

Extracting Data from the Web

Part 2: **Web Scraping**



Garrett Grolemund
Data Scientist, Educator
November 2016

- 1. APIs** (www.rstudio.com/resources/webinars/)
- 2. Web Scraping** (Today)
 - a. Basic Idea - HTML, CSS
 - b. rvest
 - c. SelectorGadget



rvest SelectorGadget Vignette

400 words, <5 minutes to read

cran.r-project.org/web/packages/rvest/vignettes/selectorgadget.html





Extracting Data from the Web

Scott Chamberlain, Karthik Ram, Me
github.com/ropensci/user2016-tutorial



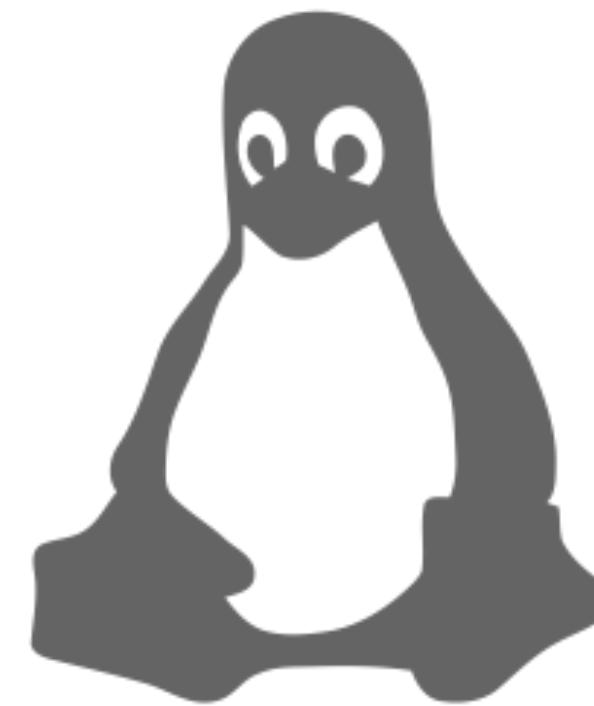
Web Scraping

APIs

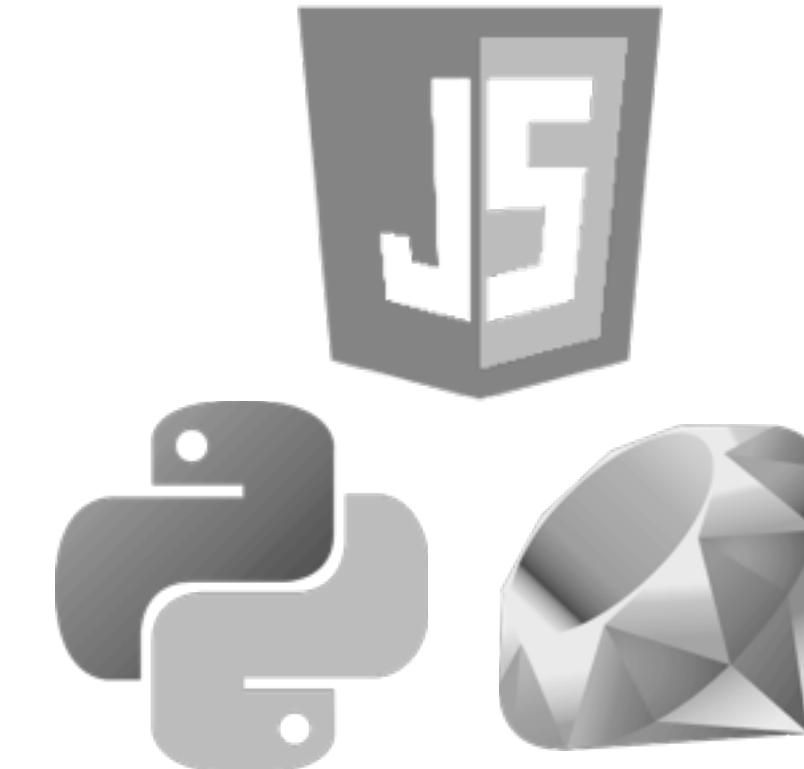
An interface for your code to interact with a piece of software:



Database



OS



Software package



Web Application

www.omdbapi.com

OMDb API - The Open Movie Database

Garrett

www.omdbapi.com

OMDb API Usage Parameters Examples Change Log Donors ▾

Donate Contact

OMDb API

The Open Movie Database

The OMDb API is a free web service to obtain movie information, all content and images on the site are contributed and maintained by our users.

If you find this service useful, please consider [donating](#).



New Poster API

The Poster API is currently in a closed beta being tested by the sites friendly donors.

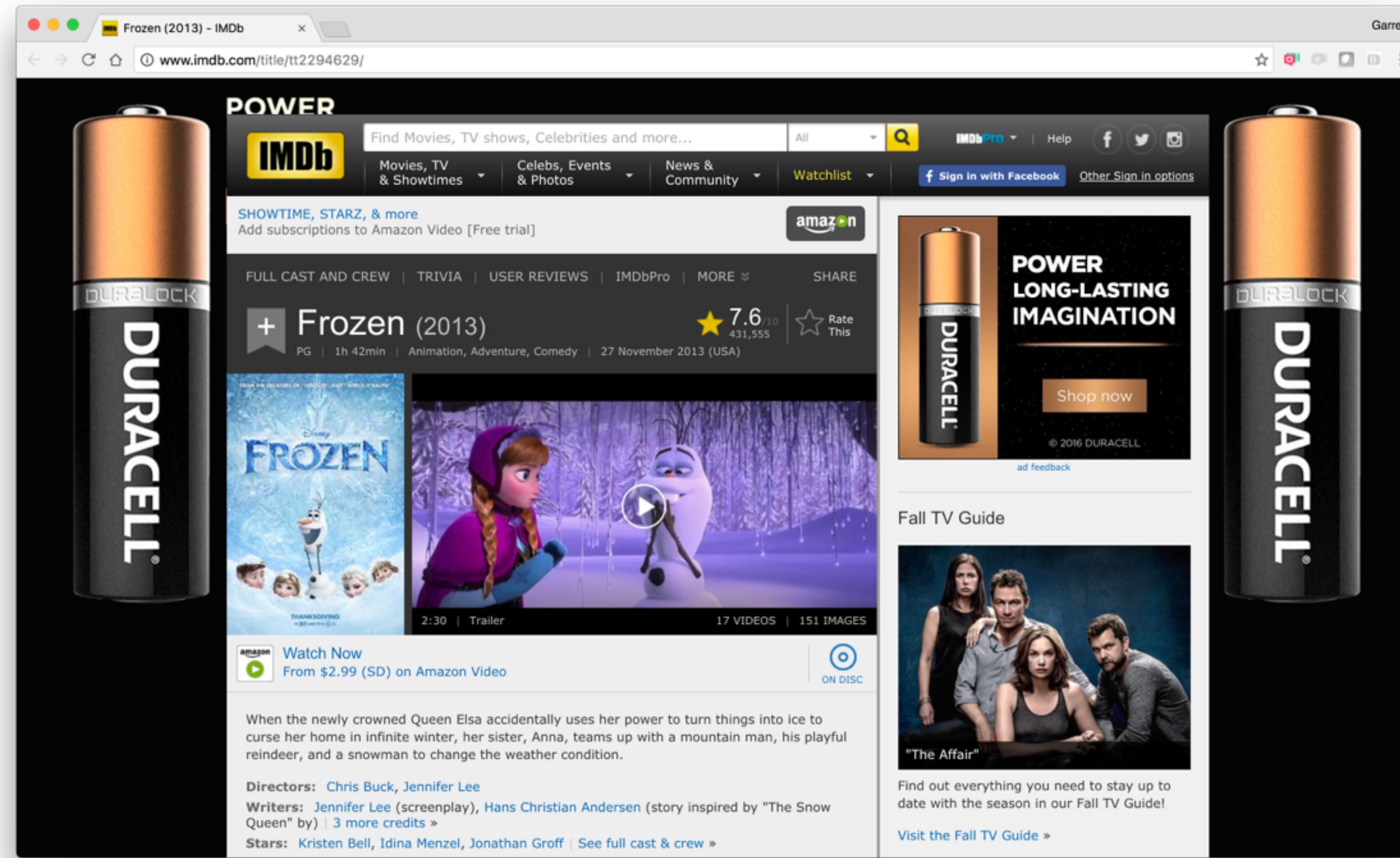
Donors get your key today and start testing!

