Teaching web scraping:

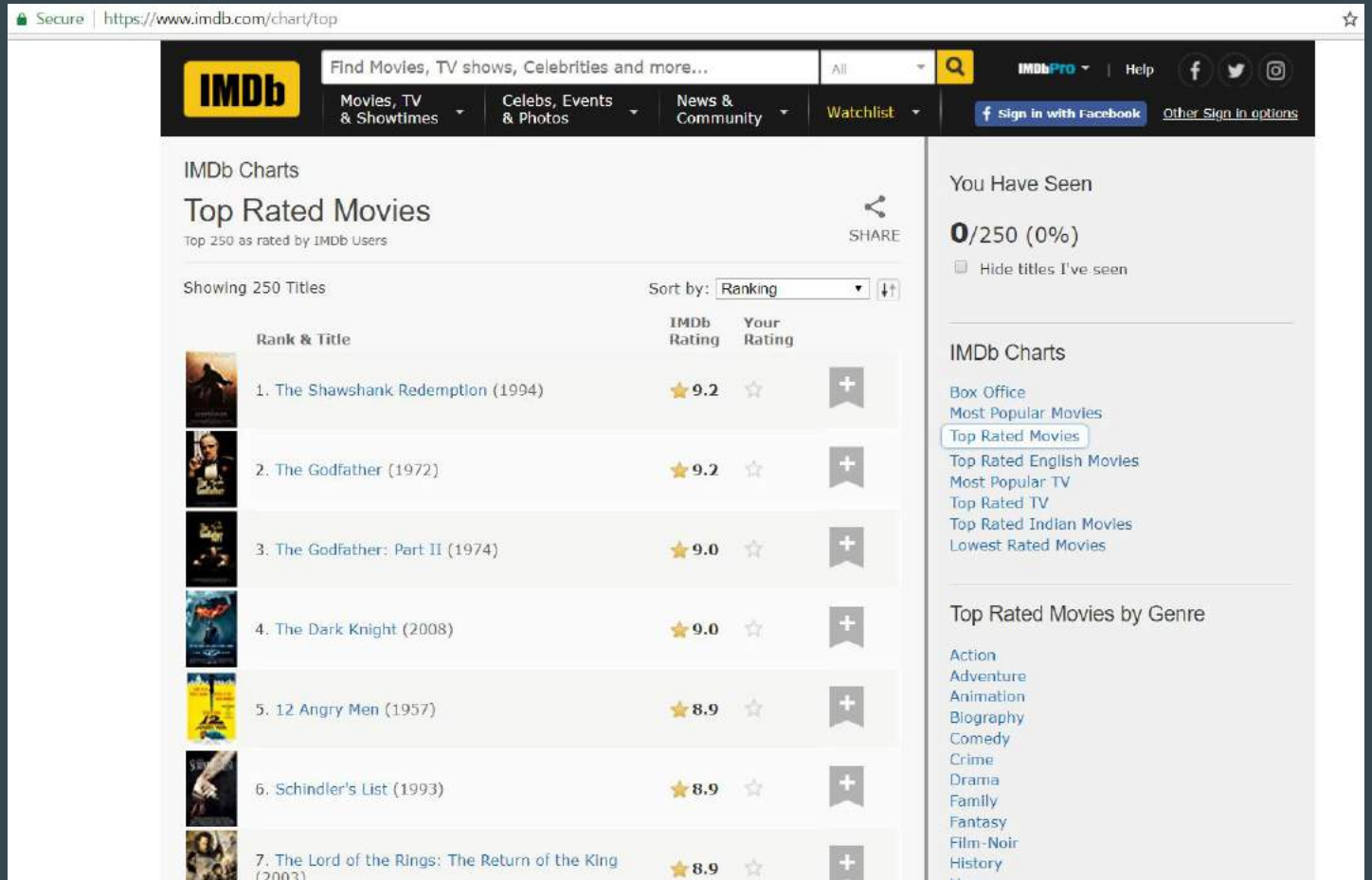
Integrating data science into statistics

•••

Mine Dogucu
New College of Florida
@MineDogucu
Mine Çetinkaya-Rundel
Duke University
@minebocek
2018-05-21

What is Web Scraping?

