

# Social Data Science

## Scraping the web I

Dr. Thomas Davidson

Rutgers University

September 27, 2021

# Plan

1. Ethics and data science
2. Introduction to webscraping
3. When to use it
4. How to scrape a web page using R (Part I)

# Ethics and data science

## New ethical questions

- ▶ Salganik discusses some examples of recent studies that raise new ethical questions
  - ▶ Emotional contagion experiment on Facebook
  - ▶ Observational study of Facebook networks
  - ▶ Browser-based study of censorship

# Ethics and data science

## Four ethical principles

- ▶ *Respect for persons*
  - ▶ Treating people as autonomous and honoring their wishes
- ▶ *Beneficence*
  - ▶ Understanding risks and benefits; finding the right balance
- ▶ *Justice*
  - ▶ Even distribution of the risks and benefits
- ▶ *Respect for law and public interest*
  - ▶ Extends beyond research participants to other stakeholders
  - ▶ Compliance and transparency-based accountability

# Ethics and data science

## Two ways of thinking about research ethics

- ▶ *Consequentialism*
  - ▶ Focus on the consequences of research
  - ▶ Ends
- ▶ *Deontology*
  - ▶ Consideration of ethical duties, irrespective of consequences
  - ▶ Means
- ▶ Salganik argues that both perspectives most useful when combined

# Ethics and data science

## Case study

- ▶ Researchers at Rutgers decide to use information from Reddit to help improve student services
- ▶ They use Reddit API to collect the complete posting history of all users who posted on r/rutgers
- ▶ A small group of these users is sent a survey. They are also asked for consent to merge their Reddit history and confidential student records
- ▶ The survey results are used to build a statistical model to predict the race, gender, sexual orientation, school year, major, and GPA of *all* r/rutgers posters
- ▶ This information is used to study how the content of posts varies across different groups of students

# Ethics and data science

## Discussion

- ▶ How might this study violate some of the four ethical principles?
- ▶ What issues arise when thinking about this study from a consequentialist or deontological perspective?
- ▶ Could we design the study in a more ethical way?

# Ethics and data science

## Four challenges in digital research

- ▶ Informed consent
  - ▶ When is it practical to get consent to participate?
  - ▶ When is it acceptable to proceed without consent?
- ▶ Managing informational risk
  - ▶ Risks of disclosure of personal information
  - ▶ Anonymization is often imperfect
- ▶ Privacy
  - ▶ What information is public or private?
  - ▶ Context-relative informational norms
- ▶ Ethical decisions and uncertainty
  - ▶ Minimal risk standard
  - ▶ Power analysis



# What is web-scraping?

## Terminology

- ▶ Web-scraping is a method to collect data from websites
  - ▶ We use the code underlying a webpage to collect data (**scraping**)
  - ▶ The process is then repeated for other pages on the same website in an automated fashion (**crawling**)

# What is web-scraping?

## Challenges

- ▶ Different websites have different structures, so a script used to scrape one website will likely have to be changed to scrape another
- ▶ Websites can be internally inconsistent, making them difficult to scrape
- ▶ Some websites are easier to crawl than others
- ▶ Some websites limit or prohibit scraping

# When should I use it?

## Commercial use cases

- ▶ Search engines
  - ▶ Google scrapes websites to create a searchable index of the internet
- ▶ Price comparison
  - ▶ Kayak scrape airlines to compare flight prices, other websites do the same for hotels and rental cars
- ▶ Recruitment
  - ▶ Recruitment companies scrape LinkedIn to get data on workers

# When should I use it?

## Social scientific use cases

- ▶ Web-scraping is a useful tool to collect data from websites without APIs
  - ▶ Large social media platforms and other sites have APIs but smaller websites do not
    - ▶ Local newspapers, forums, small businesses, educational institutions, etc.
- ▶ Often we want to collect data from a single website
  - ▶ e.g. All posts written on a forum
- ▶ Sometimes we might want to collect data from many websites
  - ▶ e.g. All schools in a school district