Donations are primarily used for server fees and hosting. Once we get enough donations to afford a second server the Poster API will be open publicly.

Leave Feedback



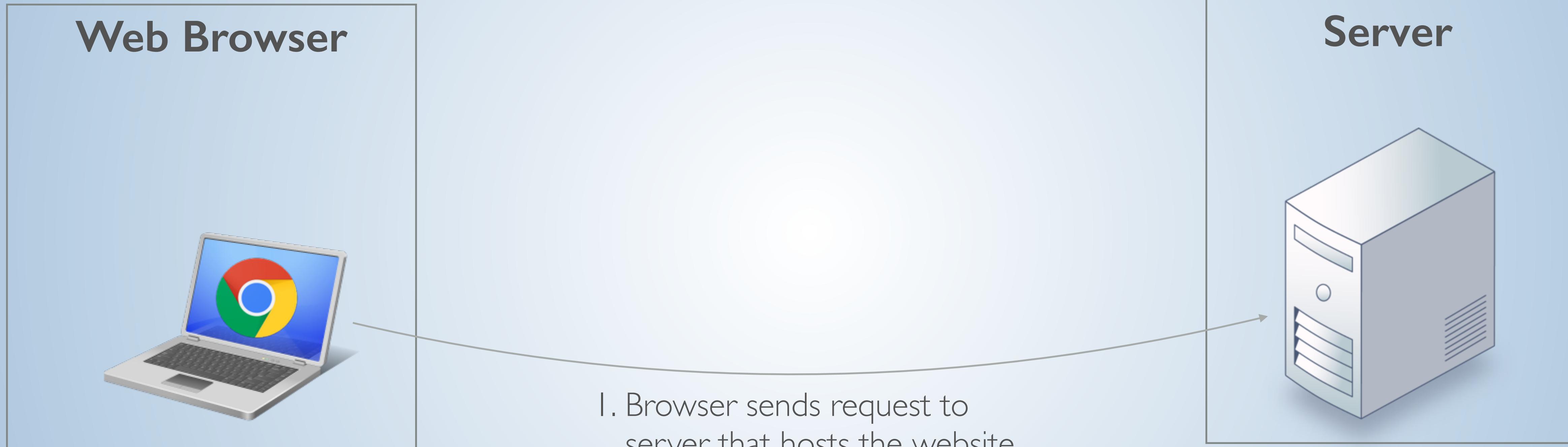
www.imdb.com



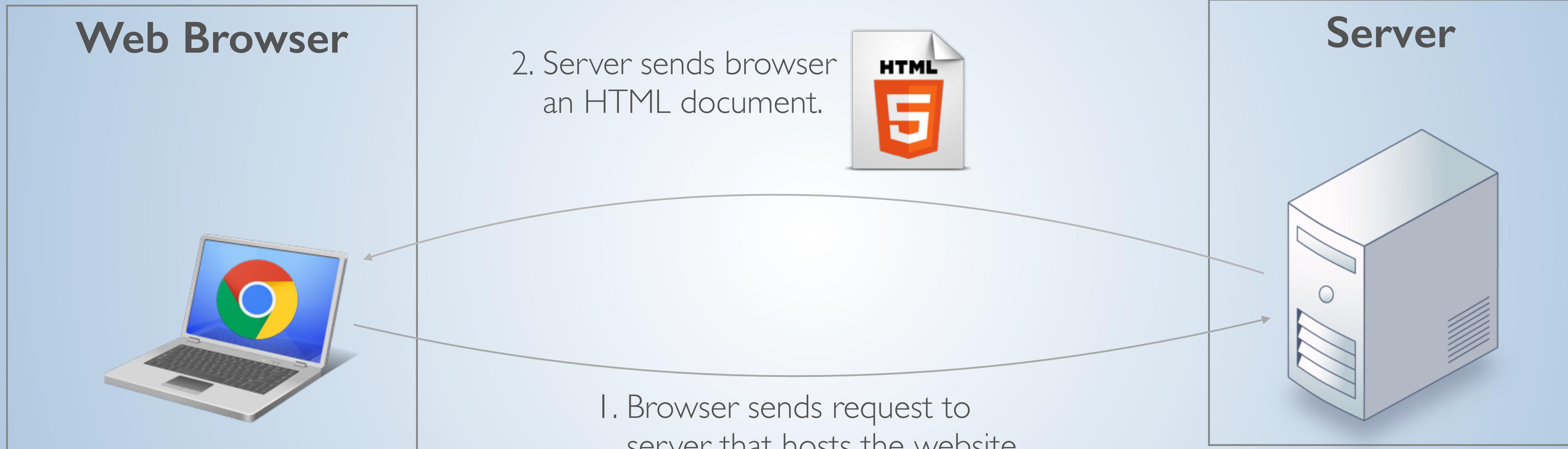
No API



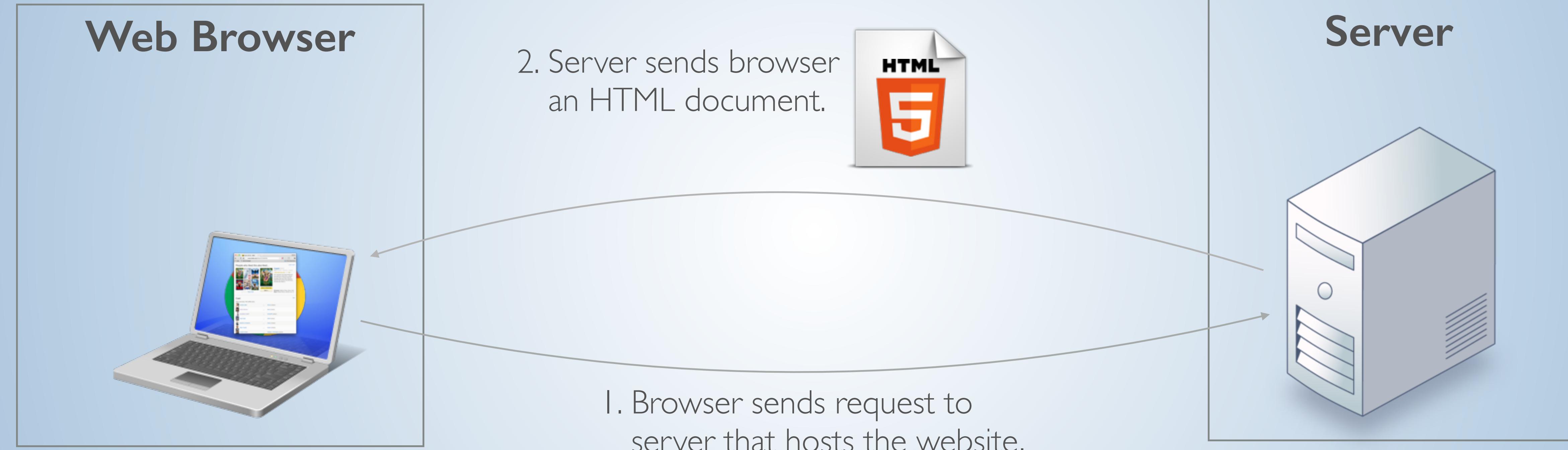
HTML (Review)



HTML (Review)



HTML (Review)



3. Browser uses instructions in HTML to render the website

HTML (Review)

Web Browser



3. Browser uses instructions
HTML to render the web!

```
<html>
  <head>
    <title>Title</title>