Example 1


Secure
https://www.imdb.com/chart/top
IMDb
Find Movies, TV shows, Celebrities and more...
All
IMDbPro
Help
Movies, TV & Showtimes
Celebs, Events & Photos
News & Community
Watchlist
Sign in with Facebook
Other Sign in options
IMDb Charts
Top Rated Movies
Top 250 as rated by IMDb Users
SHARE
Showing 250 Titles
Sort by: Ranking
Rank & Title
IMDb Rating
Your Rating
1. The Shawshank Redemption (1994) 9.2
2. The Godfather (1972) 9.2
3. The Godfather: Part II (1974) 9.0
4. The Dark Knight (2008) 9.0
5. 12 Angry Men (1957) 8.9
6. Schindler's List (1993) 8.9
7. The Lord of the Rings: The Return of the King (2003) 8.9
You Have Seen
0/250 (0%)
Hide titles I've seen
IMDb Charts
Box Office
Most Popular Movies
Top Rated Movies
Top Rated English Movies
Most Popular TV
Top Rated TV
Top Rated Indian Movies
Lowest Rated Movies
Top Rated Movies by Genre
Action
Adventure
Animation
Biography
Comedy
Crime
Drama
Family
Fantasy
Film-Noir
History

Hand Scraping

	A	B	C	D
1		Rank & Title	IMDb Rating	Your Rating
2		1. The Shawshank Redemption (1994)	9.2	
3				
4				
5		2. The Godfather (1972)	9.2	
6				
7				
8		3. The Godfather: Part II (1974)	9	
9				
10				
11		4. The Dark Knight (2008)	9	
12				
13				

.xls

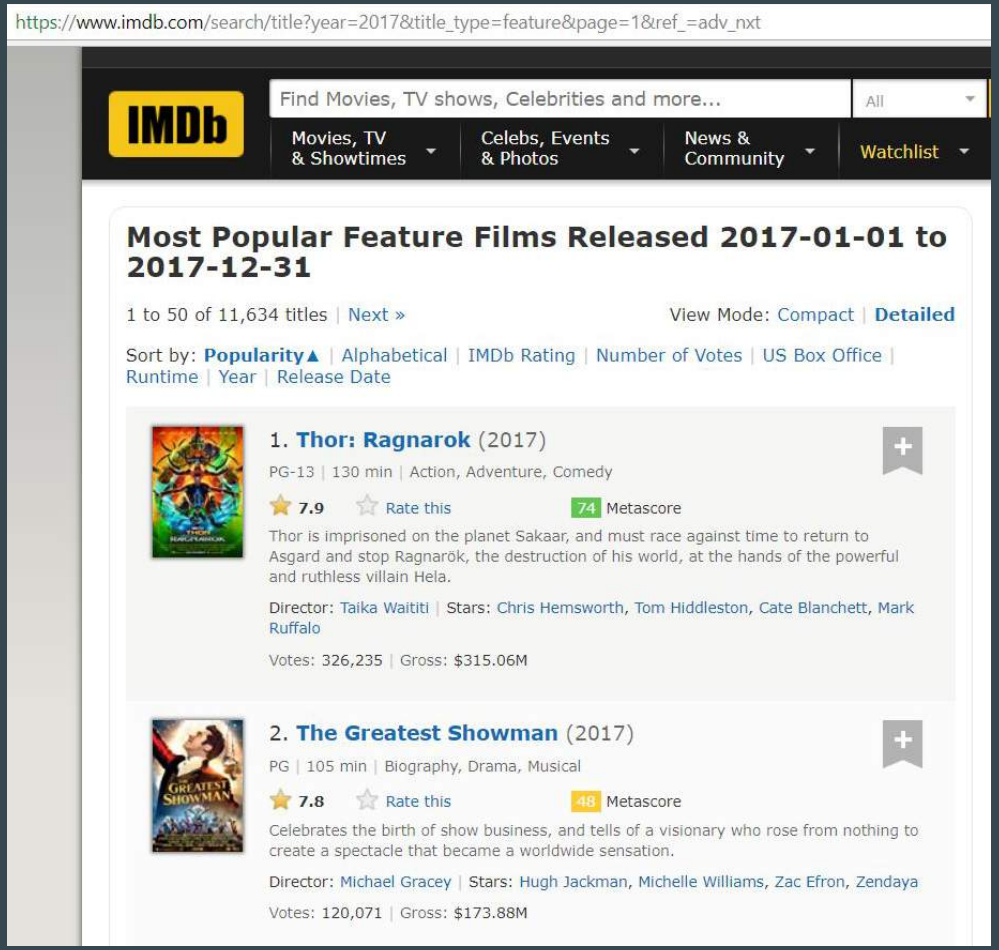
	A	B	C	D
1		Rank & Tit	IMDb Rati	Your Rating
2		1. The Sha	9.2	
3				
4				
5		2. The Goc	9.2	
6				
7				
8		3. The Goc	9	
9				
10				
11		4. The Dar	9	
12				
13				

