

Week 2 - Web Data

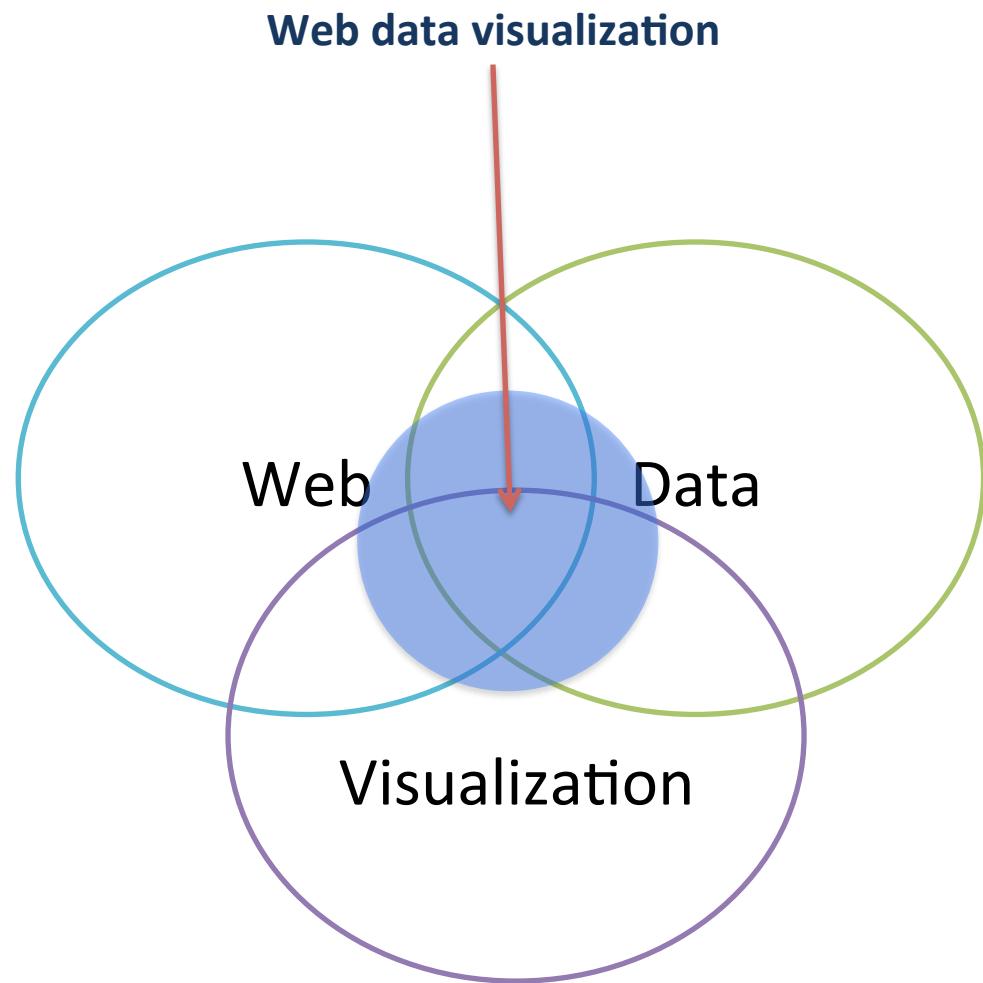
Yasmin AlNoamany

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

The image shows a web browser window with two video player interfaces side-by-side.

Left Video Player: A YouTube video player for "Jimmy Kimmel Reads Mean Comments from Trump Supporters". The video is at 0:02 / 2:19. The video frame shows Jimmy Kimmel on stage with a city skyline in the background. Subtitles at the bottom of the frame read: "THAT GOES. LAST NIGHT ON OUR SHOW, IF YOU". A "Suggested: August 15th, 2017" message is visible above the video. Below the video are standard YouTube controls (play, volume, etc.) and a progress bar. The video title is "Jimmy Kimmel Reads Mean Comments from Trump Supporters". The channel information shows "Jimmy Kimmel Live" with 9.7M subscribers. There are "Subscribe" and "Share" buttons. The video has 19,998 likes and 6,484 dislikes. A call-to-action overlay for "YouTube TV for free" is overlaid on the bottom right of the video frame.

Right Video Player: A DataCamp video player for "Learn R from the best Instructors". The video features Hadley Wickham, RStudio, smiling. The video title is "Jimmy Kimmel's Plan to Save Us from Trump" by Jimmy Kimmel Live. It has 4,410,101 views. Below it is another video thumbnail for "Celebrities Read Mean Tweets #10" by Jimmy Kimmel Live with 28,201,893 views. At the bottom, there is a thumbnail for "Donald Trump Phoenix Rally Cold Open - SNL" by 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 By using this site, you agree to the [Privacy Policy](#) and [Terms of Service](#).

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

Web content

flickr.com

Back to album You Explore Create Photos, people, or groups

TPDL 2015

These are shots from my trip to TPDL2015 in Poznan,

Show more

52 photos • 42 views

By: Yasmin Alnoamany

9

The main image shows two bronze statues of horses, one rearing and one standing behind it, both attached to chains. The background is a wire fence and foliage. Below the main image are three smaller thumbnail images showing night scenes of buildings with illuminated facades.

Web content

The screenshot shows a web browser window for soundcloud.com. The interface includes a navigation bar with Home, Collection, Search, Upgrade, Upload, and a user profile. Below the navigation is a section titled "Stream" with tabs for Charts and Discover. A message encourages users to hear the latest posts from their followed accounts. Two main posts are displayed:

- Under Armour** promoted a track titled "Lorine Chia UNLIKE ANY Commentary". The track art features a woman in a black jacket over a pink top against a blue sky with clouds. The duration is 4:11. Below the track are interaction icons for heart, repost, and more, along with statistics: 12.4K and 4 comments.
- Ahmed Haz** reposted a playlist 8 days ago. The post includes a small thumbnail image and the name "Adzeerv".

On the right side of the screen, there is a promotional banner for SoundCloud Go, which reads: "SOUNDCLOUD GO THE NEXT BIG ARTIST IS ALREADY STREAMING HERE GET SOUNDCLOUD G". Below the banner, there are sections for "Who to follow" and profiles for "Under Armour" and "Rahma Magdy.A". At the bottom, a media player shows a track by "Grace Davies" titled "Hello - Adele" at 0:09 of a 5:00 minute duration.

Web Structure

en.wikipedia.org Not logged in Talk Contributions Create account Log in Article Talk Read Edit View history Search Wikipedia

Game of Thrones

From Wikipedia, the free encyclopedia

This article is about the television series. For the novel in the series *A Song of Ice and Fire*, see *A Game of Thrones*. For other uses, see *A Game of Thrones* (disambiguation).

Game of Thrones is an American fantasy drama television series created by David Benioff and D. B. Weiss. It is an adaptation of *A Song of Ice and Fire*, George R. R. Martin's series of fantasy novels, the first of which is *A Game of Thrones*. It is filmed in Belfast and elsewhere in the United Kingdom, Canada, Croatia, Iceland, Malta, Morocco, Spain, and the United States. The series premiered on HBO in the United States on April 17, 2011, and its sixth season concluded on June 26, 2016. The series was renewed for a seventh season,^[1] which premiered on July 16, 2017,^[2] and will conclude with its eighth season in 2018 or 2019.^[3]

Set on the fictional continents of Westeros and Essos, *Game of Thrones* has several plot lines and a large ensemble cast but centers on three primary story arcs. The first story arc centers on the Iron Throne of the Seven Kingdoms and follows a web of alliances and conflicts among the dynastic noble families either vying to claim the throne or fighting for independence from the throne. The second story arc focuses on the last descendant of the realm's deposed ruling dynasty, exiled and in hiding while plotting a return to the throne. The

Game of Thrones

Genre Fantasy Serial drama
Created by David Benioff D. B. Weiss
Based on *A Song of Ice and Fire* by George R. R. Martin
Starring see List of *Game of Thrones* characters
Theme music composer Ramin Djawadi



Graph Theory: Pages are nodes & links are directed edges

homeboxoffice.com About HBO | C

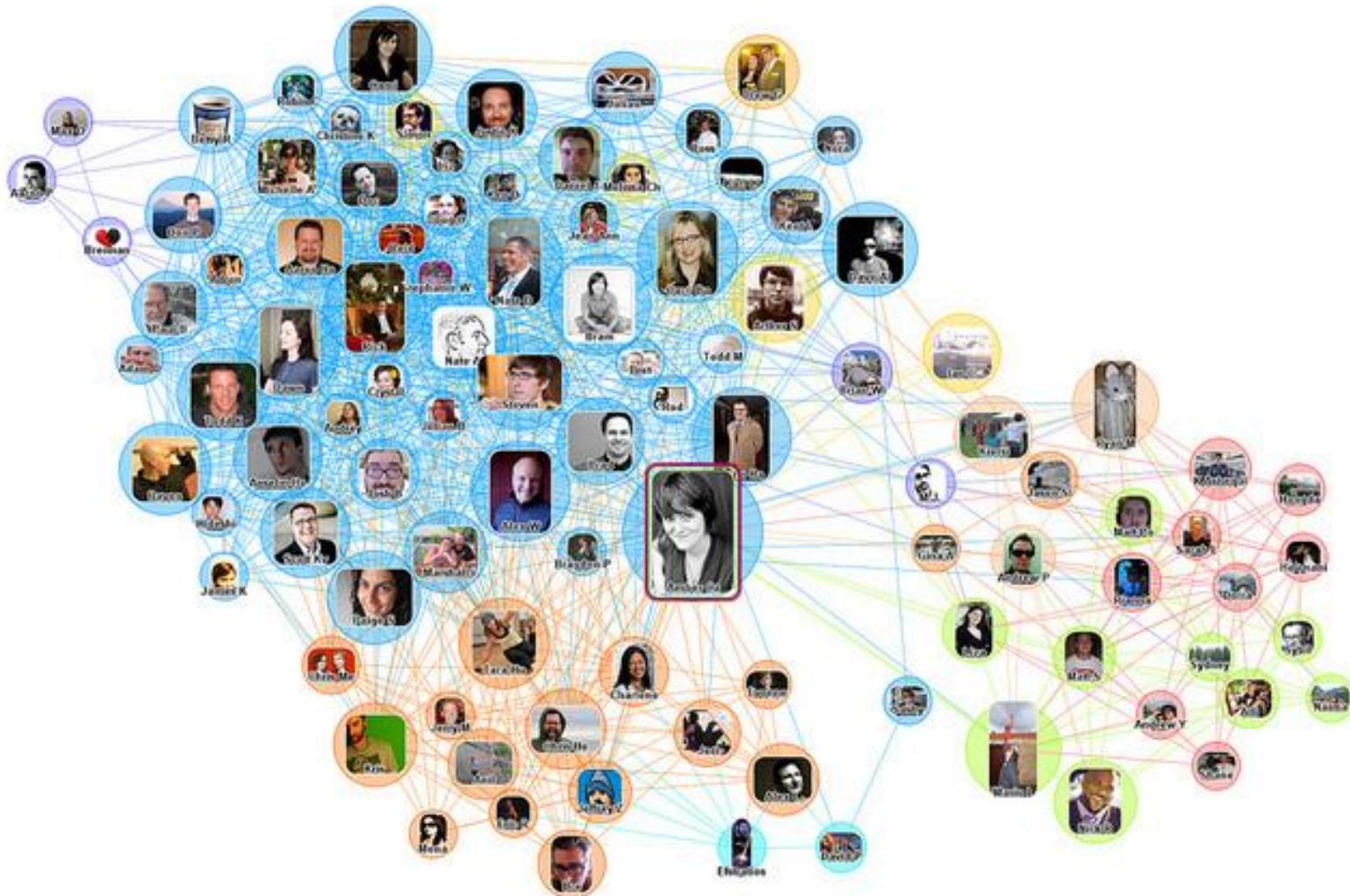
HBO HOME BOX OFFICE **CINEMAX**

Corporate Site

Affiliate Sites / B2B B2B PMD Media Relations International Press FTP Services Tech Ops HOD Reporting