    <link rel="icon" type="icon" href="http://a" />
    <link rel="icon" type="icon" href="http://b" />
    <script type="text/javascript">
      var ue_t0=window.ue_t0||+new Date();
    </script>
  </head>
  <body>
    <div>
      <p>Click <b>here</b> now.</p>
      <span>Frozen</span>
    </div>
    <table style="width:100%">
      <tr>
        <td>Kristen</td>
        <td>Bell</td>
      </tr>
      <tr>
        <td>Idina</td>
        <td>Menzel</td>
      </tr>
    </table>
    
  </body>
</html>
```

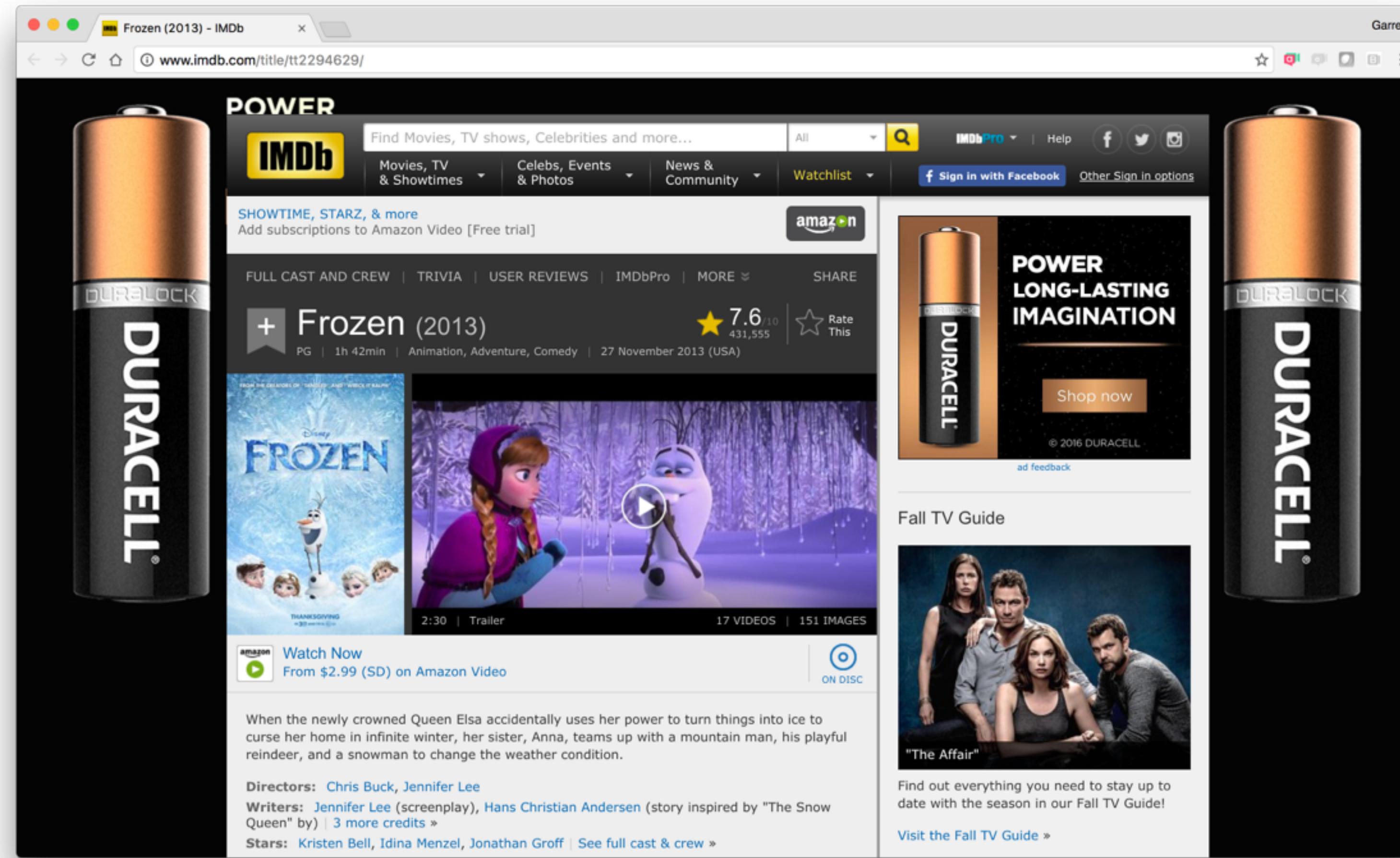


quest to
the website.