# When should I use it?

## Ethical and legal considerations

### No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service

Casey Fiesler,<sup>1\*</sup> Nathan Beard,<sup>2</sup> Brian C. Keegan<sup>1</sup>

<sup>1</sup>Department of Information Science, University of Colorado Boulder

<sup>2</sup>College of Information Studies, University of Maryland

#### Abstract

Researchers from many different disciplines rely on social media data as a resource. Whereas some platforms explicitly allow data collection, even facilitating it through an API, others explicitly forbid automated or manual collection processes. A current topic of debate within the social computing research community involves the ethical (or even legal) implications of collecting data in ways that violate Terms of Service (TOS). Using a sample of TOS from over one hundred social media sites from around the world, we analyze TOS language and content in order to better understand the landscape of prohibitions on this practice. Our findings show that

opportunities for digital social research, with new ways of collecting, analyzing, and visualizing data; it also allows for ordered collection, so that messy online data can become usable, well-ordered data sets (Marres and Weltevrede 2013).

However, even when data collection is possible technically, sometimes it is prohibited by terms of service (TOS), which restrict certain behaviors and uses of a site. Whether it is permissible, or ethical, for researchers to violate TOS in the course of collecting data is currently an open question within the social computing research community (Vaccaro et al. 2015; Vitak, Shilton, and Ashktorab 2016).

# When should I use it?

## Ethical and legal considerations

- ▶ Fiesler, Beard, and Keegan (2020)s review the legal cases related to web-scraping and analyze website terms of service
  - ▶ “In short, it is an unsettled question as to whether it is explicitly illegal (or even a criminal act) to violate TOS.”
  - ▶ No academic or journalist has ever been prosecuted for violating a website terms of service to collect data for research
- ▶ They analyze terms of service of over 100 social media websites
  - ▶ Terms of service are ambiguous, inconsistent, and lack context

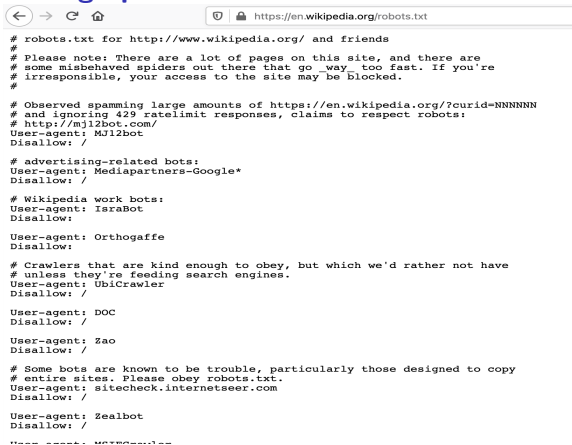
# When should I use it?

## Best-practices

- ▶ Only scrape publicly available data
  - ▶ i.e. You can access the page on the web without logging in
- ▶ Do not scrape copyright protected data
- ▶ Try not to violate website terms of service
- ▶ Do not burden the website
  - ▶ Limit the number of calls you make (similar to rate-limiting in APIs)
- ▶ Avoid using the data in a way that may interfere with business
  - ▶ i.e. Don't copy valuable data from a small business and share it on Github