ADDICTIVE ORIGINAL SERIES
GAME OF THRONES Sundays at 9PM on HBO*

Web Structure



Web usage

http://www.cnn.com

The screenshot shows the CNN homepage with a prominent orange "NEWS ALERT" banner at the top stating, "There are fears of catastrophic flooding as Harvey moves inland. Watch CNN". Below the banner, the CNN logo and "Home" link are visible, along with "Live TV" and "U.S. Edition" buttons. The main headline is "Widespread damage after Harvey hits Texas", accompanied by a large image of a damaged building and a "BREAKING NEWS" box. A sidebar on the left lists "Latest" news items about Harvey, while the right sidebar features "Top stories" about North Korea launching missiles. At the bottom, there's a "LIVE UPDATES" section, a "WATCH LIVE" video player showing a portrait of Donald Trump, and a legal disclaimer about privacy and terms of service.

```
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

The screenshot shows the CNN homepage with a prominent orange news alert banner at the top stating "NEWS ALERT: There are fears of catastrophic flooding as Harvey moves inland. Watch CNN". Below the banner, the CNN logo and "Home" link are visible, along with "Live TV" and "U.S. Edition" buttons. The main headline reads "Widespread damage after Harvey hits Texas". To the left of the headline is a large image of a damaged building with debris scattered around. A red "BREAKING NEWS" box is overlaid on the image. Below the image, a sub-headline states "Debris is scattered across a wide swath of Southeast Texas after the Category 4 hurricane roared ashore". Underneath this, there are "LIVE UPDATES" and "WATCH LIVE" sections. On the right side, there is a sidebar titled "Latest" which lists recent stories about Hurricane Harvey. Below the latest section is a "Top stories" section featuring a photo of Kim Jong-un and a headline about North Korea launching missiles. At the bottom of the page, there is a legal disclaimer about privacy and terms of service.

```
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

The screenshot shows the CNN homepage with a prominent orange news alert banner at the top stating "NEWS ALERT: There are fears of catastrophic flooding as Harvey moves inland. Watch CNN". Below the banner, the CNN logo and "Home" link are visible, along with "Live TV" and "U.S. Edition" buttons. The main headline reads "Widespread damage after Harvey hits Texas". To the left of the headline is a large image of a damaged building with debris scattered around. A red "BREAKING NEWS" box is overlaid on the image. Below the image, a sub-headline states "Debris is scattered across a wide swath of Southeast Texas after the Category 4 hurricane roared ashore". A "LIVE UPDATES" section indicates "Harvey now a tropical storm. Heavy rain continues". A "WATCH LIVE" section shows a video thumbnail of Donald Trump. At the bottom of the main content area, there is a legal notice: "By using this site, you agree to the [Privacy Policy](#) and [Terms of Service](#).
This says" followed by a close button.

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

```
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

The screenshot shows the CNN homepage with a prominent orange news alert banner at the top stating "NEWS ALERT: There are fears of catastrophic flooding as Harvey moves inland. Watch CNN". Below the banner, the CNN logo and "Home" link are visible, along with "Live TV" and "U.S. Edition" buttons. The main headline is "Widespread damage after Harvey hits Texas", accompanied by a large image of a damaged building and a "BREAKING NEWS" badge. A sidebar on the left lists the latest news stories about Hurricane Harvey, including: "Texas governor: Flooding is our top priority right now", "Harvey has been downgraded to a tropical storm", "Rockport resident: Harvey 'sounded like a freight train with square wheels'", and "Soon: Texas Governor gives hurricane update". To the right, a "Top stories" section features a photo of Kim Jong-un and the headline "North Korea launches missiles days after US praised restraint". Other stories in the sidebar include: "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", and "Opinion: Trump, the imperial president". At the bottom, a legal notice states: "By using this site, you agree to the [Privacy Policy](#) and [Terms of Service](#).
This says" followed by a close button.

```
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>

The screenshot shows the CNN homepage with a red banner at the top reading "NEWS ALERT: There are fears of catastrophic flooding as Harvey moves inland. Watch CNN". Below the banner, the CNN logo is on the left, followed by "Home", "Live TV", "U.S. Edition +", and a menu icon. The main headline is "Widespread damage after Harvey hits Texas" with a sub-image of a destroyed building and a person walking through debris. A "BREAKING NEWS" box is visible. To the right, there's a "Latest" section with several news items and a "Top stories" section featuring a photo of Kim Jong-un and Ri Yong-ho.

Latest

- 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

'This says By using this site, you agree to the [Privacy Policy](#) and [Terms of Service](#). X

```
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

The screenshot shows the CNN homepage with a red banner at the top reading "NEWS ALERT: There are fears of catastrophic flooding as Harvey moves inland. Watch CNN". Below the banner, the CNN logo is on the left, followed by "Home", "Live TV", "U.S. Edition +", and a menu icon. The main headline is "Widespread damage after Harvey hits Texas". To the left of the headline is a large image of a damaged building with debris scattered around. A red "BREAKING NEWS" box is overlaid on the image. Below the image is a sub-headline: "Debris is scattered across a wide swath of Southeast Texas after the Category 4 hurricane roared ashore". Underneath this are "LIVE UPDATES" and "WATCH LIVE" sections. On the right side, there's a sidebar titled "Latest" with several news items and a "Top stories" section featuring a photo of Kim Jong-un and Ri Yong-ho.

LATEST

- 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

'This says By using this site, you agree to the [Privacy Policy](#) and [Terms of Service](#). X

```
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

The screenshot shows the CNN homepage with a red banner at the top reading "NEWS ALERT: There are fears of catastrophic flooding as Harvey moves inland. Watch CNN". Below the banner, the CNN logo is on the left, followed by "Home", "Live TV", "U.S. Edition +", and a menu icon. The main headline is "Widespread damage after Harvey hits Texas". To the left of the headline is a large image of a damaged building with debris scattered around. A red "BREAKING NEWS" box is overlaid on the image. Below the image is a sub-headline: "Debris is scattered across a wide swath of Southeast Texas after the Category 4 hurricane roared ashore". Underneath this are "LIVE UPDATES" and "WATCH LIVE" sections. On the right side, there's a sidebar titled "Latest" with several news items and a "Top stories" section featuring a photo of Kim Jong-un and Ri Yong-ho.

LATEST

- 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

This says By using this site, you agree to the [Privacy Policy](#) and [Terms of Service](#).

```
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"
```