Server



www.imdb.com



No API



Frozen (2013) - IMDb

www.imdb.com/title/tt2294629/

Garrett

Apps SelectorGadget Other Bookmarks

People who liked this also liked...

Tangled (2010)
PG Animation | Adventure | Comedy
★★★★★ 7.8/10

The magically long-haired Rapunzel has spent her entire life in a tower, but now that a runaway thief has stumbled upon her, she is about to discover the world for the first time, and who she really is.

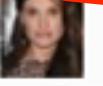
Add to Watchlist

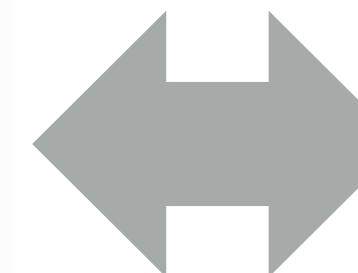
Next »

◀ Prev 6 ▶

Cast

Cast overview, first billed only:

 Kristen Bell	... Anna (voice)
 Idina Menzel	... Elsa (voice)
 Jonathan Groff	... Kristoff (voice)
 Josh Gad	... Olaf (voice)
 Santino Fontana	... Hans (voice)
 Alan Tudyk	... Duke (voice)
 Ciarán Hinds	... Pabbie / Grandpa (voice)



Frozen (2013) - IMDb

view-source:www.imdb.com/title/tt2294629/

Garrett

Apps SelectorGadget Other Bookmarks

```
3933
3934
3935     <table class="cast_list">
3936     <tr><td colspan="4" class="castlist_label">Cast
3937     overview, first billed only:</td></tr>
3938     <tr class="odd">
3939         <td class="primary_photo">
3940             <a href="/name/nm0068338/?ref_=tt_cl_i1"
3941                 >
3949         </td>
3950         <td class="itemprop itemprop=actor"
3951             itemscope itemtype="http://schema.org/Person">
3952             <a href="/name/nm0068338/?ref_=tt_cl_t1"
3953                 itemprop='url'><span class="itemprop"
3954                 itemprop="name">Kristen Bell</span>
3955             </a>
3956         <td class="ellipsis">
3957             ...
3958         </td>
3959         <td class="character">
3960             <div>
3961                 <a href="/character/ch0307445/?ref_=tt_cl_t1" >Anna</a>
3962             </div>
3963             (voice)
3964         </td>
3965     </tr>
```



Strategy

1. Extract information from the HTML DOM
2. Use the structure of HTML to find the information



HTML (Review)

Each element in the page is created by a tag.

```
<a href="http://github.com">GitHub</a>
```

tag name

attribute
(name)

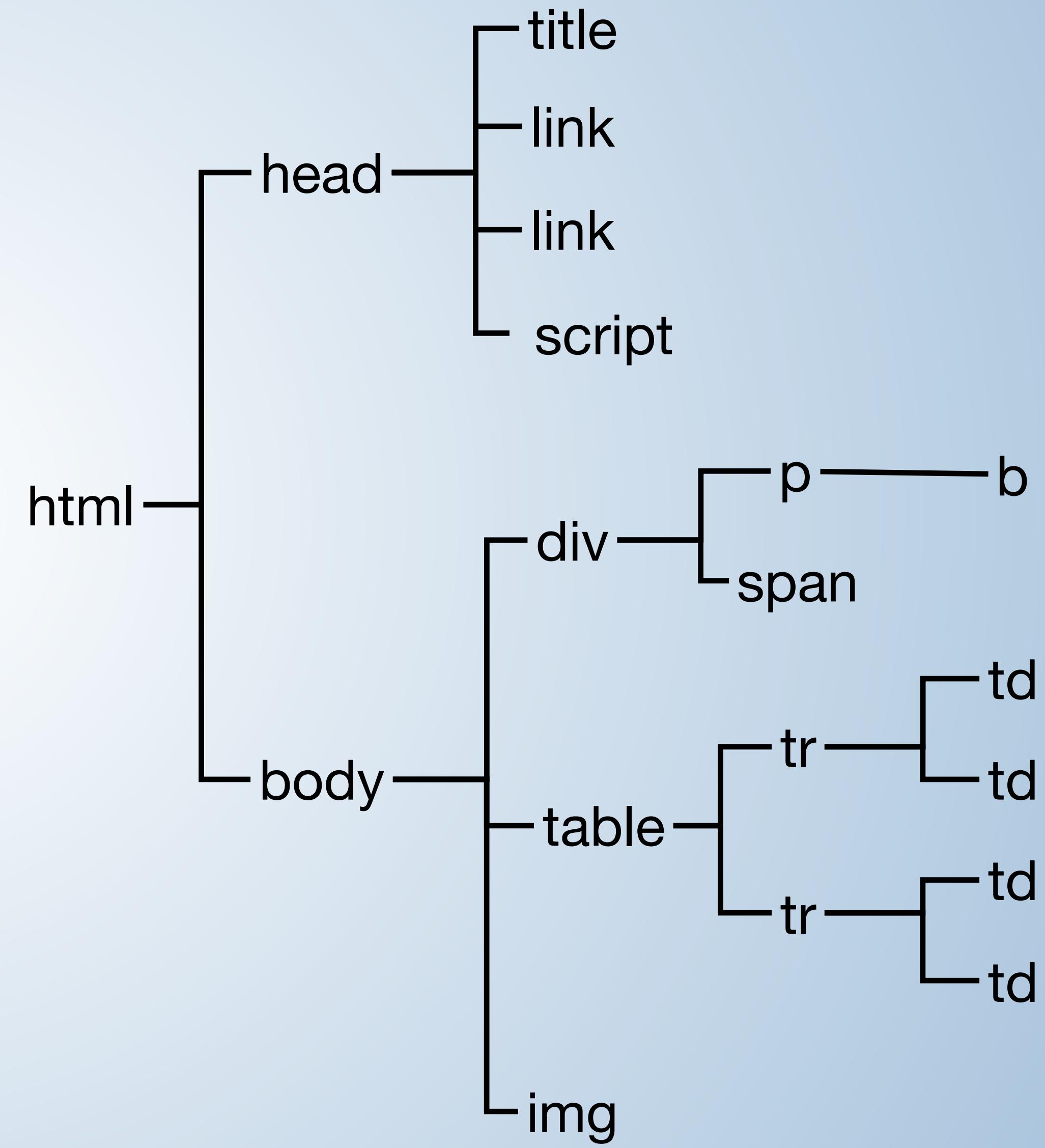
attribute
(value)

content

HTML (Review)