# How to scrape a web page

## Start by looking up “robots.txt”



The screenshot shows a web browser window with the address bar displaying `https://en.wikipedia.org/robots.txt`. The page content is a text file with the following text:

```
# robots.txt for http://www.wikipedia.org/ and friends
# Please note: There are a lot of pages on this site, and there are
# some misbehaved spiders out there that go _way_ too fast. If you're
# irresponsible, your access to the site may be blocked.
#
# Observed spamming large amounts of https://en.wikipedia.org/?curid=NNNNNN
# and ignoring 429 ratelimit responses, claims to respect robots:
# http://mj12bot.com/
User-agent: MJ12bot
Disallow: /

# advertising-related bots:
User-agent: Mediapartners-Google*
Disallow: /

# Wikipedia work bots:
User-agent: IsraBot
Disallow: /

User-agent: Orthogaffe
Disallow: /

# Crawlers that are kind enough to obey, but which we'd rather not have
# unless they're feeding search engines.
User-agent: UbiCrawler
Disallow: /

User-agent: DOC
Disallow: /

User-agent: Zao
Disallow: /

# Some bots are known to be trouble, particularly those designed to copy
# entire sites. Please obey robots.txt.
User-agent: sitecheck.internetseer.com
Disallow: /

User-agent: Zealbot
Disallow: /

User-agent: MJ12bot
```



# How to scrape a web page

## Decoding robots.txt

- ▶ **User-agent** = the name of the scraper
  - ▶ \* = All scrapers
- ▶ **Allow: /path/** = OK to scrape
- ▶ **Disallow: /path/** = Not OK to scrape
  - ▶ **Disallow: /** = Not OK to scrape any pages
- ▶ **Crawl-Delay: N** = Wait N milliseconds between each call to the website

# How to scrape a web page

## Exercise

- ▶ Find a website of interest
- ▶ Locate the robots.txt file
  - ▶ Does the website allow webscraping?
  - ▶ Are there any restrictions on which pages can be accessed?

# How to scrape a web page

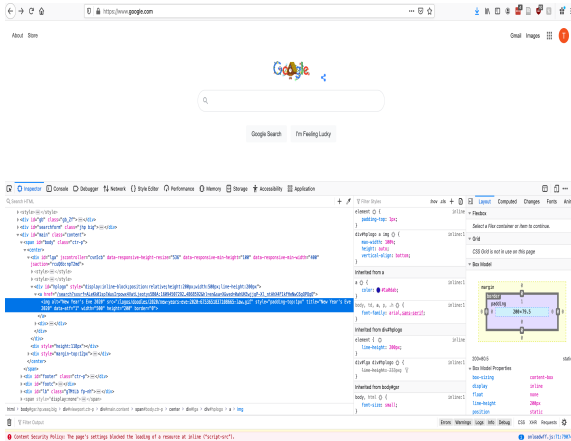
## Terminology

- ▶ A web-page is loaded using a **URL** (Uniform Resource Locator)
- ▶ The underlying code we are interested in is usually **HTML** (Hypertext Markup Language)
- ▶ Many websites use **CSS** (Cascading Style Sheets) to structure HTML
  - ▶ This will help us to find what we are interested in
    - ▶ See <https://flukeout.github.io/> for an interactive tutorial on using CSS selectors
    - ▶ Chrome Plugin to help find CSS elements: <https://selectorgadgets.com/>

# How to scrape a web page

## Inspecting HTML

- ▶ Open up a website and right click on any text or image on the screen
  - ▶ You should see an option `Inspect Element`
  - ▶ This will allow you to see the code used to generate the page



# How to scrape a web page


the catsite

BROWSE BY CAT TOPICS CONNECT WITH US


CAT ARTICLES FORUMS MEDIA MEMBERS

LOG IN REGISTER SEARCH

Too many ads? Join our community of cat lovers now to reduce ads by 90%! [Click here to join for free!](#)






What Are Siamese Flame-Point Cats?  
[FAQ & Pictures]  
How to identify these gorgeous felines...




Does Coat Color Affect Cat Personality?  
Can you tell how friendly a cat is by fur color and pattern?...