.csv

Web scraping

	title	year	rating
1	The Shawshank Redemption	1994	9.2
2	The Godfather	1972	9.2
3	The Godfather: Part II	1974	9.0
4	The Dark Knight	2008	9.0
5	12 Angry Men	1957	8.9
6	Schindler's List	1993	8.9
7	The Lord of the Rings: The Return of the King	2003	8.9
8	Pulp Fiction	1994	8.9
9	The Good, the Bad and the Ugly	1966	8.8
10	Fight Club	1999	8.8
11	The Lord of the Rings: The Fellowship of the Ring	2001	8.8

Example 2



Beginner's Guide on Webscraping in R using rvest on Analytics Vidhya - http://bit.ly/rvest-example

Hypertext Markup Language (HTML) Nodes

```
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>My First Heading</h1>
<p>My first paragraph.</p>

</body>
</html>
```

https://www.w3schools.com/html/tryit.asp?filename=tryhtml_intro

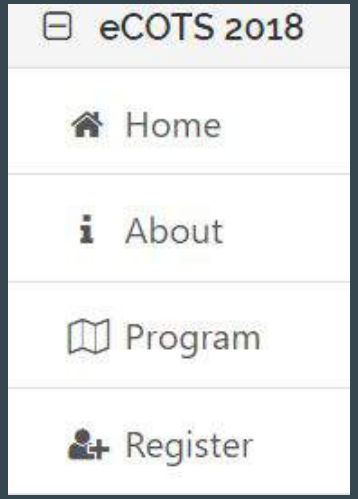
Electronic Conference On Teaching Statistics

eCOTS has been made possible by:

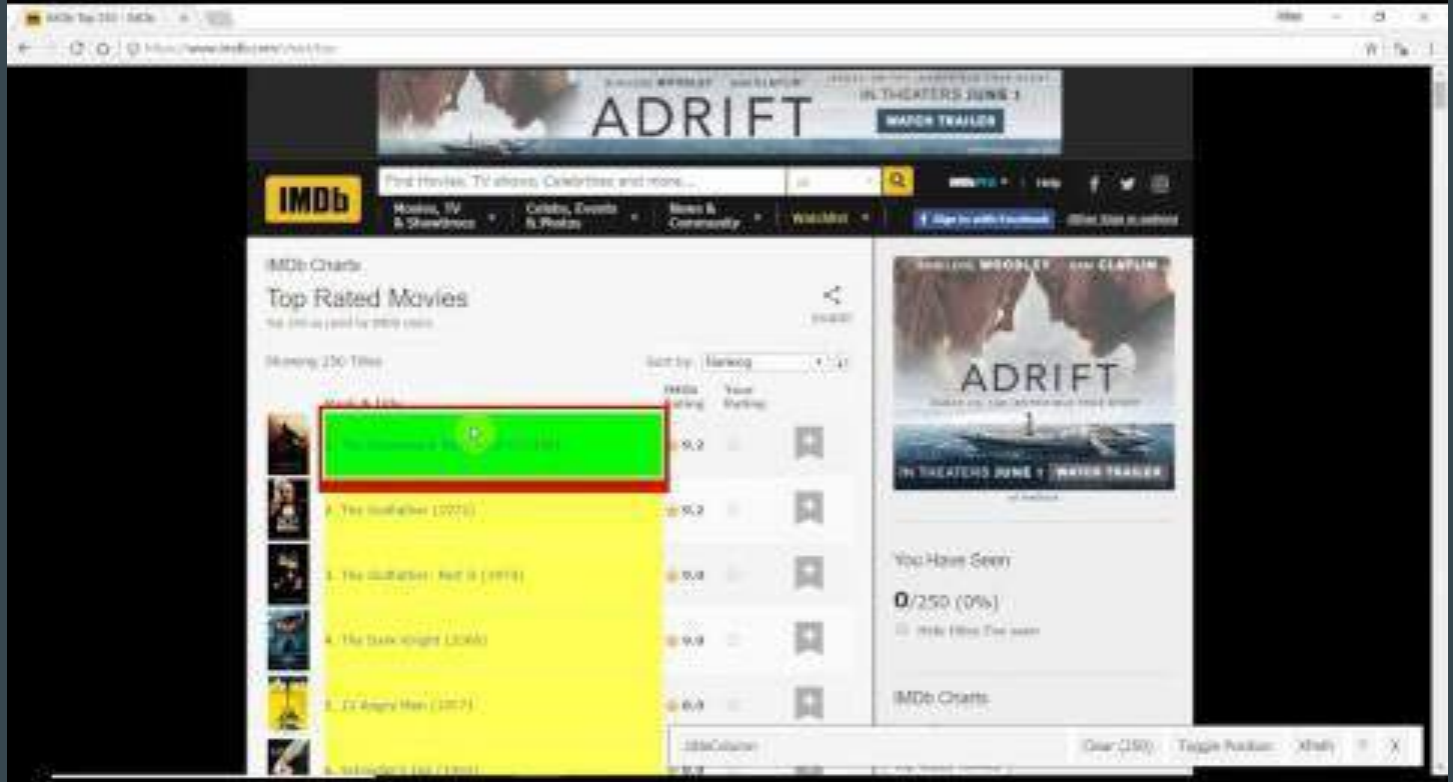
```
<div class="panel panel-default">
    <div class="panel-heading">
        <h4 class="panel-title">
            <a data-toggle="collapse" data-parent="#accordion" href="#ecots18">
                <i id="ecots18_toggle" class="fa fa-minus-square-o fa-fw"></i>
                </span> eCOTS 2018</a>
        </h4>
    </div>
    <div id="ecots18" class="panel-collapse collapse in">
        <div class="list-group">
            <a href="/cause/ecots/ecots18/" class="list-group-item">
                <i class="fa fa-fw fa-home"> </i> Home</a>
            <a href="/cause/ecots/ecots18/about" class="list-group-item">
                <i class="fa fa-fw fa-info"> </i> About</a>
            <!--<a href="/cause/ecots/ecots18/proposals/submit" class="list-group-item">
                <i class="fa fa-fw fa-file-text"> </i> Call for Proposals</a>-->
           <a href="/cause/ecots/ecots18/program" class="list-group-item">
                <i class="fa fa-fw fa-map-o"> </i> Program</a>
            <a href="/cause/ecots/ecots18/register" class="list-group-item">
                <i class="fa fa-fw fa-user-plus"> </i> Register</a>

            <a href="/cause/ecots/ecots18/program/birds-of-a-feather" class="list-group-item">
                <i class="fa fa-fw fa-leaf"> </i> Birds of a Feather</a>
            <a href="/cause/ecots/ecots18/program/breakouts" class="list-group-item">
            <i class="fa fa-fw fa-comments"> </i> Breakout Sessions</a>

<a href="/cause/ecots/ecots18/keynotes" class="list-group-item">
                <i class="fa fa-fw fa-microphone"> </i> Keynotes</a>
```