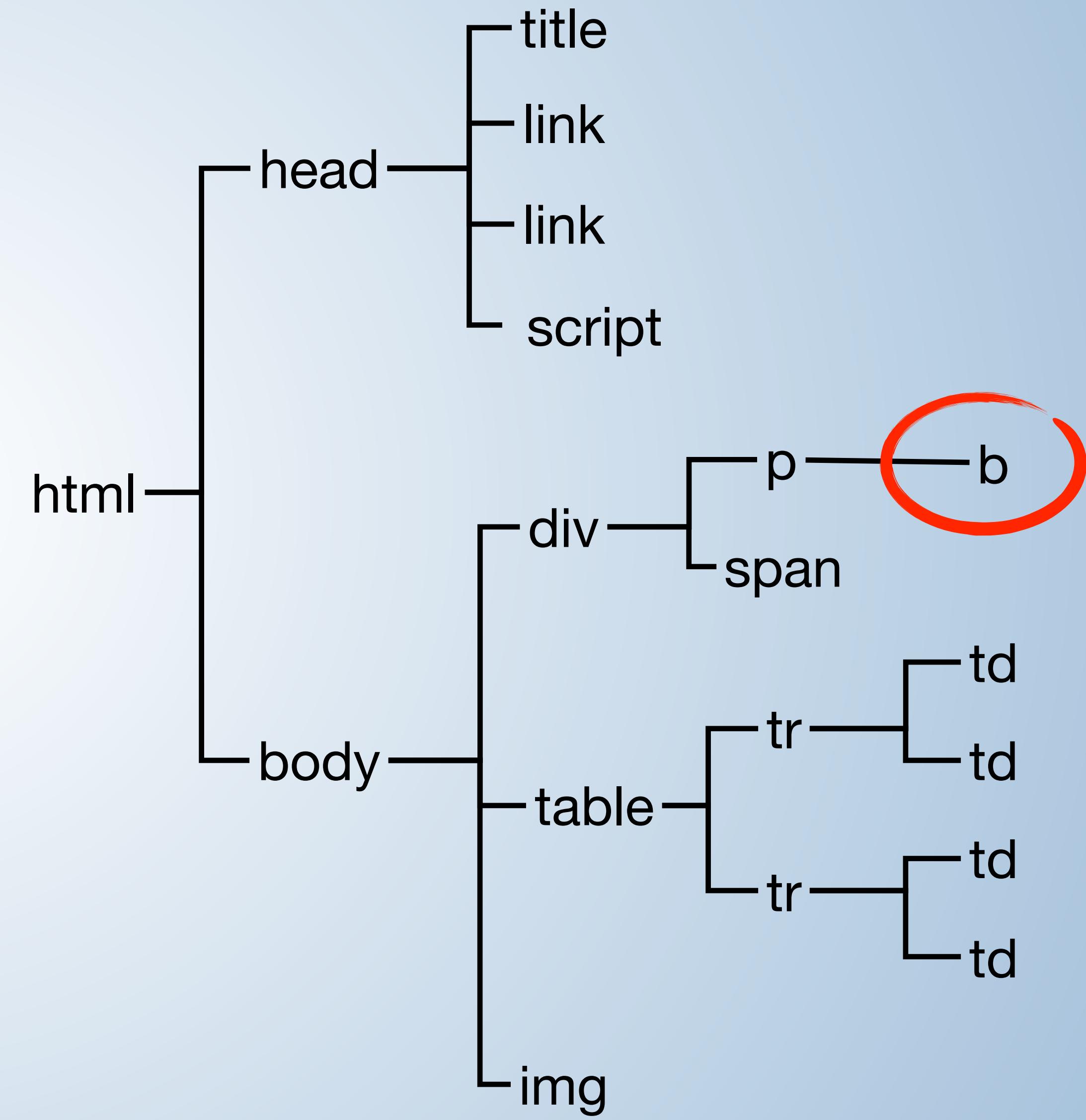
```
<html>
  <head>
    <title>title</title>
    <link rel="icon" type="icon" href="http://a" />
    <link rel="icon" type="icon" href="http://b" />
    <script type="text/javascript">
      var ue_t0=window.ue_t0||+new Date();
    </script>
  </head>
  <body>
    <p>21 <b>here</b> now.</p>
    <div>
      <span>From zero</span>
    </div>
    <table style="width:100%">
      <tr>
        <td>Kristen</td>
        <td>Bell</td>
      </tr>
      <tr>
        <td>Lina</td>
        <td>Menzel</td>
      </tr>
    </table>
    
  </body>
</html>
```



HTML (Review)



```
<html>
  <head>
    <title>Title</title>
    <link rel="icon" type="icon" href="http://a" />
    <link rel="icon" type="icon" href="http://b" />
    <script type="text/javascript">
      var ue_t0=window.ue_t0||+new Date();
    </script>
  </head>
  <body>
    <div>
      <p>Click <b>here</b> now.</p>
      <span>Frozen</span>
    </div>
    <table style="width:100%">
      <tr>
        <td>Kristen</td>
        <td>Bell</td>
      </tr>
      <tr>
        <td>Idina</td>
        <td>Menzel</td>
      </tr>
    </table>
    
  </body>
</html>
```



Which HTML tag surrounds Kristin Bell's name?



Which HTML tag surrounds Kristin Bell's name?

span



A screenshot of a browser window titled "view-source:www.imdb.com". The address bar shows "view-source:www.imdb.com/title/tt2294629/". The content area displays the raw HTML code for a movie title page. A large blue arrow points from the word "span" at the top left to the word "Kristen" in the highlighted line of code. The line of code is:

```
3943     itemprop='url'> <span class="itemprop" itemprop="name">Kristen Bell</span>
```

The code also includes other elements like td, a, and div tags.