Featured Threads




 <b>Featured</b> <b>Rescued Senior Care</b> saltslime Yesterday at 8:41 PM - Cat Health	Replies: 6 Views: 110	Today at 8:31 AM FurryDancer
 <b>Featured</b> <b>Recent Rescue With Mange.</b> Christina Yesterday at 9:32 PM - Cat Health	Replies: 2 Views: 102	Yesterday at 11:56 PM Mario Bayote
 <b>Featured</b> <b>Fostering a Feral</b> baldy Yesterday at 9:36 AM - Caring for Strays and Ferals	Replies: 5 Views: 122	Yesterday at 1:23 PM fousamom

Unanswered Threads

 <b>Two Separate Issues: Enlarged Kidney(?) and</b>	Replies: 0 Views: 1	38 minutes ago
--	------------------------	----------------

NEW POSTS

STAFF ONLINE


-  Furbalmon  
Cat Fan especially Black Cats
-  nubysmama  
Forum Helper
-  Kat0121  
Advisor

MEMBERS ONLINE

Furnance02, nolly60, arny, Furbalmon, deficat15, kacie, C12, cat nap, nubysmama, glinnag, BellaGood, CatladyJan, Wherethehellarekisses, em, Kat0121, Junkie66, Fels, movinonline, trunderseed

Total: 947 (members: 27, guests: 920)

LATEST POSTS

 Advice on Cat Introductions -  
Feeling a Bit Lost  
Latest: Furnance02 - A moment ago

# How to scrape a web page

The screenshot shows a web browser displaying the 'thecatsite' forum. The URL in the address bar is <https://thecatsite.com/threads/advice-on-cat-introductions-feeling-a-bit-lost.422848/>. The forum header includes navigation links for 'CAT ARTICLES', 'FORUMS', 'MEDIA', and 'MEMBERS', along with buttons for 'BROWSE BY CAT TOPICS', 'CONNECT WITH US', 'LOG IN', 'REGISTER', and 'SEARCH'. A yellow banner promotes joining the community to reduce ads by 90%.

The thread title is 'Advice on Cat Introductions - Feeling a Bit Lost' by user 'Fumama22', dated Dec 22, 2020. The thread starter's profile shows they joined Dec 22, 2020, have 19 messages, and a reaction score of 12.

The first post, dated Dec 22, 2020, at 11:01, contains the following text:

Hi all,

I'm new to the forum and have been reading all of your excellent thoughts and posts on cat introductions. Our new cat is Florence, a 4 year old spayed female from the local humane society. Our resident cat is Howthorne, a 10 1/2 year old neutered male. Howthorne has always been an anxious and not particularly cuddly cat (he has moments, but is NOT a lap cat - he is still scared of my stepkids after four years) and his friend and my beloved cat Tennyson passed away in August. They were not bonded, and fought occasionally, but generally speaking got along and did friendly nose touches.

We have had Florence for about 6 weeks. We lost one week (week #2) of introduction time because she was very sick with a URI (she is healthy now). She is declawed (something we found out after we knew we wanted her) and was surrendered by her family for "not getting along with their kids." They did not provide any additional information. From my perspective, she's one of the sweetest cats I've ever known. She has spirit but she is quite confident now and good-natured.

The right sidebar features a 'LATEST POSTS' section with the following entries:

- Happy New Year!**  
Latest: lozlie - 2 minutes ago  
The Cat Lounge
- Adult Cats Fighting - Play, Territory, What to Do?**  
Latest: cat nap - 9 minutes ago  
Cat Behavior
- Thread Devoted To Sharing Your Favorite Albums And Songs - 2020**  
Latest: rubysnana - 13 minutes ago  
The Cat Lounge
- post funny picture and memes here**  
Latest: Whenaithelbreaks - 14 minutes ago  
Fur Pictures and Video Only!
- Growing a Human - it's a gift!**  
Latest: Photo: Lisa - 17 minutes ago

# How to scrape a web page

## Using rvest to scrape HTML

```
library(rvest)
library(tidyverse)
library(stringr)
```



# How to scrape a web page

## Using rvest to scrape HTML

```
url <- "https://thecatsite.com/threads/advice-on-cat-introductions-feel  
thread <- read_html(url)
```