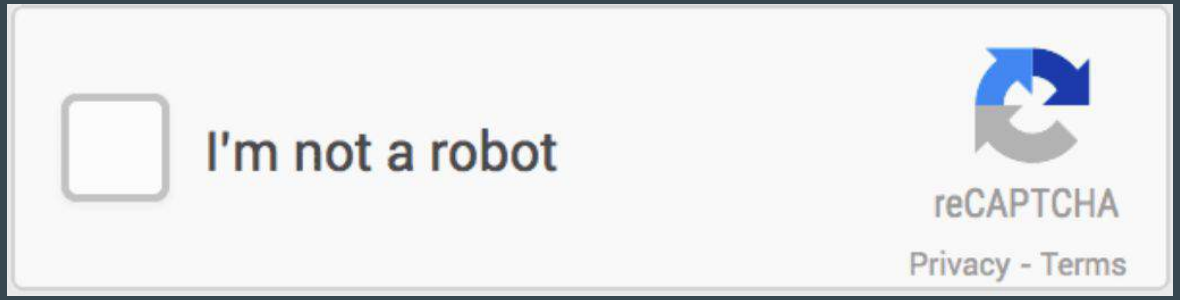


bit.ly/SelectorGadget


ADRIFT
IN THEATERS JUNE 1
WATCH TRAILER
IMDb
Top Rated Movies
ADRIFT
You Have Seen
0/250 (0%)
IMDb Charts


IMDb
Most Popular Feature
2017-12-31
2. The Greatest Showman (2017)
3. Revenge (II) (2017)

Things to Consider



```
library(robotstxt)


paths_allowed("http://www.imdb.com")
#>
 www.imdb.com
#> [1] TRUE
```


Created on 2018-05-20 by the reprex package (v0.2.0).

Things to Consider

```
library(rvest)
library(tidyverse)
page <- read_html("http://www.imdb.com/chart/top")
titles <- page %>%
  html_nodes(".titleColumn a") %>%
  html_text()

head(titles)
#> [1] "The Shawshank Redemption" "The Godfather"
#> [3] "The Godfather: Part II"   "The Dark Knight"
#> [5] "12 Angry Men"             "Schindler's List"
```

Reads an HTML or XML object

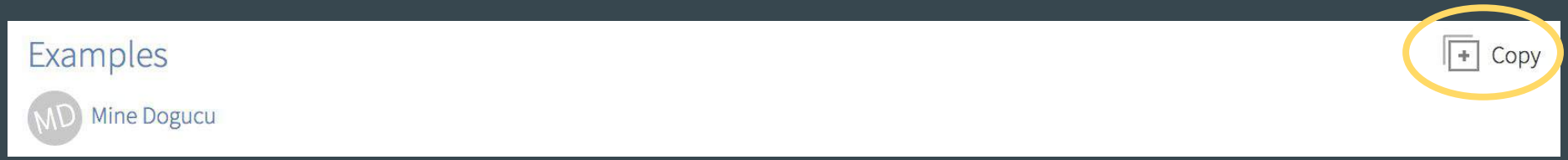
Selector

Text value

bit.ly/ecots2018

Click on 

Make your own copy of the project called Examples



Benefits

- Students get exposed to non-standard (non-rectangular) data format.
- Students get large amounts of data in a short span of time and in a tidy format.
- Students get exposure to working with strings.
- Students can have more diverse sources of data for statistics projects.
- Web scraping can bring computing topics (e.g. HTML, functions, loops) into the statistics classroom.
- Instructors can use web scraping to curate datasets for classroom use.

Potential Problems

- Website can be down
- NA values

Notes

- More complex scraping is possible
- Timing in the semester
- Web APIs
- Terms of Use

QUESTIONS?

Mine Dogucu
mdogucu@ncf.edu

@MineDogucu

Mine Çetinkaya-Rundel
mine@stat.duke.edu

@minebocek

bit.ly/ecots2018-webscraping
https://github.com/mdogucu/eCOTS2018