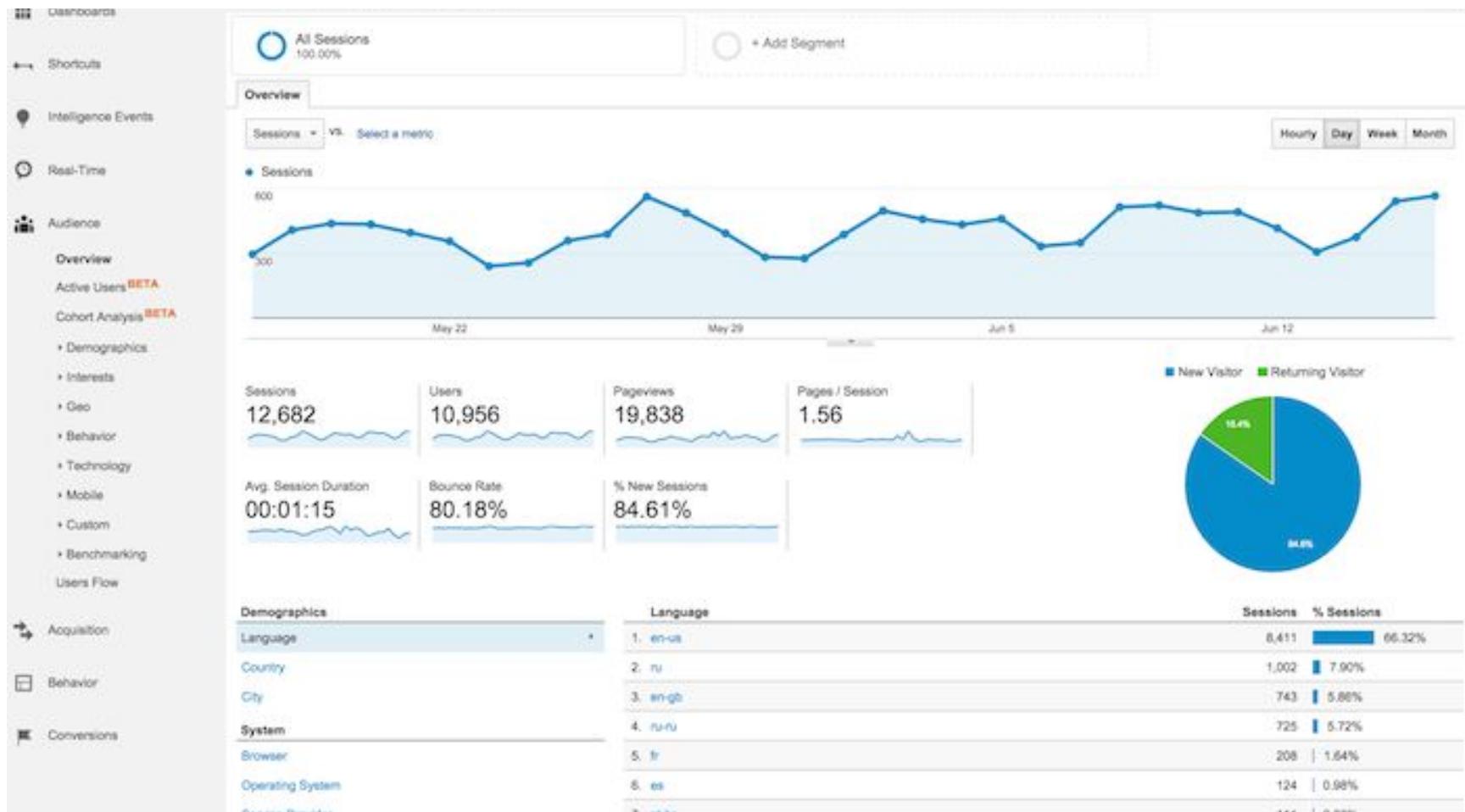
Web usage

http://www.cnn.com

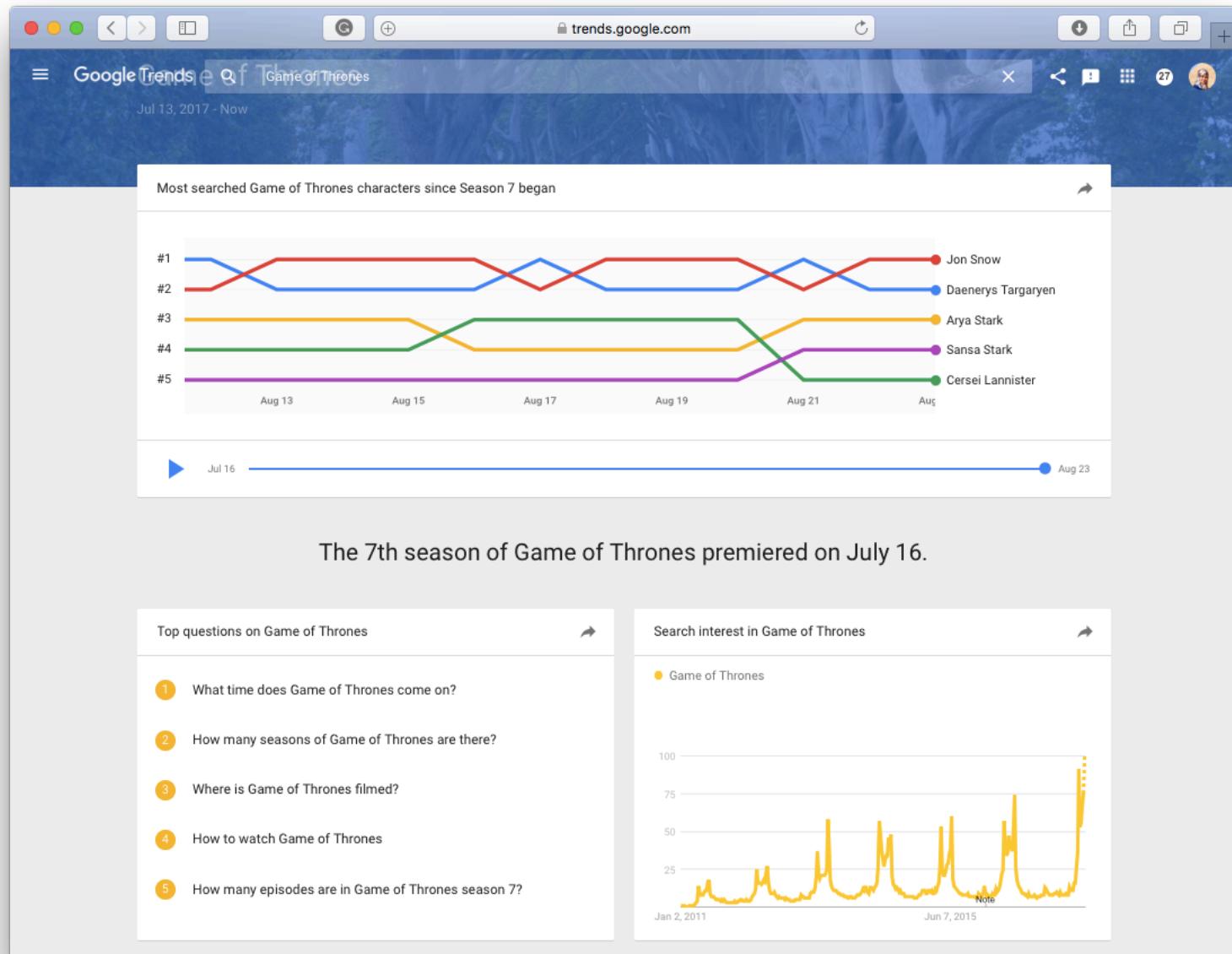
The screenshot shows the CNN homepage with a prominent orange banner at the top reading "NEWS ALERT" and "There are fears of catastrophic flooding as Harvey moves inland. Watch CNN". Below the banner, the CNN logo and "Home" link are visible, along with "Live TV" and "U.S. Edition" buttons. The main headline is "Widespread damage after Harvey hits Texas", accompanied by a photograph of a damaged building and a "BREAKING NEWS" badge. A sidebar on the left lists "Latest" news items: "Texas governor: Flooding is our top priority right now" (pink dot), "Harvey has been downgraded to a tropical storm" (grey dot, 38 min ago), "Rockport resident: Harvey 'sounded like a freight train with square wheels'" (grey dot, 1 hr ago), and "Soon: Texas Governor gives hurricane update" (grey dot, 1 hr ago). To the right, a "Top stories" section features a photograph of Kim Jong-un and Ri Yong-ho with the headline "North Korea launches missiles days after US praised restraint". Other stories in the sidebar include "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", and "Opinion: Trump, the imperial president". At the bottom, a message says "By using this site, you agree to the [Privacy Policy](#) and [Terms of Service](#).
This says" followed by a close button.

```
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"
```

