and services, but also about interactions among people, organizations and automatic systems, i.e., communities.

Getting data from the Web

Gathering Data

- Data to be visualized can come from a variety of sources and in a variety of formats
  - our own experiments/research
  - provided by others (libraries, data warehouses)
  - Web content

Gathering Data
Sources – Universities

- Berkeley Social Science Data Search
  - http://lib.berkeley.edu/wikis/datalab/Main/GoogleSearch
  - targets 800+ web sites that provide high-quality, downloadable data sets
  - For help, email joshua.quan@berkeley.edu

- Stanford Large Network Dataset Collection
  - http://snap.stanford.edu/data
  - social network data

- CMU's Data and Story Library
  - http://lib.stat.cmu.edu/DASL
  - data files and stories that illustrate use of basic statistics methods

- UCLA Stats Data Sets
  - http://www.stat.ucla.edu/data/
  - used in UCLA Stats Dept labs and assignments

Gathering Data
Sources - General Data Apps

- Amazon Public Data Sets
  - http://aws.amazon.com/
  - Large scientific datasets
- New York Times AP
  - http://developer.nytimes.com
  - grab results of queries in XML, JSON, ...
- Freebase
  - http://www.freebase.com
  - like Wikipedia for data
  - can download data or use API
- Stats about webpages over time
  - http://httparchive.org/
  - http://httparchive.org/downloads.php

Gathering Data
Sources – Government

- Census Bureau
  - http://www.census.gov
- Data.gov
  - http://data.gov
- Current Employment Statistics
  - http://bls.gov/ces/
- Data.gov.uk
  - http://data.gov.uk

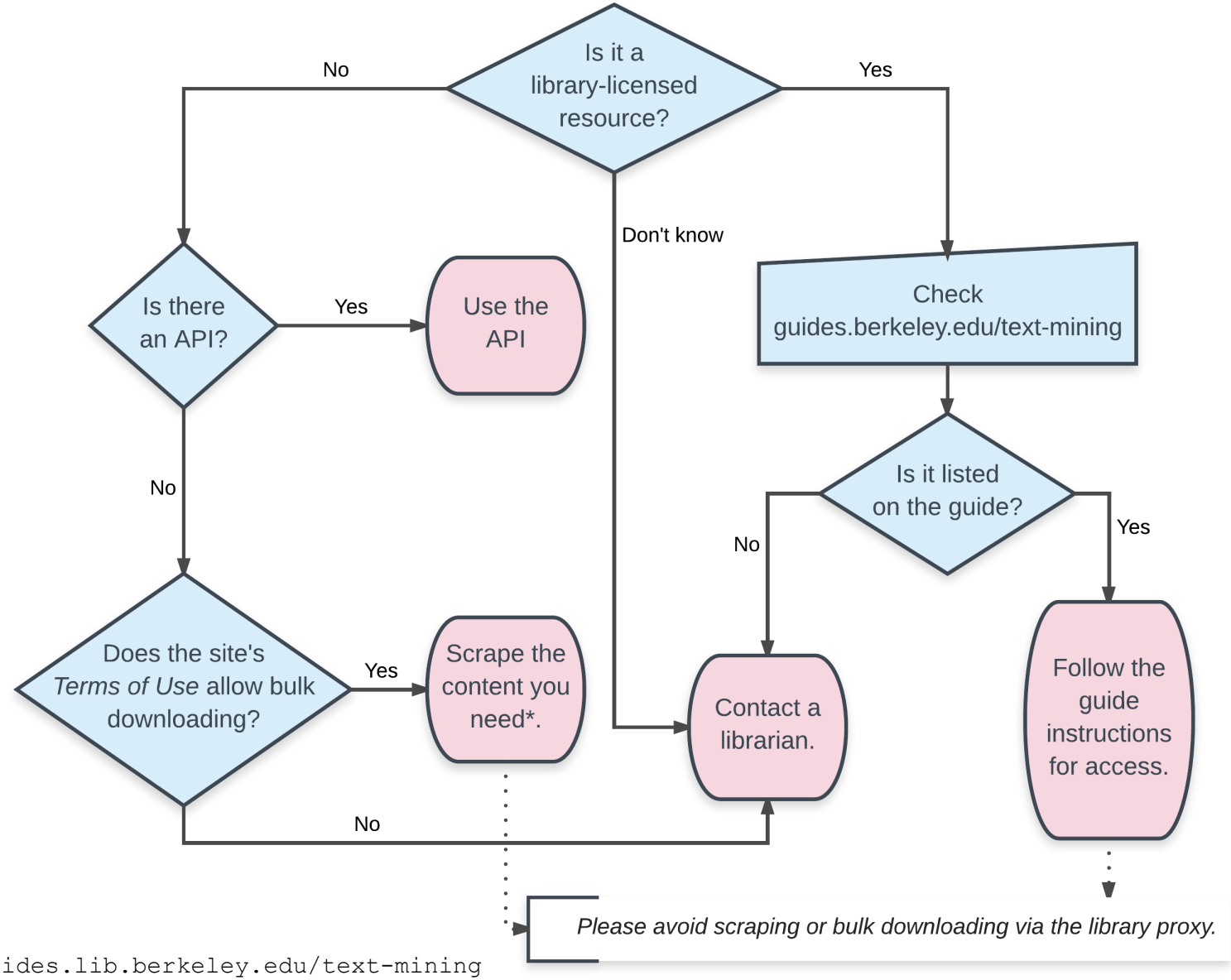
Check out *Visualize This*, Ch 2!