# How to scrape a web page

## Using rvest to scrape HTML

```
class(thread)
```

```
## [1] "xml_document" "xml_node"
```

```
print(thread)
```

```
## {html_document}
```

```
## <html id="XF" lang="en-US" dir="LTR" data-app="public" data-template
```

```
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; char
```

```
## [2] <body data-template="thread_view">\n<div class="p-pageWrapper" i
```

# How to scrape a web page

## Collecting messages

First, we parse the HTML to obtain the text of each message on the page. Here we use the CSS selector `.message-body`, which selects all elements with class `message-body`. The `html_nodes` function in `rvest` allows us to retrieve these nodes.

```
message.data <- thread %>% html_nodes(".message-body")  
print(message.data[1])
```

```
## {xml_nodeset (1)}
```

```
## [1] <article class="message-body js-selectToQuote"><div class="bbWra
```

# How to scrape a web page

## Collecting messages

Next we use `html_text()` to extract the text from the HTML.

```
messages <- thread %>% html_nodes(".message-body") %>%  
  html_text() %>% str_trim()  
messages[1]
```

```
## [1] "Hi all,\nI'm new to the forum and have been reading all of your"
```

# How to scrape a web page

## Collecting messages

As expected, there are twenty messages.

```
print(length(messages))
```

```
## [1] 20
```

```
print(substr(messages[1], 1, 250)) # print a substring
```

```
## [1] "Hi all,\nI'm new to the forum and have been reading all of your
```

# How to scrape a web page

## Getting user names

Next we collect the name of each user using the same logic. User information is found by parsing the `.message-userDetails` node.

```
users <- thread %>% html_nodes(".message-userDetails") %>%  
  html_text() %>% str_trim()  
print(length(users))
```

```
## [1] 20
```

```
class(users)
```

```
## [1] "character"
```

```
users[1]
```

```
## [1] "Furmama22\nTCS Member\nThread starter\nAdult Cat"
```

# How to scrape a web page

## Getting user names

Let's add some more elements to the pipe to extract the user name from this string. Note how the elements in the string returned in the previous chunk are separated by the newline symbol (`\n`).

```
users <- thread %>% html_nodes(".message-userDetails") %>%  
  html_text() %>% str_trim() %>% str_split('\n')  
class(users)
```

```
## [1] "list"
```

```
users[1:2]
```

```
## [[1]]
```

```
## [1] "Furmama22"      "TCS Member"      "Thread starter" "Adult Cat"
```

```
##
```

```
## [[2]]
```

```
## [1] "calicosrspecial" "TCS Member"      "Top Cat"
```

# How to scrape a web page

## Getting user names

The final step is to get the name from each list. This can be done by using the map command.

```
users <- thread %>% html_nodes(".message-userDetails") %>%  
  html_text() %>% str_trim() %>% str_split('\n') %>% map(1)  
class(users)
```

```
## [1] "list"
```

```
users[1:2]
```

```
## [[1]]
```

```
## [1] "Furmama22"
```

```
##
```

```
## [[2]]
```

```
## [1] "calicosrspecial"
```



# How to scrape a web page

## Collecting timestamps

Finally, we also want to get the time-stamp of each message. While the forum only displays dates, we can actually get the full timestamp. What's the problem here?

```
dates <- thread %>% html_nodes("time.u-dt")
print(dates[1])
```

```
## {xml_nodeset (1)}
```

```
## [1] <time class="u-dt" dir="auto" datetime="2020-12-22T11:26:12-0800
```

```
length(dates)
```

```
## [1] 27
```

# How to scrape a web page

## Collecting timestamps

I went back to the HTML and found this CSS selector `.u-concealed .u-dt` is selected instead. It returns the datetime for each post in the thread, along with the date time at the top indicating when the thread was created.

```
dates <- thread %>% html_nodes(".u-concealed .u-dt")
length(dates)
```

```
## [1] 21
```

```
dates[1]
```

```
## {xml_nodeset (1)}
```

```
## [1] <time class="u-dt" dir="auto" datetime="2020-12-22T11:26:12-0800
```

```
class(dates[1])
```

```
## [1] "xml_nodeset"
```