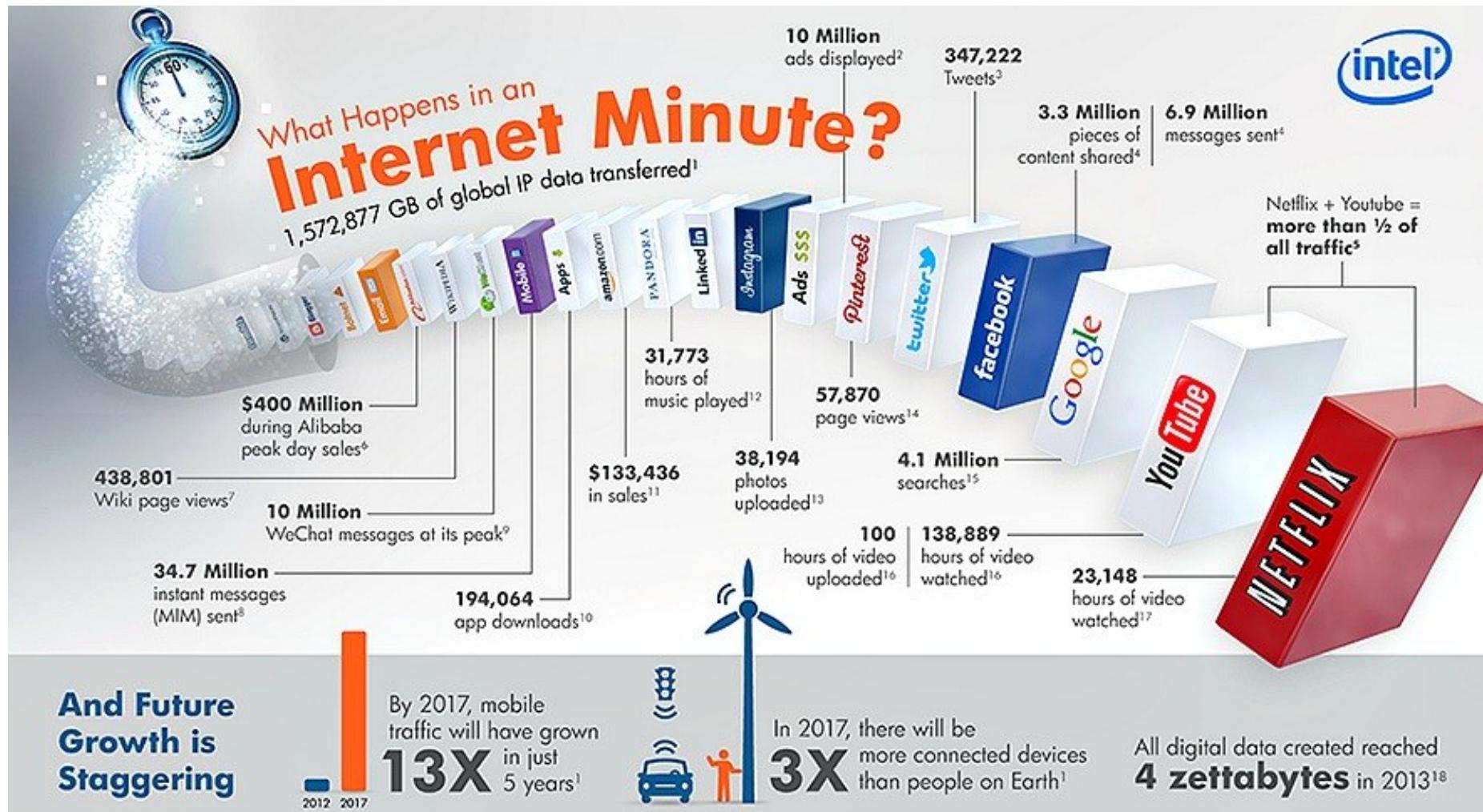
Web usage



Using the Web to understand the Web



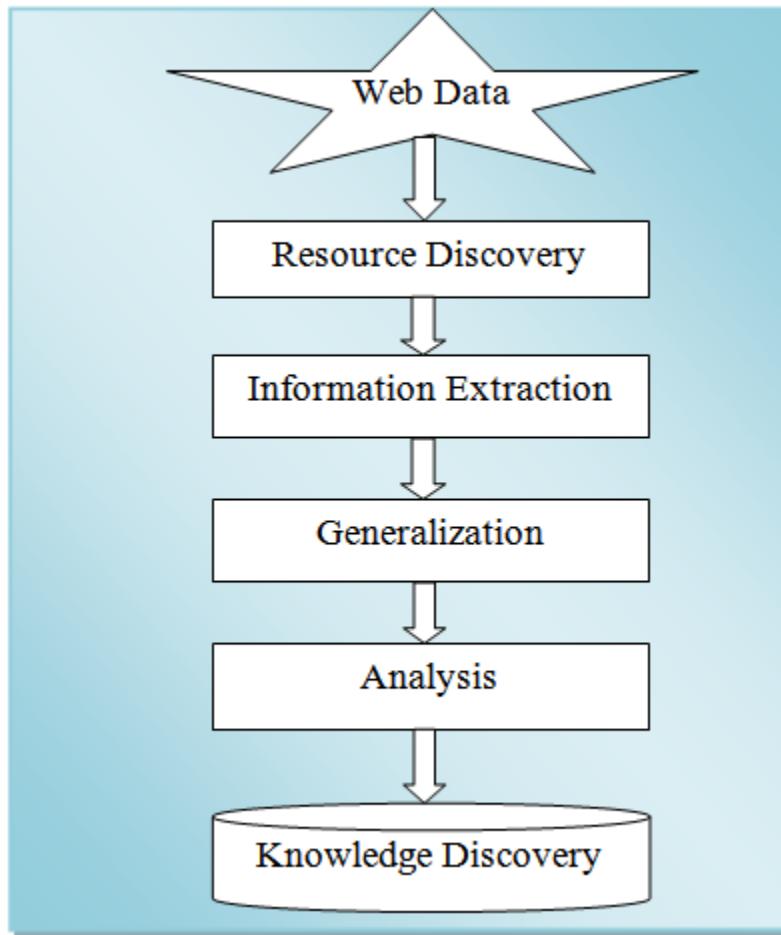
What happens in the Internet minute



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

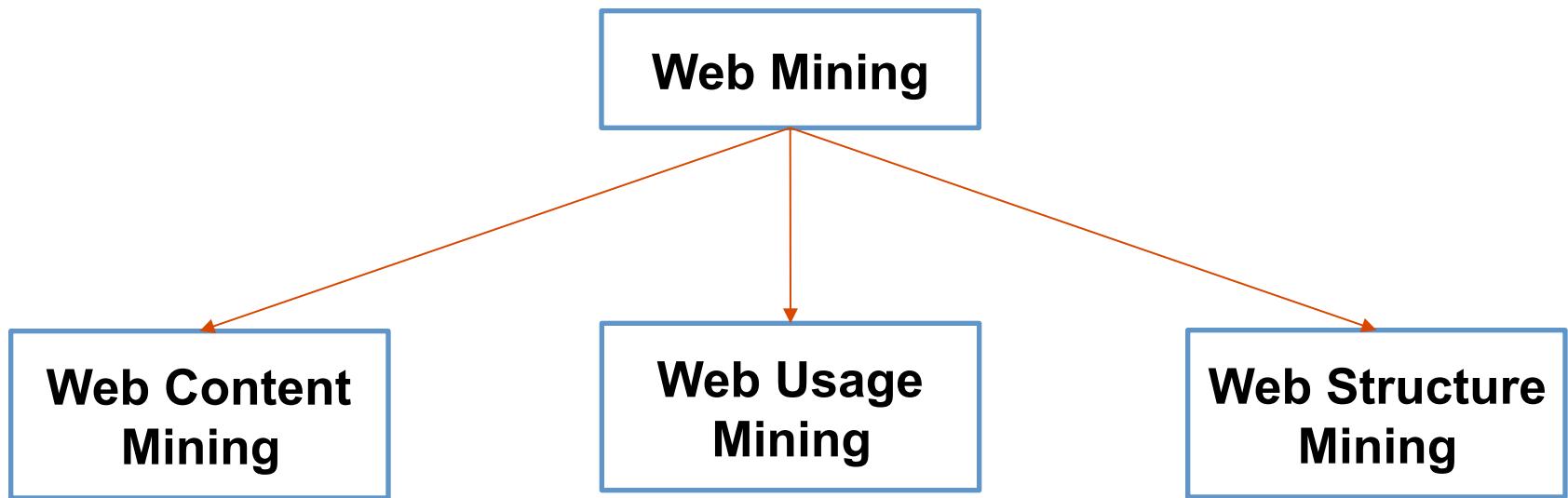
Decomposition of Web mining



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

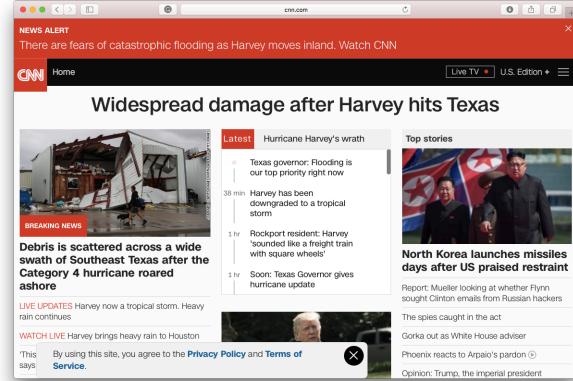
Types of Web Mining