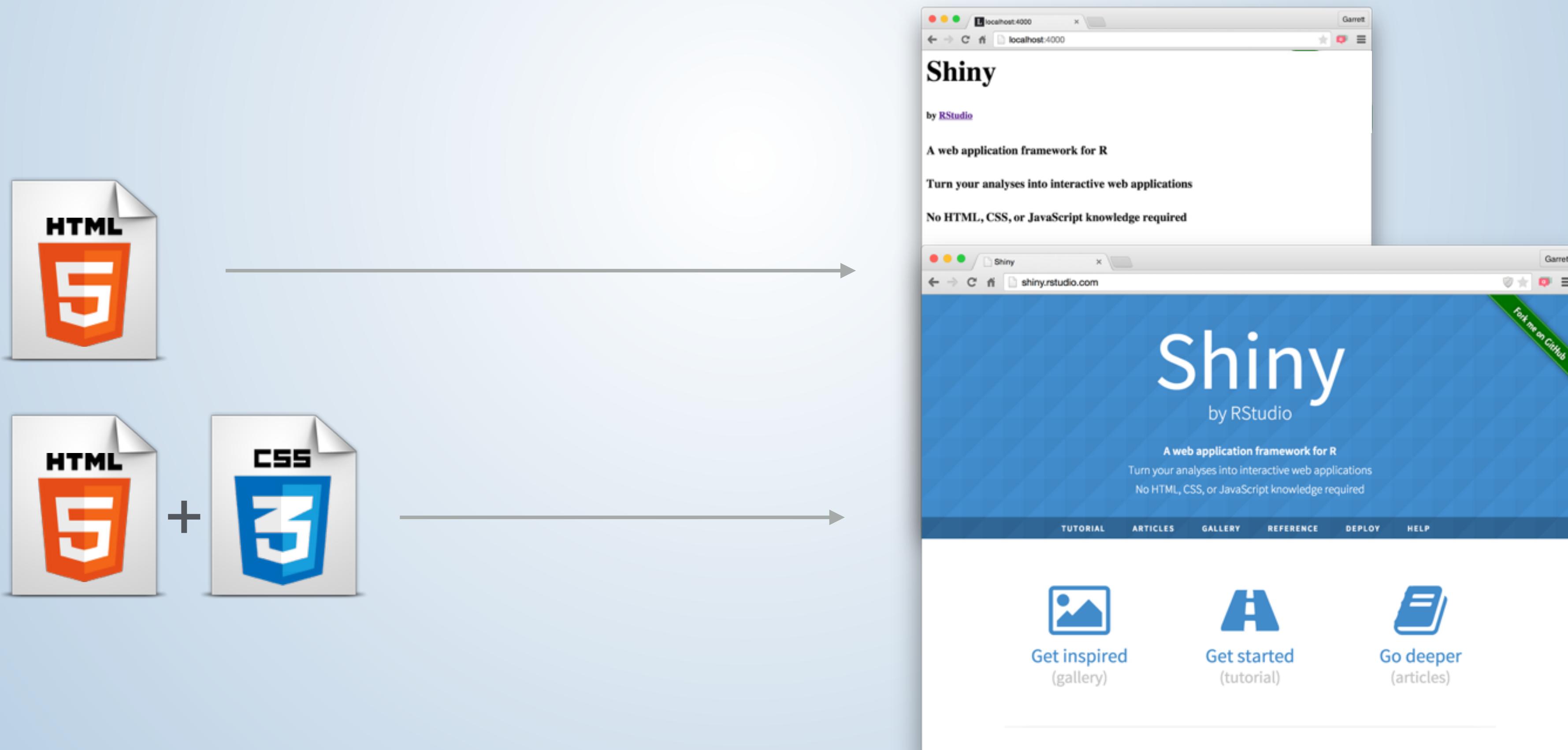
But there are ~600
Span tags



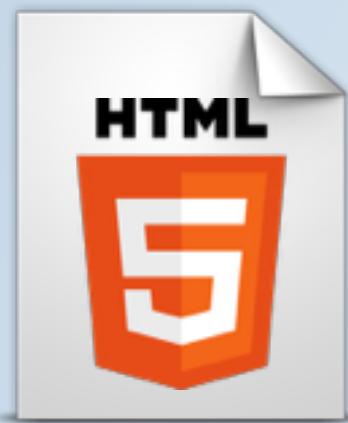
css selectors

CSS (Review)

Cascading Style Sheets (CSS) are a framework for customizing the appearance of elements in a web page.



CSS (Review)



Shiny

by RStudio

A web application framework for R

Turn your analyses into interactive web applications

No HTML, CSS, or JavaScript knowledge required

- [Tutorial](#)
- [Articles](#)
- [Gallery](#)
- [Reference](#)
- [Deploy](#)
- [Help](#)

[Get inspired](#)
(gallery)

[Get started](#)
(tutorial)

[Go deeper](#)
(articles)



Shiny

by RStudio

A web application framework for R

Turn your analyses into interactive web applications

No HTML, CSS, or JavaScript knowledge required

TUTORIAL ARTICLES GALLERY REFERENCE DEPLOY HELP

Get inspired
(gallery)

Get started
(tutorial)

Go deeper
(articles)

Fork me on GitHub

CSS (Review)



```
span {  
    color: #ffffff;  
}  
  
.num {  
    color: #a8660d;  
}  
  
table.data {  
    width: auto;  
}  
  
#firstname {  
    background-color: yellow;  
}
```

← selector

← styling

← selector

← styling

← selector

← styling

← selector

← styling

CSS (Review)

A CSS script describes an element by its tag, class, and/or ID.

```
<span class="bigname" id="shiny">Shiny</span>
```

tag name

class
(optional)

id
(optional)

CSS (Review)

A CSS script describes an element by its tag, class, and/or ID.

```
<span class="bigname" id="shiny">Shiny</span>
```

```
span
```

CSS selector for **ALL** elements with:

- the **span tag**

CSS (Review)

A CSS script describes an element by its tag, class, and/or ID.

```
<span class="bigname" id="shiny">Shiny</span>
```

```
.bigname
```

CSS selector for **ALL** elements with:

- the **bigname class**

CSS (Review)

A CSS script describes an element by its tag, class, and/or ID.

```
<span class="bigname" id="shiny">Shiny</span>
```

```
span.bigname
```

CSS selector for **ALL** elements with:

- the **span tag**

AND

- the **bigname class**

CSS (Review)

A CSS script describes an element by its tag, class, and/or ID.

```
<span class="bigname" id="shiny">Shiny</span>
```

```
#shiny
```

CSS selector for **ALL** elements with:

- the **shiny id**

CSS (Review)



```
span {  
    color: #ffffff;  
}  
  
.num {  
    color: #a8660d;  
}  
  
table.data {  
    width: auto;  
}  
  
#firstname {  
    background-color: yellow;  
}
```

← selector

← styling

← selector

← styling

← selector

← styling

← selector

← styling

CSS (*Review*)

Prefix	Matches
none	tag
.	class
#	id

Which CSS identifiers are associated
with Kristen Bell's name?