# How to scrape a web page

## Collecting timestamps

Each HTML node contains several different attributes related to the time. In this case we can select the datetime attribute using the `html_attr` function.

```
dates <- thread %>% html_nodes(".u-concealed .u-dt") %>% html_attr("data-datetime")
dates[1]
```

```
## [1] "2020-12-22T11:26:12-0800"
```

```
class(dates[1])
```

```
## [1] "character"
```

# How to scrape a web page

## Collecting timestamps

Finally, its often useful to clean up timestamps. We can do this using the `lubridate` package. In this case we extract the year, month, day, hour, minutes, and seconds, converted to EST. The result is a special type of object used to represent dates and times.

```
library(lubridate)
dates <- dates %>% ymd_hms(tz = "EST")
dates[1]
```

```
## [1] "2020-12-22 14:26:12 EST"
```

```
class(dates)
```

```
## [1] "POSIXct" "POSIXt"
```

# How to scrape a web page

## Putting it all together

```
length(users)
```

```
## [1] 20
```

```
class(users)
```

```
## [1] "list"
```

```
length(messages)
```

```
## [1] 20
```

```
class(messages)
```

```
## [1] "character"
```

```
length(dates)
```

```
## [1] 21
```

```
class(dates)
```

```
## [1] "POSIXct" "POSIXt"
```

# How to scrape a web page

## Putting it all together

```
data <- as_tibble(cbind(messages, unlist(users), dates[-1]))
colnames(data) <- c("message", "user", "timestamp")
head(data)
```

```
## # A tibble: 6 x 3
```

##	message	user
##	<chr>	<chr>
## 1	"Hi all,\nI'm new to the forum and have been reading a~	Furmama22
## 2	"Furmama22 said:\n\n\n\nHi all,\nI'm new to the forum ~	calicosrsp
## 3	"Thank you SO much for taking the time to reply. I rea~	Furmama22
## 4	"I don't think I can add a thing to \nC\n calicosrspec~	Mamanyt195
## 5	"Thanks so much for your thoughts and comments! And th~	Furmama22
## 6	"You certainly came to the right place! And, in my exp~	Mamanyt195

# How to scrape a web page

## Creating a function to collect and store data

```
get.posts <- function(thread) {  
  messages <- thread %>% html_nodes(".message-body") %>%  
    html_text() %>% str_trim()  
  users <- thread %>% html_nodes(".message-userDetails") %>%  
    html_text() %>% str_trim() %>% str_split('\n') %>% map(1)  
  timestamps <- thread %>% html_nodes(".u-concealed .u-dt") %>%  
    html_attr("datetime") %>% ymd_hms(tz="EST")  
  timestamps <- timestamps[-1] # remove first timestamp  
  data <- as_tibble(cbind(messages, unlist(users), timestamps))  
  colnames(data) <- c("message", "user", "timestamp")  
  return(data)  
}
```

# How to scrape a web page

## Using the function

We can now easily run all the code to extract information using a single function call:

```
results <- get.posts(thread)
head(results)
```

```
## # A tibble: 6 x 3
```

```
##   message
```

```
user
```

```
##   <chr>
```

```
<chr>
```

```
## 1 "Hi all,\nI'm new to the forum and have been reading a~ Furmama22
```

```
## 2 "Furmama22 said:\n\n\nHi all,\nI'm new to the forum ~ calicosrsp
```

```
## 3 "Thank you SO much for taking the time to reply. I rea~ Furmama22
```

```
## 4 "I don't think I can add a thing to \nC\n calicosrspec~ Mamanyt195
```

```
## 5 "Thanks so much for your thoughts and comments! And th~ Furmama22
```

```
## 6 "You certainly came to the right place! And, in my exp~ Mamanyt195
```