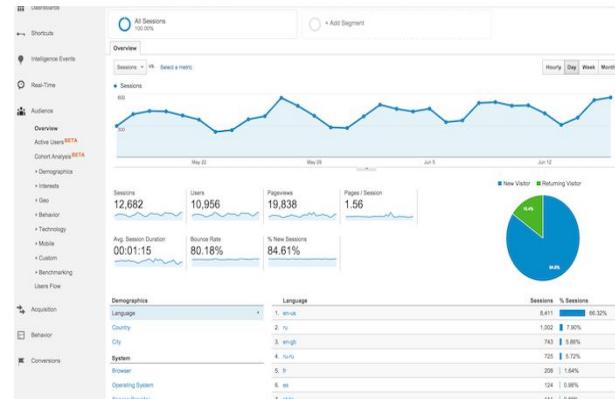
Types of Web Mining

Web Mining

Web Content Mining



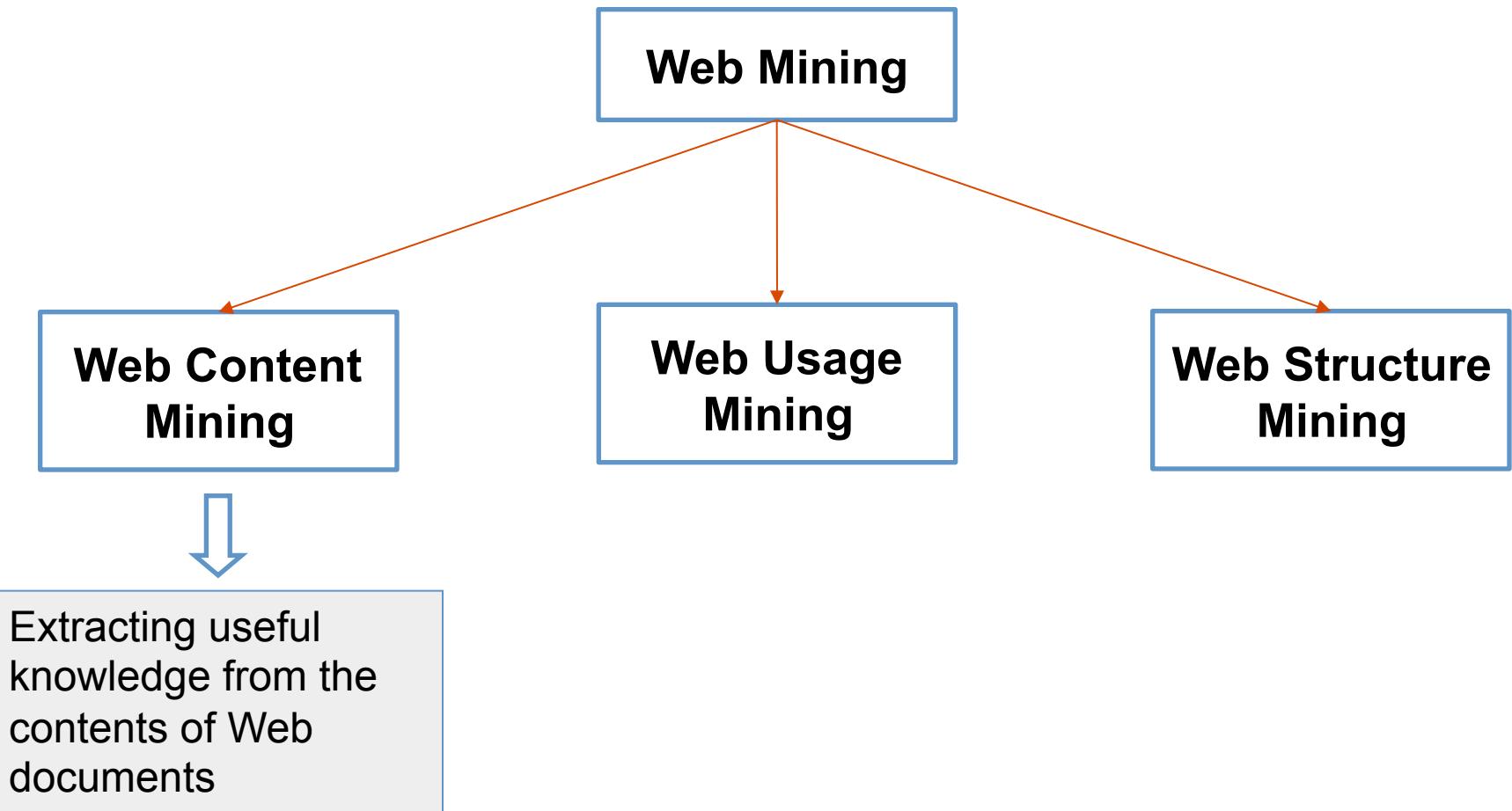
Web Usage Mining



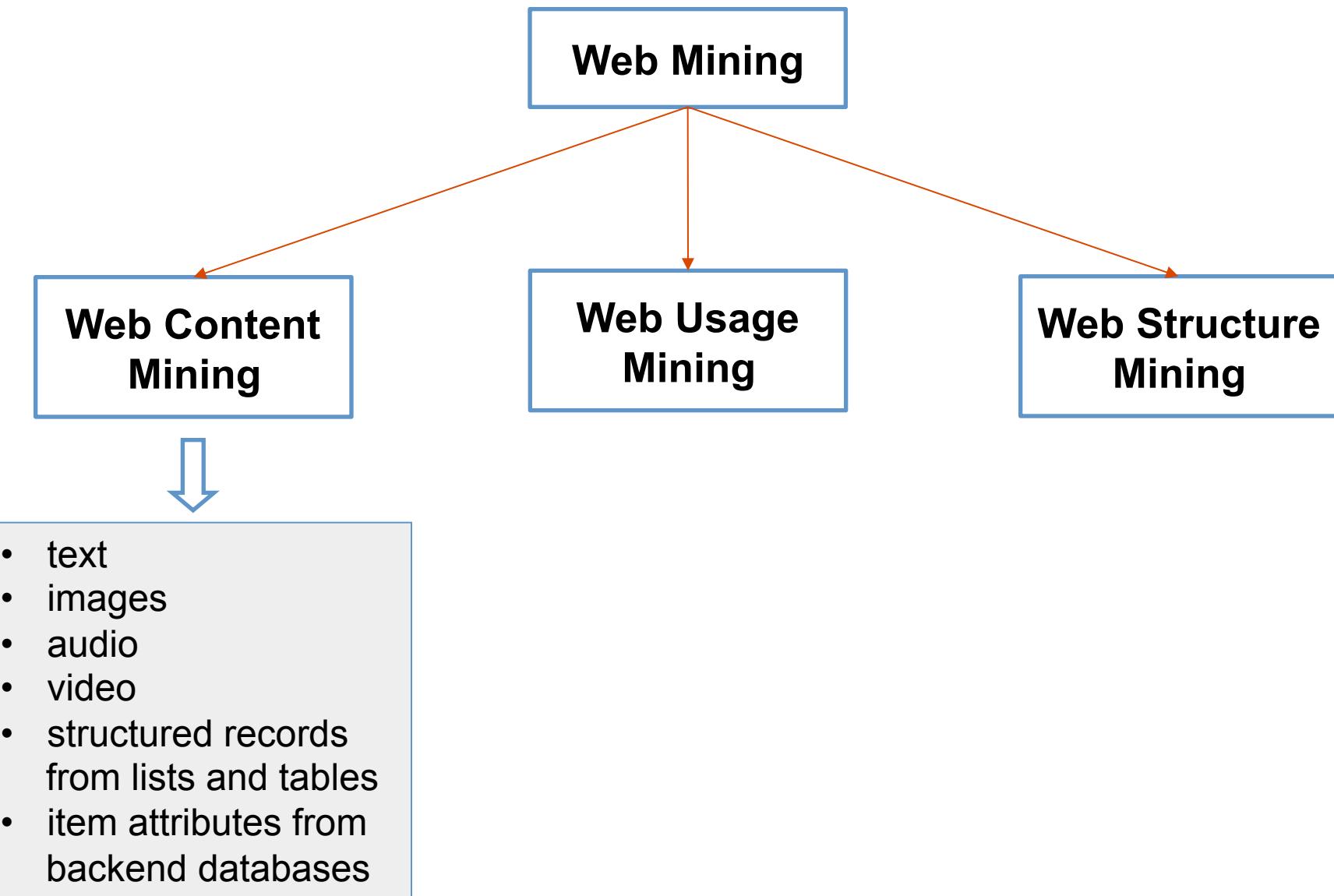
Web Structure Mining



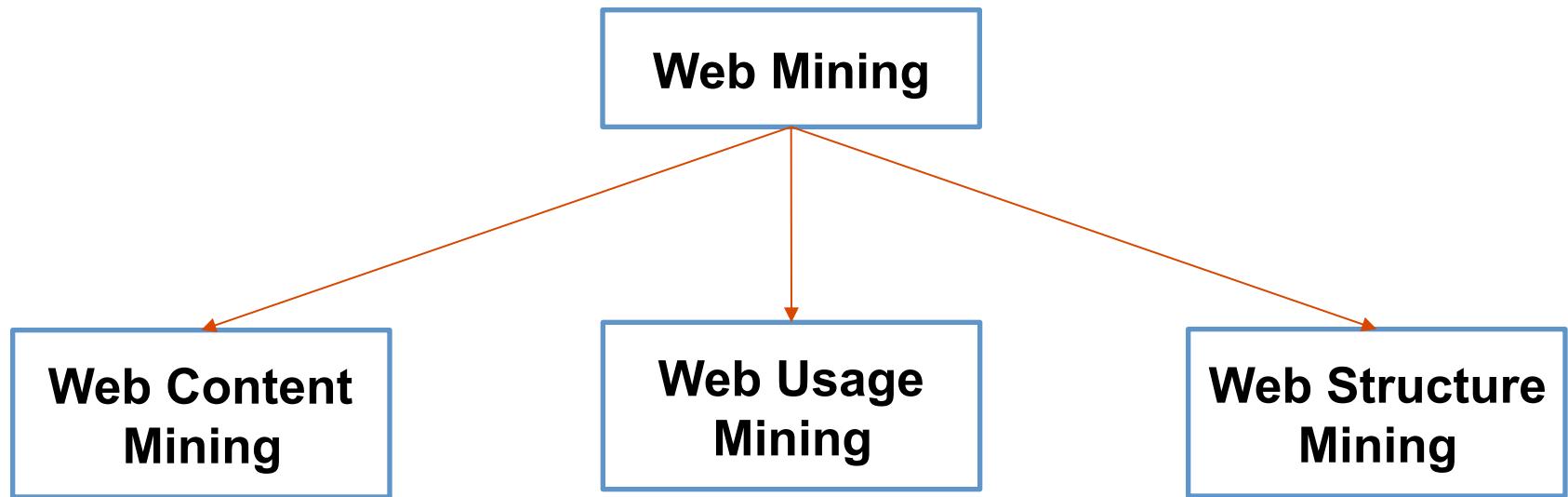
Types of Web Mining



Types of Web Mining



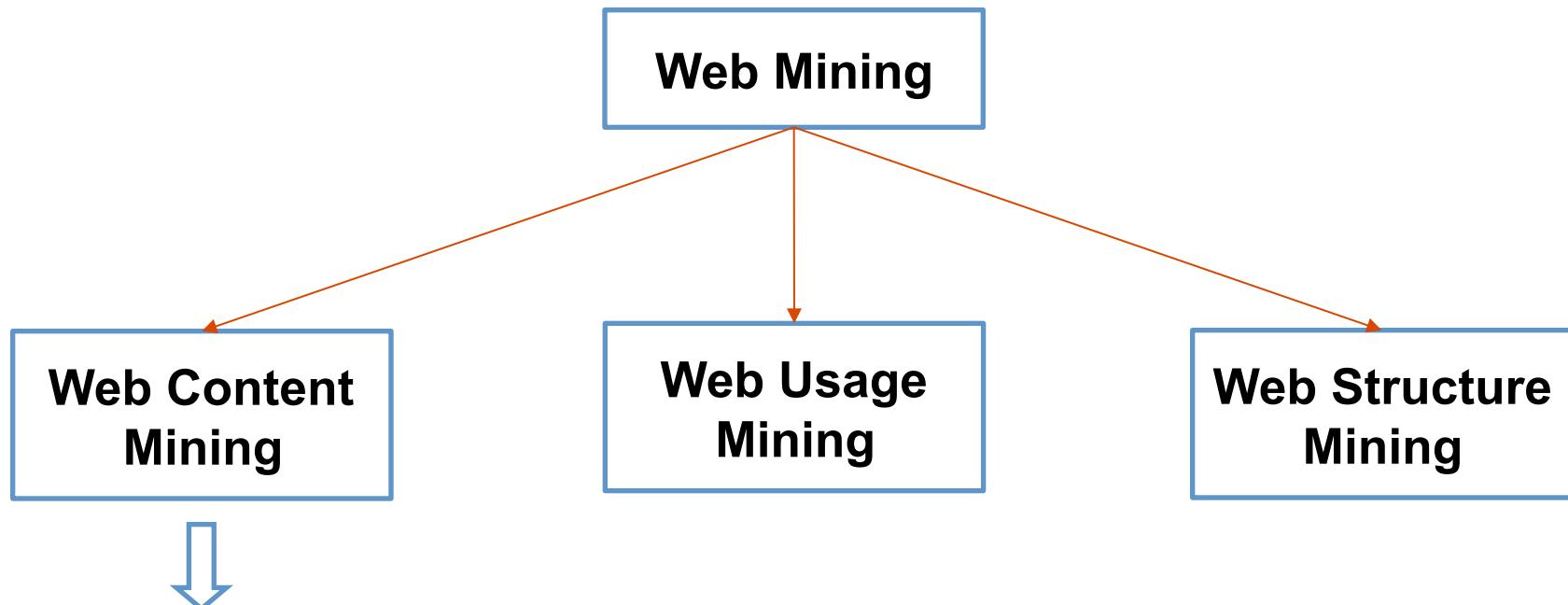
Types of Web Mining



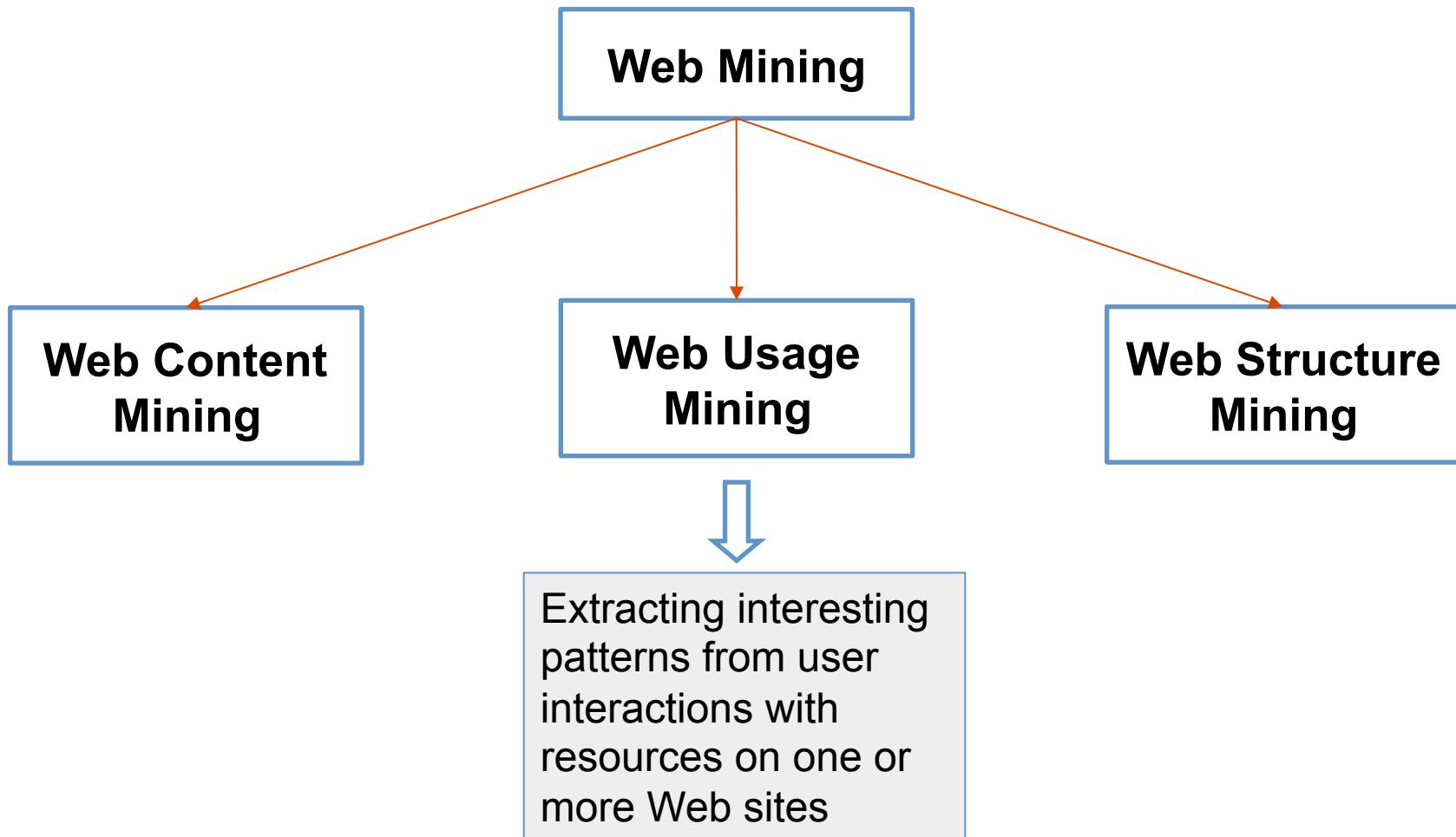
Applications:

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

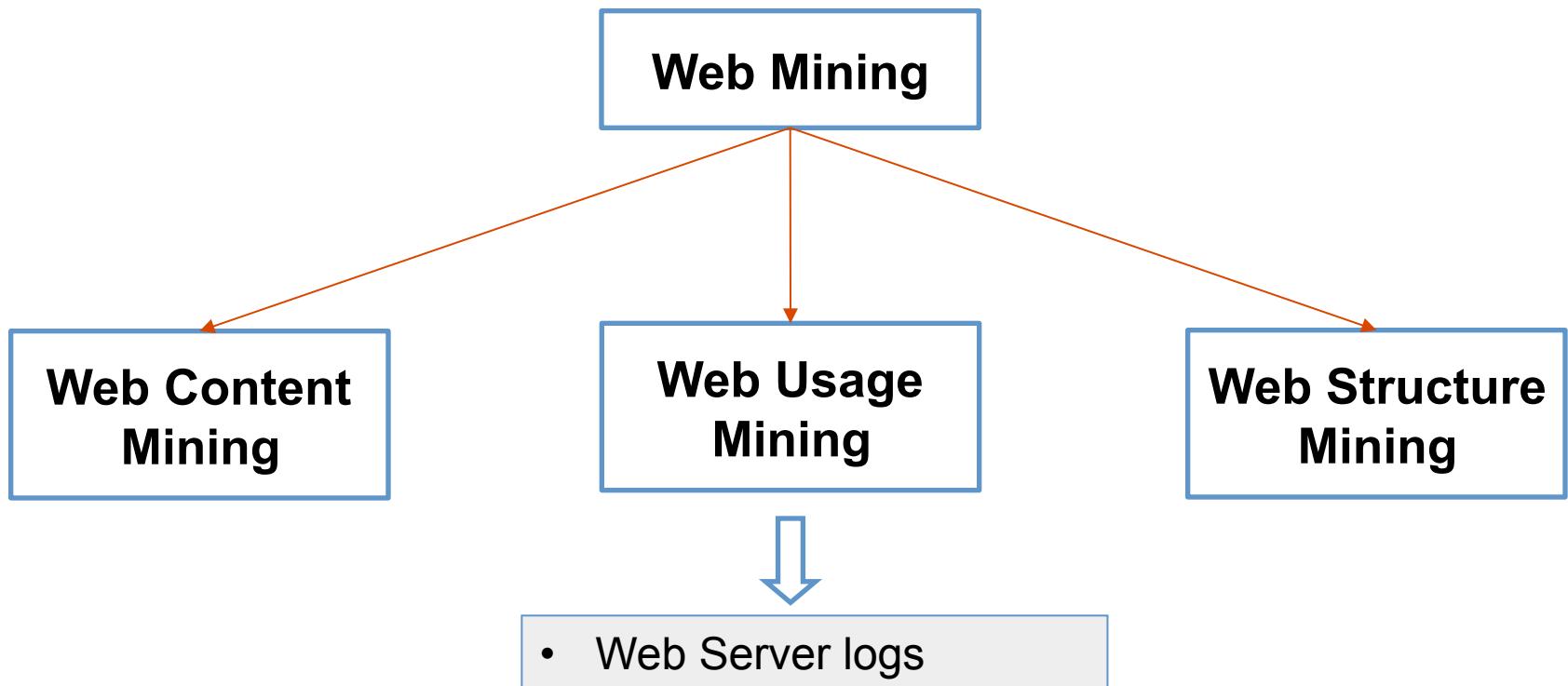
Types of Web Mining



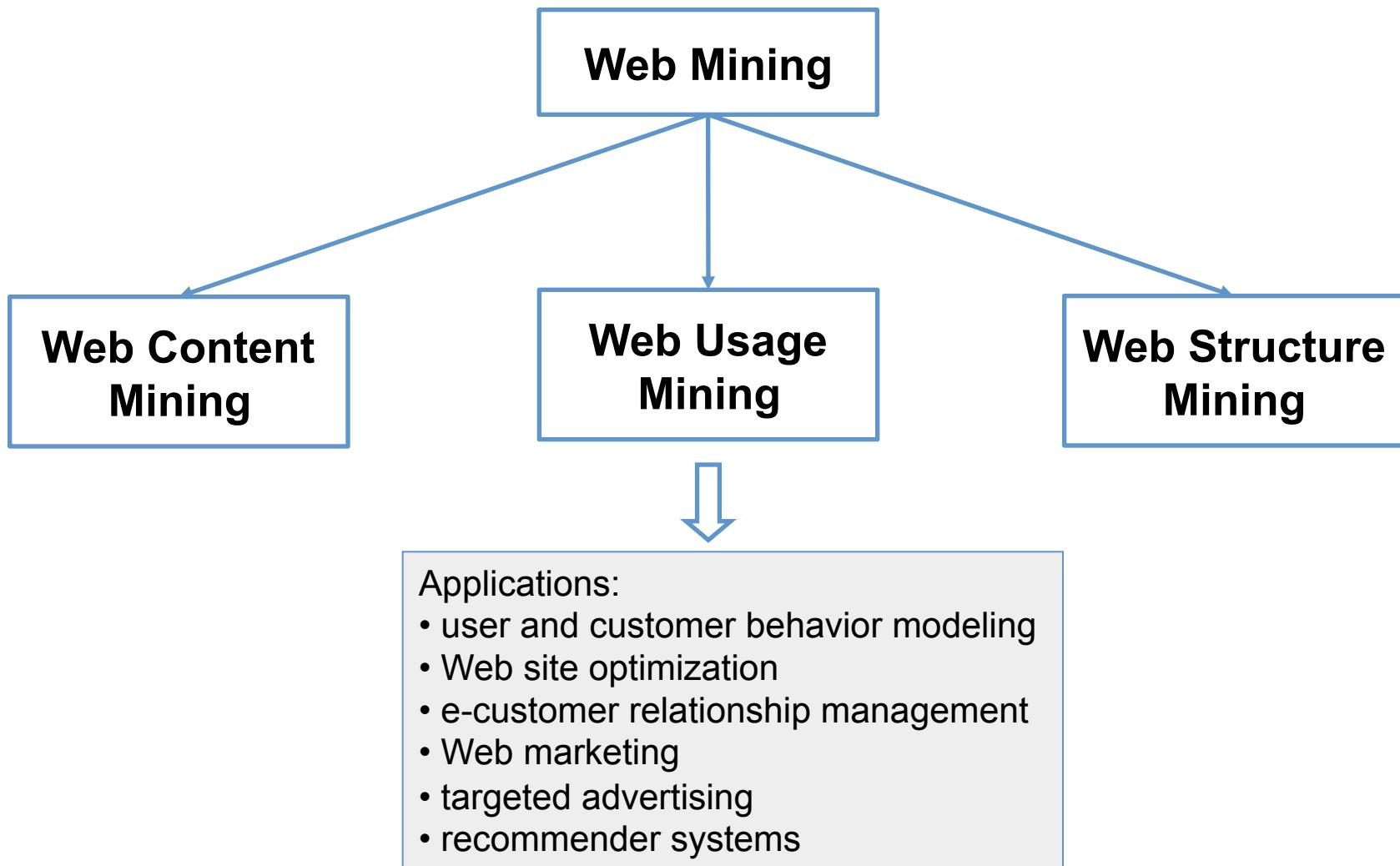
Types of Web Mining



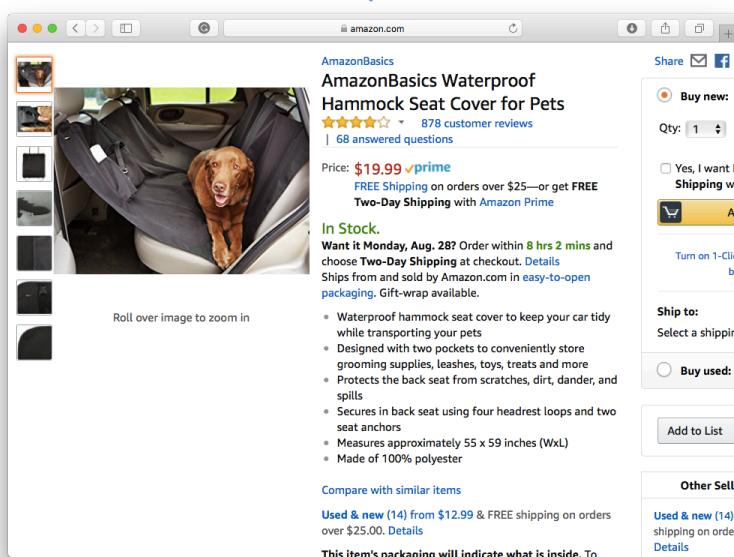
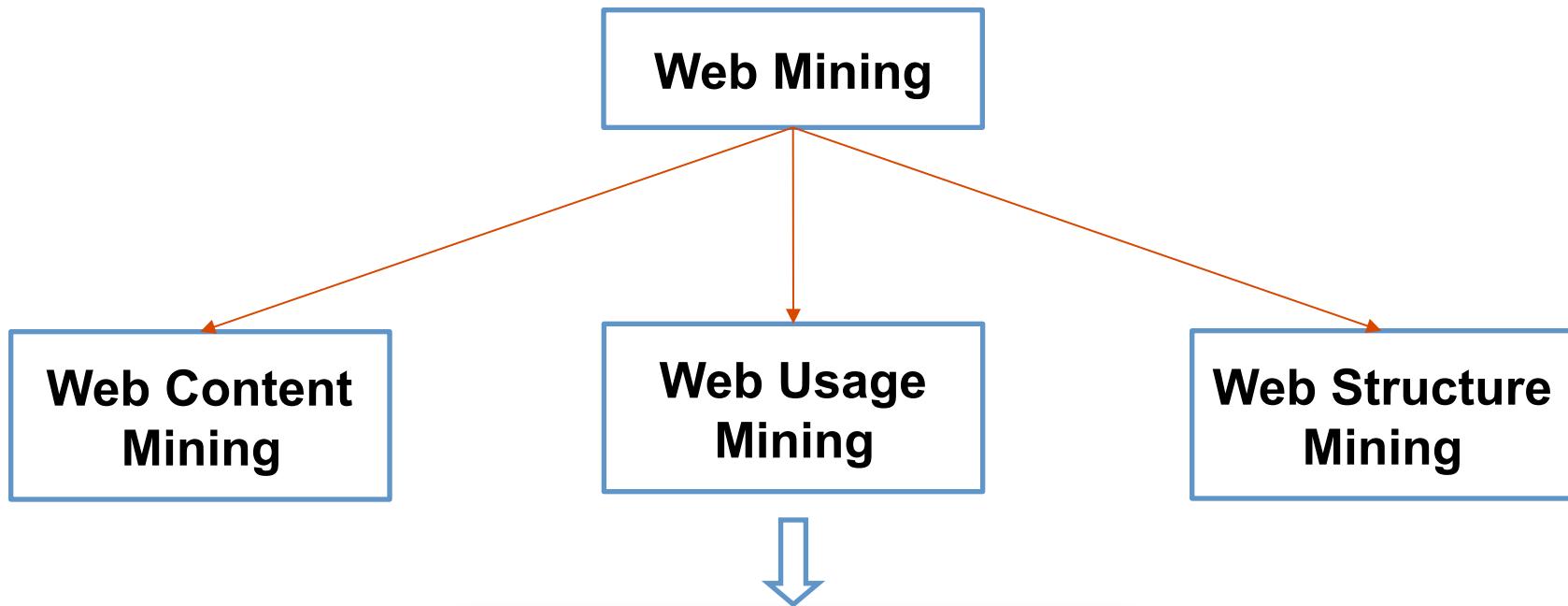
Types of Web Mining