Which CSS identifiers are associated with Kristen Bell's name?

span (the element)

itemprop (the class)



```
view-source:www.imdb.com
view-source:www.imdb.com/title/tt2294629/
Apps SelectorGadget Other Bookmarks
3941      <td class="itemprop" itemprop="actor" itemscope
3942      itemtype="http://schema.org/Person">
3943      <a href="/name/nm0068338/?ref_=tt_cl_t1"
3944          itemprop='url'> <span class="itemprop" itemprop="name">Kristen Bell</span>
3945      </a>
3946      <td class="ellipsis">
3947          ...
3948      </td>
3949      <td class="character">
            <div>
```



Which CSS identifiers are associated with Kristen Bell's name?

span (the element)

itemprop (the class)

```
view-source:www.imdb.com
view-source:www.imdb.com/title/tt2294629/
SelectorGadget
Garrett
Apps Other Bookmarks
3941 <td class="itemprop" itemprop="actor" itemscope
itemtype="http://schema.org/Person">
3942 <a href="/name/nm068338/?ref_=tt_cl_t1"
3943 itemprop='url'> <span class="itemprop" itemprop="name">Kristen Bell</span>
3944 </a>
3945 <td class="ellipsis">
3946 ...
3947 </td>
3948 <td class="character">
3949 <div>
```

Span.itemprop



Recap

1. Extract information from the HTML DOM
2. Use the structure of HTML **and CSS** to find the information



rvest

rvest



A package that makes it easy to extract
info from a webpage.

```
install.packages("rvest")
```

* This will also install *xml2*, a package that *rvest* relies on.



Basic Workflow

- 1.** Download the HTML and turn it into an XML file with `read_html()`
- 2.** Extract specific nodes with `html_nodes()`
- 3.** Extract content from nodes with `html_text()`,
`html_name()`, `html_attrs()`, `html_children()`,
`html_table()`



Basic Workflow

1. Download the HTML and turn it into an XML file with `read_html()`

```
library(rvest)  
frozen <- read_html("http://www.imdb.com/title/tt2294629/")
```

read_html

URL

* `read_html()` comes in the `xml2` package loaded with `rvest`



Basic Workflow

1. Download the HTML and turn it into an XML file with `read_html()`

2. Extract specific nodes with `html_nodes()`

```
cast <- html_nodes(frozen, "span.itemprop")
```

XML

CSS selector



Basic Workflow

1. Download the HTML and turn it into an XML file with `read_html()`
2. Extract specific nodes with `html_nodes()`
3. Extract content from nodes with `html_text()`,
`html_name()`, `html_attrs()`, `html_children()`,
`html_table()`



```
library(rvest)
frozen <- read_html("http://www.imdb.com/title/tt2294629/")
cast <- html_nodes(frozen, "span.itemprop")
html_text(cast)

## [1] "Animation"           "Adventure"
## [3] "Comedy"              "Chris Buck"
## [5] "Jennifer Lee"        "Jennifer Lee"
## [7] "Hans Christian Andersen" "Kristen Bell"
## [9] "Idina Menzel"         "Jonathan Groff"
## [11] "Kristen Bell"          "Idina Menzel"
## [13] "Jonathan Groff"       "Josh Gad"
## [15] "Santino Fontana"       "Alan Tudyk"
## [17] "Ciarán Hinds"          "Chris Williams"
## [19] "Stephen J. Anderson"   "Maia Wilson"
## [21] "Edie McClurg"          "Robert Pine"
## [23] "Maurice LaMarche"      "Livvy Stubenrauch"
## [25] "Eva Bella"             "snowman"
## [27] "sister love"           "sister sister relationship"
## [29] "magic"                 "snow"
## [31] "Walt Disney Animation Studios" "Walt Disney Pictures"
```

We've scraped
too much info



tables

www.bestplaces.net

Best Places to Live in Orlando | Garrett

www.bestplaces.net/city/florida/orlando

About Bert Sperling | Bert's Blog | [f](#) [t](#) [G+](#) [Y](#) [i](#)

sperling's
BEST PLACES Info on Cost of Living, Schools, Crime
Rates, House Prices, and more...

Enter a City, Town, Or Z **Search** Find A Place

Home Page Best Places Quiz Find a Place City Rankings Research Tools Top Lists Sign Up Log In

CHOOSE FROM OVER 600 HOTELS WORLDWIDE. HYATT BOOK NOW ▶

Home / United States / Florida / Orlando-Kissimmee-Sanford Metro Area / Orange County / Orlando / Zip Codes

Orlando, Florida

221 Reviews | Leave a Comment | Add Favorite

Use This Checklist:

- Place Overview
- Cost of Living
- Crime Rates
- School Ratings
- Education Stats
- Climate
- Economy
- Health
- Religion
- People Stats

270,934 Up 64.5%
[Population](#)

33 Median Age

38.7% Married Population

2.4 Household Size

3.9% Unemployment Rate

\$157,600 Median Home Price

24.55 minutes Average Commute Time

Real Estate For Sale For Rent

Side-by-Side Comparison

Compare Orlando, Florida to any other place in the USA.

Compare Now

Ask A Orlando Expert

 Angel C.
[How To Become An Expert Here](#)

Real Estate in Orlando

Newly Listed Homes

HOMES UNDER \$250,000



Tables

Use `html_table()` to scrape whole tables of data as a data frame.

```
url <- "http://www.bestplaces.net/climate/city/florida/orlando"  
orlando <- read_html(url)  
tables <- html_nodes(orlando, css = "table")  
html_table(tables, header = TRUE)[[2]]
```



selectorGadget

```
library(rvest)
frozen <- read_html("http://www.imdb.com/title/tt2294629/")
cast <- html_nodes(frozen, "span.itemprop")
html_text(cast)

## [1] "Animation"                      "Adventure"
## [3] "Comedy"                          "Chris Buck"
## [5] "Jennifer Lee"                   "Jennifer Lee"
## [7] "Hans Christian Andersen"       "Kristen Bell"
## [9] "Idina Menzel"                   "Jonathan Groff"
## [11] "Kristen Bell"                   "Idina Menzel"
## [13] "Jonathan Groff"                "Josh Gad"
## [15] "Santino Fontana"                "Alan Tudyk"
## [17] "Ciarán Hinds"                  "Chris Williams"
## [19] "Stephen J. Anderson"           "Maia Wilson"
## [21] "Edie McClurg"                  "Robert Pine"
## [23] "Maurice LaMarche"              "Livvy Stubenrauch"
## [25] "Eva Bella"                     "snowman"
## [27] "sister love"                   "sister sister relationship"
## [29] "magic"                          "snow"
## [31] "Walt Disney Animation Studios" "Walt Disney Pictures"
```

We've scraped
too much info



selectorGadget

A GUI tool to identify CSS selector combinations

Clear (1) Toggle Position XPath Help X



To Install

1. Run `vignette("selectorgadget")`
2. Drag **SelectorGadget** link into your browser's bookmark bar
3. Or visit SelectorGadget.com