What if your data is online, but isn't given in a nice Excel/CSV file?

Get data from the Web

- Two kinds of ways to get data off the Web
  1. Web APIs - i.e. application-facing, for computers
  2. Web scraping - i.e. user-facing Websites for humans

UC Berkeley's Library access workflow



http://guides.lib.berkeley.edu/text-mining

Web APIs

- API stands for **A**pplication **P**rogramming **I**nterface
- a set of rules and procedures that facilitate interactions between computers and their applications

- A WebAPI = HTTP based API
  - allows users to query a remote database over the internet
  - majority adhere to a particular style known as Representational State Transfer or REST
  - "RESTful" APIs are convenient because we can use them to query databases using URLs

Benefits of APIs

- Regulate access to the data
  - Traceable accounts
  - Enable access to only a portion of the available data
    - By Time
    - By Location
- Avoid direct access to the platform website
- Avoid server congestion
- Provide paid access to full (or higher volume) data

Which Web APIs have you used?

Try this

- http://maps.googleapis.com/maps/api/geocode/json?address=Berkeley
- https://en.wikipedia.org/w/api.php?action=query&titles=Main%20Page&prop=revisions&rvprop=content&format=json
- https://www.googleapis.com/books/v1/volumes?q=isbn:0747532699
- https://itunes.apple.com/search?term=jack+johnson

JSON Parser: http://jsonparseronline.com/

More than 18.000 APIs available

Based on https://www.programmableweb.com/

How do GET Requests work?

- Surfing the Web = Making a bunch of GET Requests
- For instance, I open my web browser and type in http://www.wikipedia.org. Once I hit return, I'd see a webpage
- Several different processes occurred:

How do GET Requests work?

- Surfing the Web = Making a bunch of GET Requests
- For instance, I open my web browser and type in http://www.wikipedia.org. Once I hit return, I'd see a webpage
- Several different processes occurred:

```
curl -I "https://www.wikipedia.org"
HTTP/1.1 200 OK
Date: Tue, 05 Sep 2017 16:27:40 GMT
Content-Type: text/html
Connection: keep-alive
Server: mw1269.eqiad.wmnet
```

Steps of getting data from Web APIs

- Establish GET Request
  1. a base URL for the API
  2. (usually) some authorization code or key,
  3. a format for the response (e.g., .json)

- Get the response
  - Example: response = requests.get(http://climatedataapi.worldbank.org/climateweb/rest/v1/country/cru/tas/year/CAN.csv)

- Parse the response text
  - response.text[ ]

Demo

Assignment 2

Due Sept. 19th

Working on this assignment

- The easiest way is to log into the http://datahub.berkeley.edu. If you have a @berkeley.edu email address, you already have full access to the programming environment hosted on that site.

Submission details

- Once you have completed the notebook that is given to you, save and download it

- To download the notebook as a pdf, From within your notebook, click:
  - File >> Download as.. >> PDF via LaTeX (.pdf)

- Upload the PDF on bCourses

Next Lecture

- Intro to Web Scraping
- Systems and toolkits for visualizing data