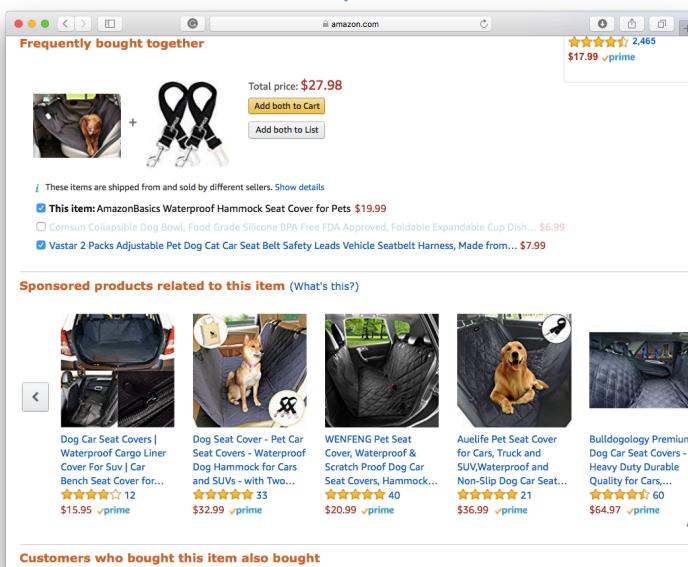
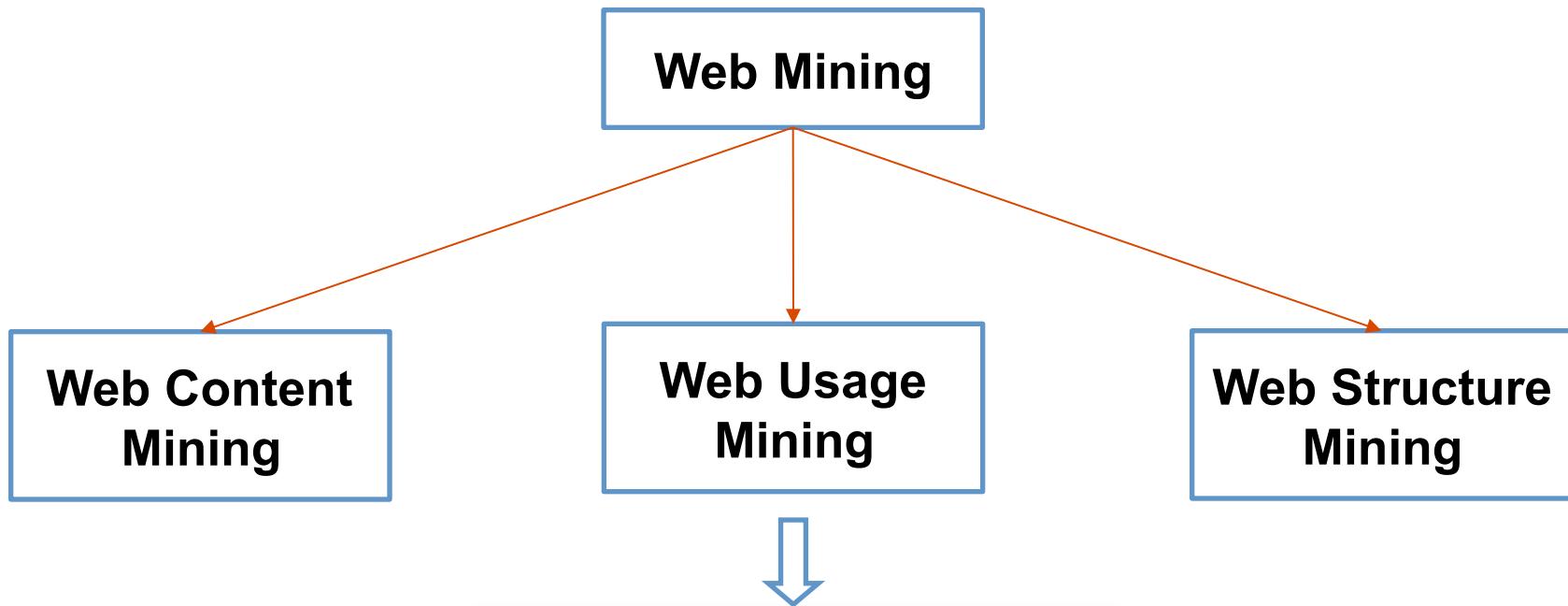
Types of Web Mining