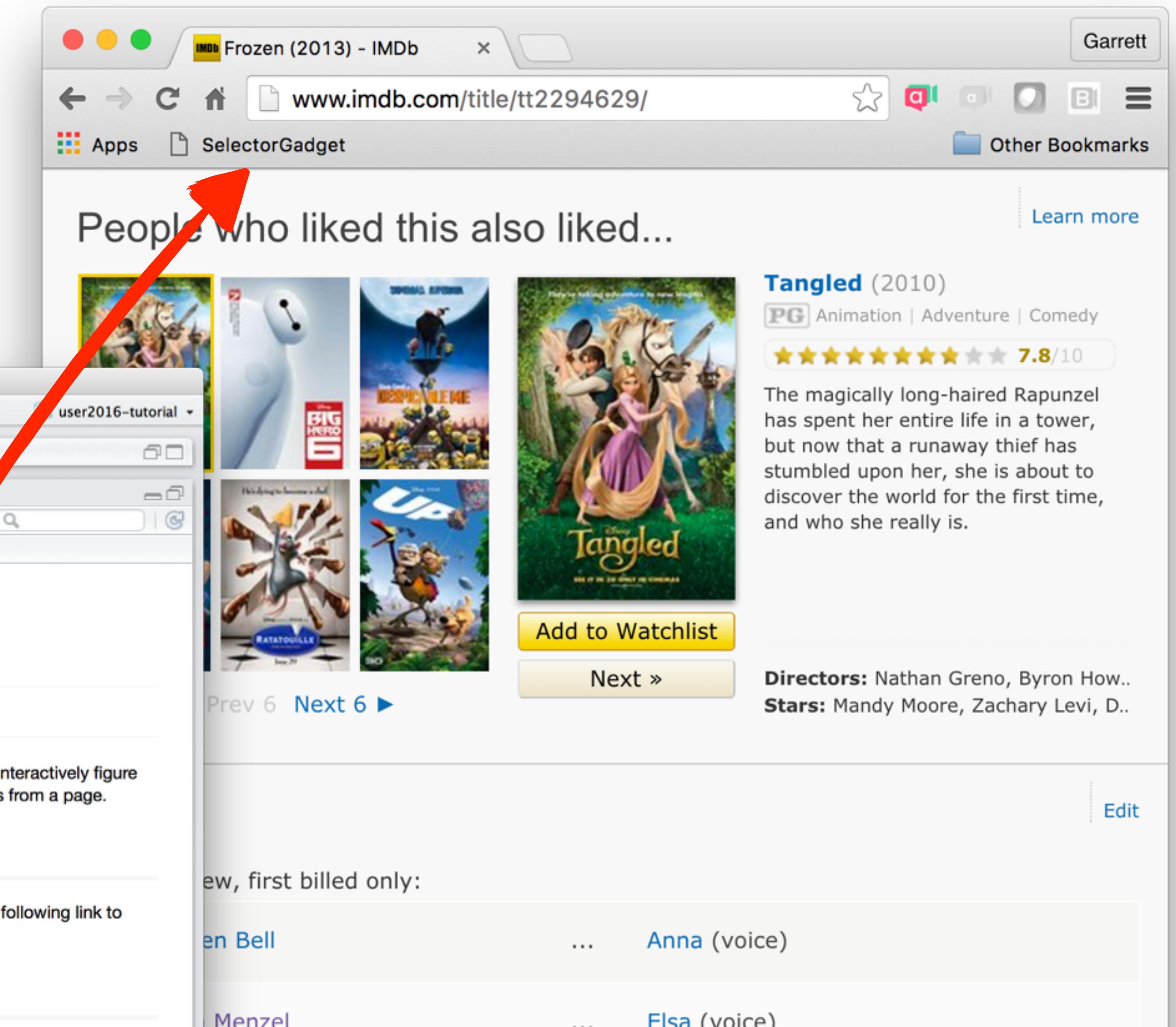
The screenshot shows the RStudio interface. In the top-left pane, there is an R script named 'scraping-outline.Rmd' with the following code:

```
1 # Read in web page
2
3 library(rvest)
4 frozen <- read_html("http://www.imdb.com/title/tt2294629/")
5
6 # Look at web page
7
8 frozen
9 html_structure(frozen)
10 as_list(frozen)
11
12 xml_children(frozen)
13 xml_children(frozen)[[2]]
```

In the bottom-left pane, the console shows the command:

```
> vignette("selectorgadget")
```

In the main pane, there is a section titled "Selectorgadget" by Hadley Wickham (2016-06-16). It contains a description of the SelectorGadget bookmarklet and an "Installation" section. The "Installation" section includes the text: "To install it, open this page in your browser and then drag the following link to your bookmark bar: [SelectorGadget](#)". This text is circled with a red oval.



To Use

1. **Navigate** to a webpage
2. **Open** the SelectorGadget bookmark
3. **Click** on item to scrape
4. **Click** on yellow items You do not want to scrape
5. **Click** on additional items that you do want to scrape
6. **Copy** selector to use with html_nodes()



.fa-bolt

Clear (1) Toggle Position XPath Help X

CSS selector to use

start over

move gadget

show XPath

help

close gadget



```
cast2 <- html_nodes(frozen, "#titleCast span.itemprop")
html_text(cast2)
```

```
cast3 <- html_nodes(frozen, ".itemprop .itemprop")
html_text(cast3)
```

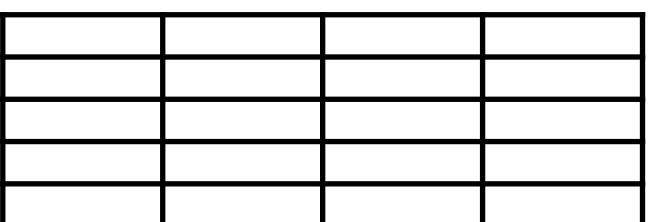


Recap

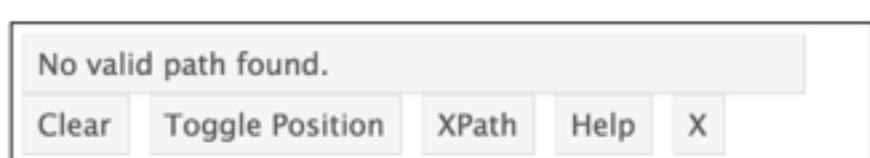
Pull from HTML, use HTML/CSS structure



`read_html()` → `html_nodes()` → `html_text()`, etc.



`html_table()` for tables



`selectorGadget` for finding useful selector combinations



Thank You