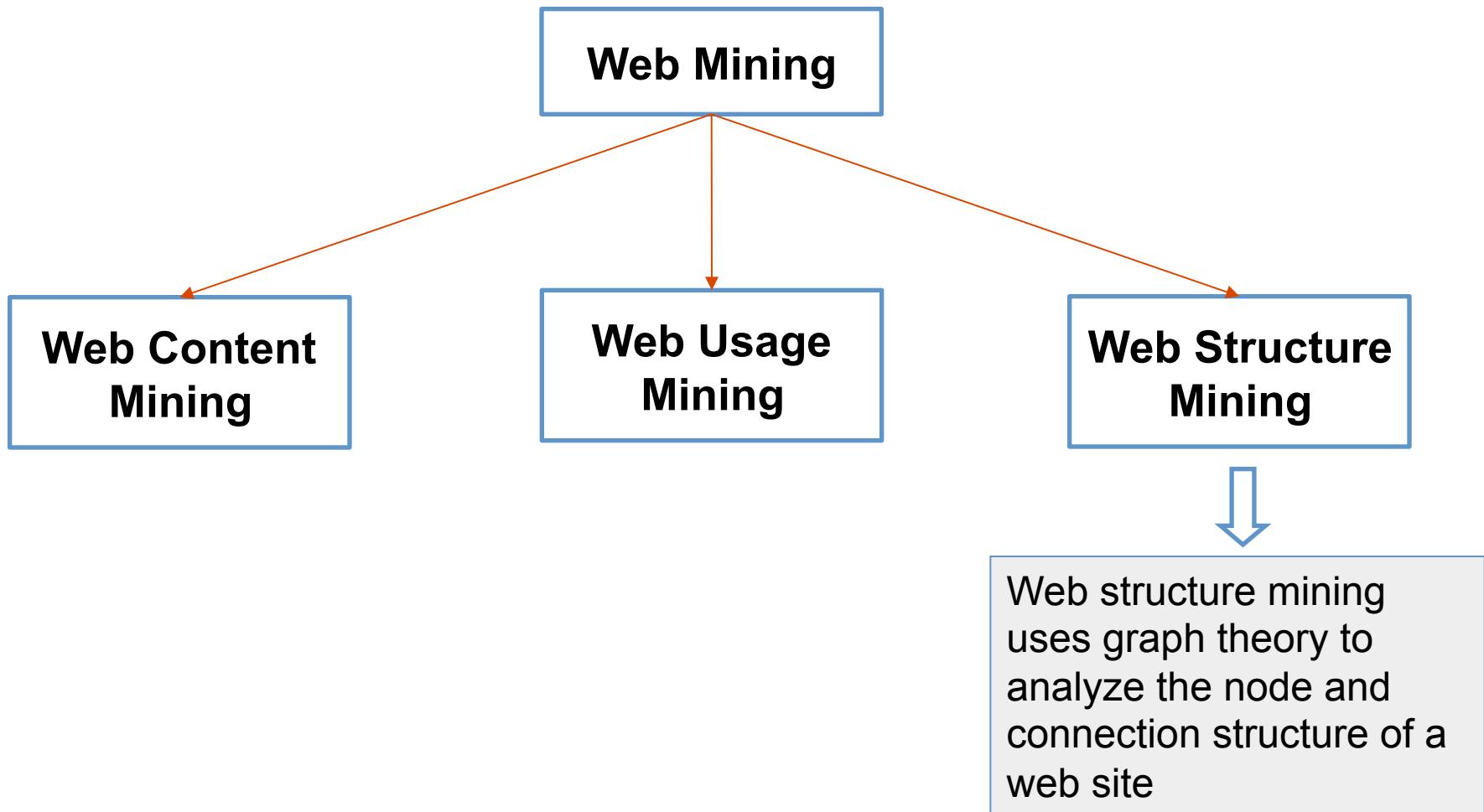
Types of Web Mining



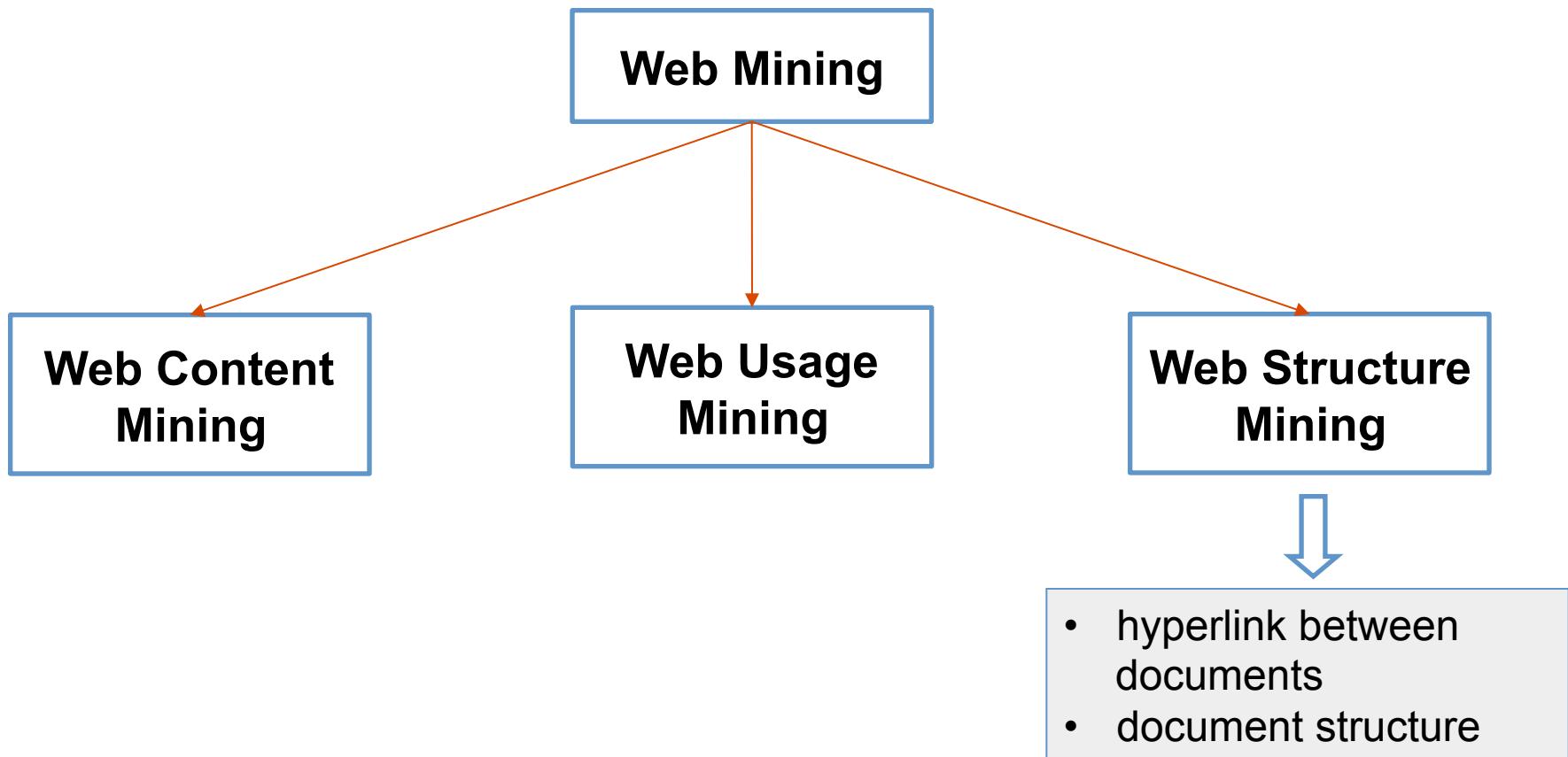
Types of Web Mining



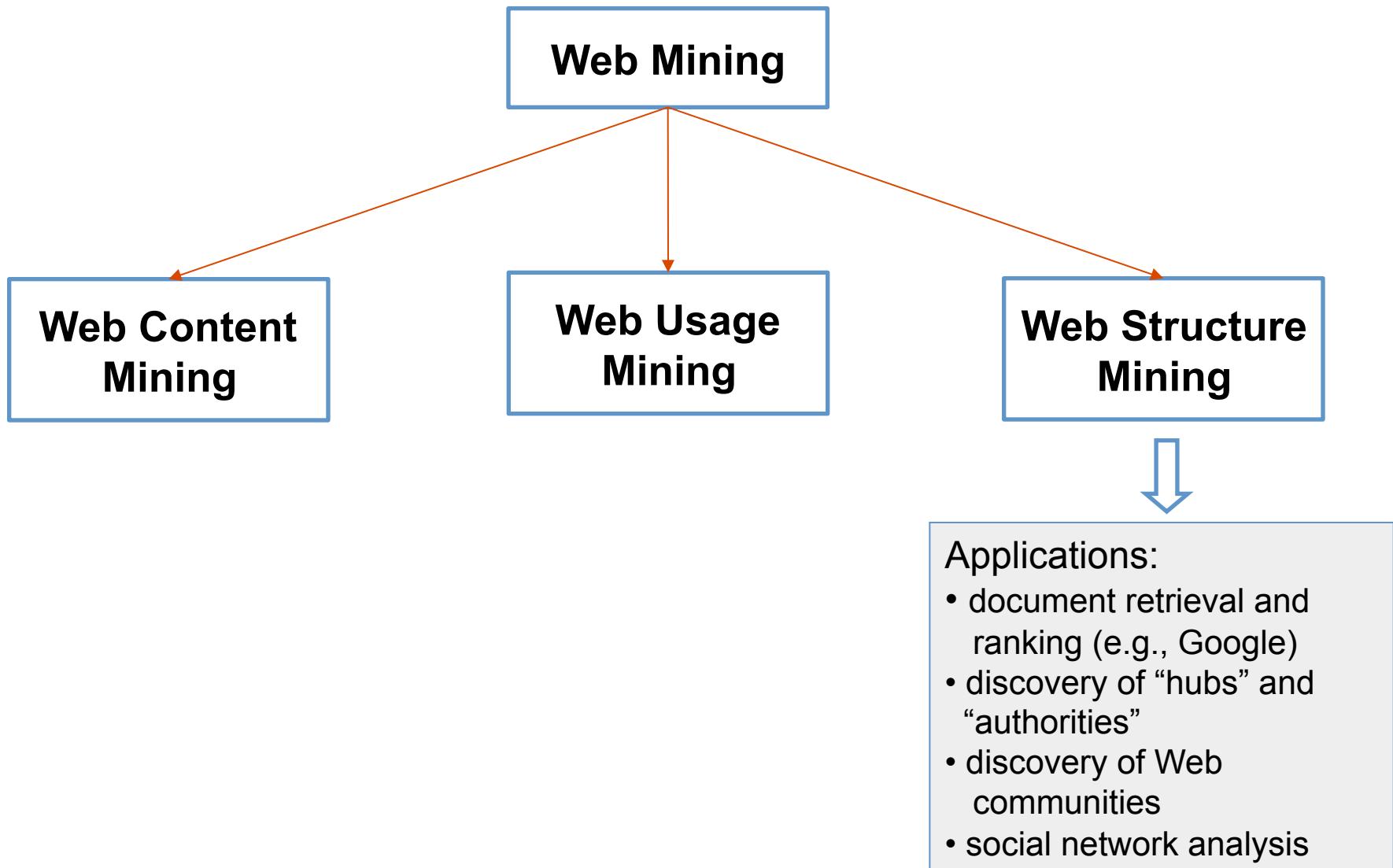
Types of Web Mining



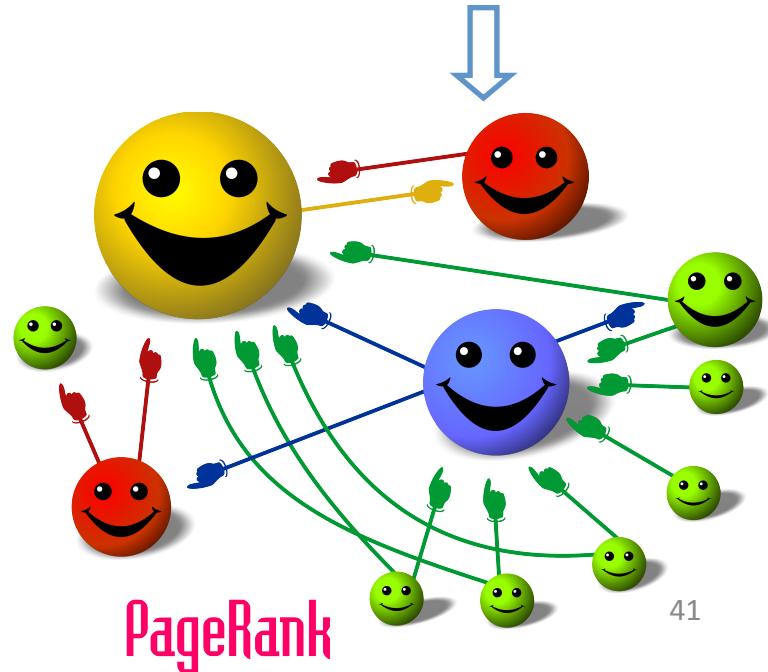
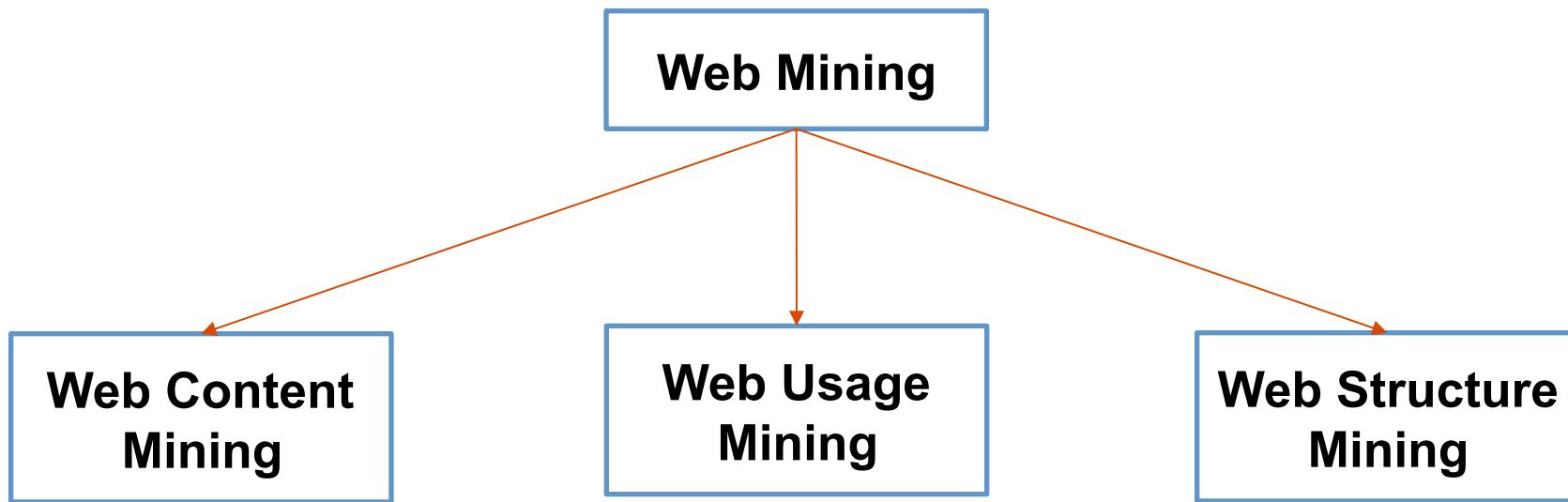
Types of Web Mining



Types of Web Mining



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.firebaseio.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>

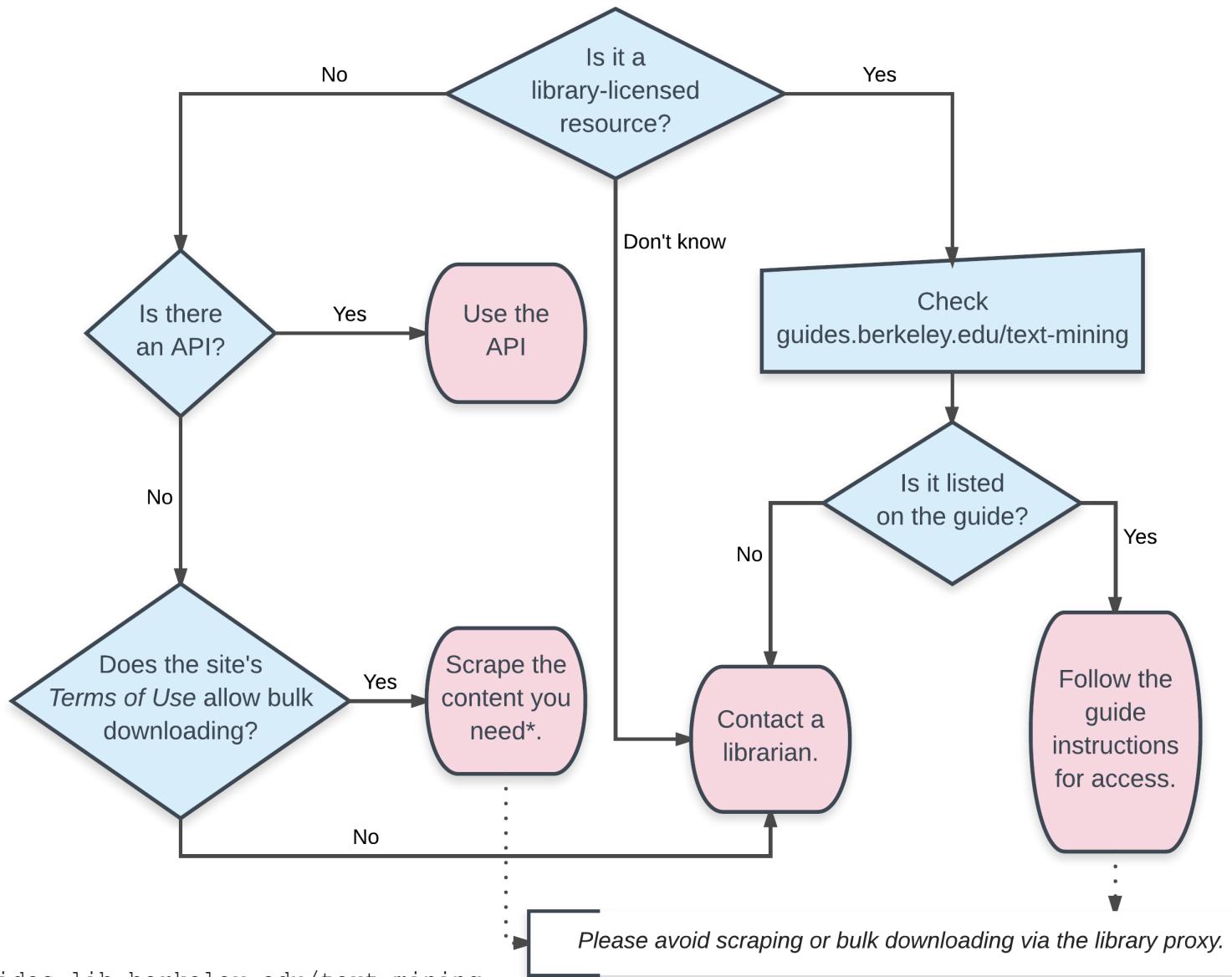
Check out *Visualize This*, Ch 2!

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



Web APIs

- API stands for Application Programming Interface
- 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](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://en.wikipedia.org/w/api.php?action=query&titles>Main%20Page&prop=revisions&rvprop=content&format=json)
- [https://www.googleapis.com/books/v1/volumes?qisbn:
0747532699](https://www.googleapis.com/books/v1/volumes?qisbn:0747532699)
- <https://itunes.apple.com/search?term=jack+johnson>

More than 18.000 APIs available



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