Web Scraping mit R

Stefan Schmidt

16.04.2020

Agenda

- Web Scraping
- Review: HTML[1] und CSS[2] Selektoren
- R Package: rvest
- Werkzeug: selectorGadget
- Rechtliche Bedenken

[1] Hypertext Markup Language
[2] Cascading Style Sheets

Web Scraping

Web Scraping

Was ist das?

- engl. “aus dem Web kratzen” (auch “Web Ha**rvest**ing”)
- Verfahren zum automatisierten Auslesen von Texten / Daten von Webseiten (HTML)

Web Scraping

Wann wendet man es an?

- Wenn zur Extraktion der Informationen keine API [3] zur Verfügung steht
- Also keine Möglichkeit besteht definiert und sauber (z.B. via Web Service, Connector, SQL...) auf die Information zuzugreifen

[3] Application Programming Interface

Web Scraping

Die Idee dahinter

- Informationen aus dem HTML Dokument extrahieren (Achtung: Schnell sehr umfangreich!)
- mit Hilfe von CSS Selektoren die interessanten Information identifizieren

Web Scraping

In R:

- steht mit **rvest** im tidyverse bereits ein Package bereit

Review: HTML und CSS Selektoren

Extracting data from the web

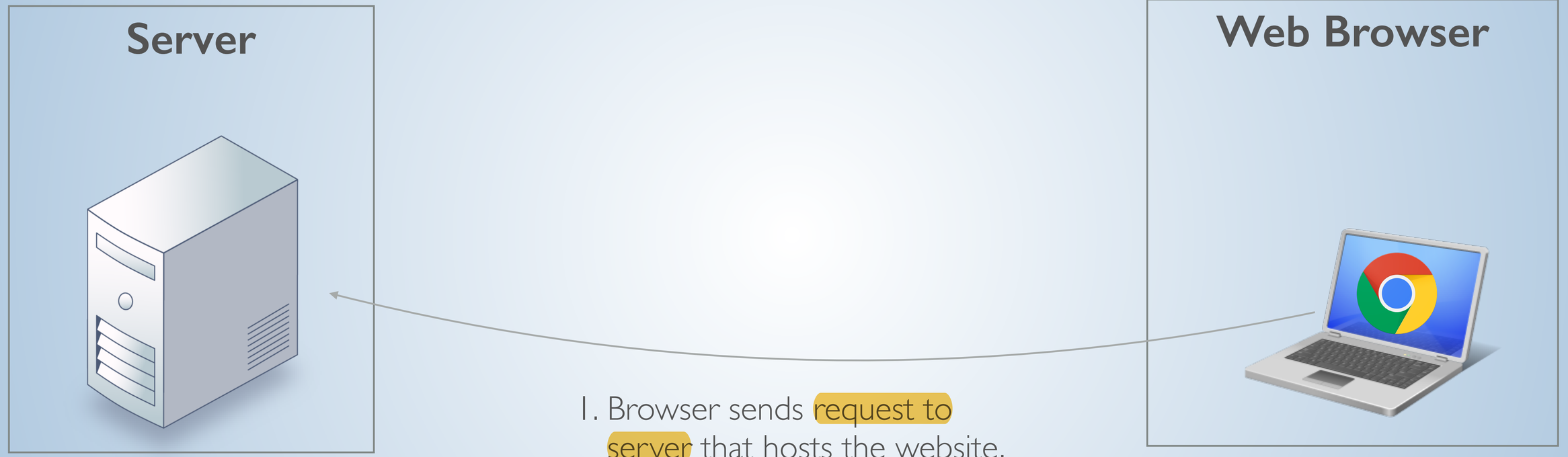
APIs and beyond

Scott Chamberlain
Karthik Ram
Garrett Grolemund

June 2016

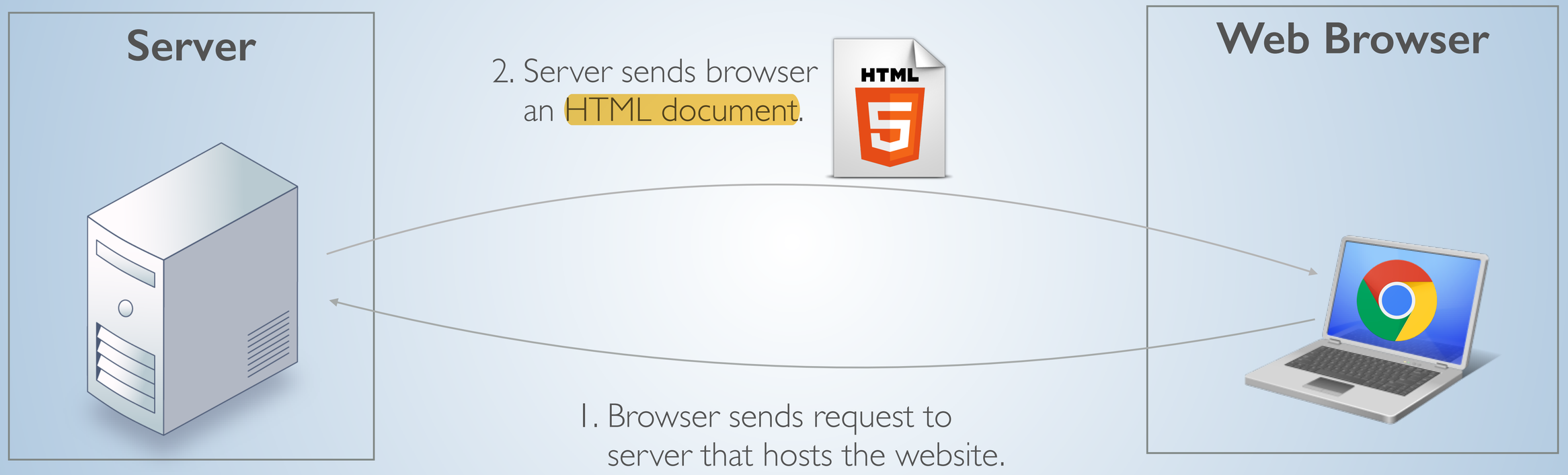
USER 2016 CONFERENCE

HTML (Review)



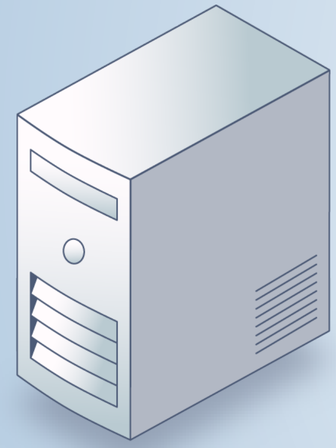
1. Browser sends request to server that hosts the website.

HTML (Review)


Server
2. Server sends browser an HTML document.
HTML
Web Browser
1. Browser sends request to server that hosts the website.

HTML (Review)

Server



2. Server sends browser an HTML document.

Web Browser

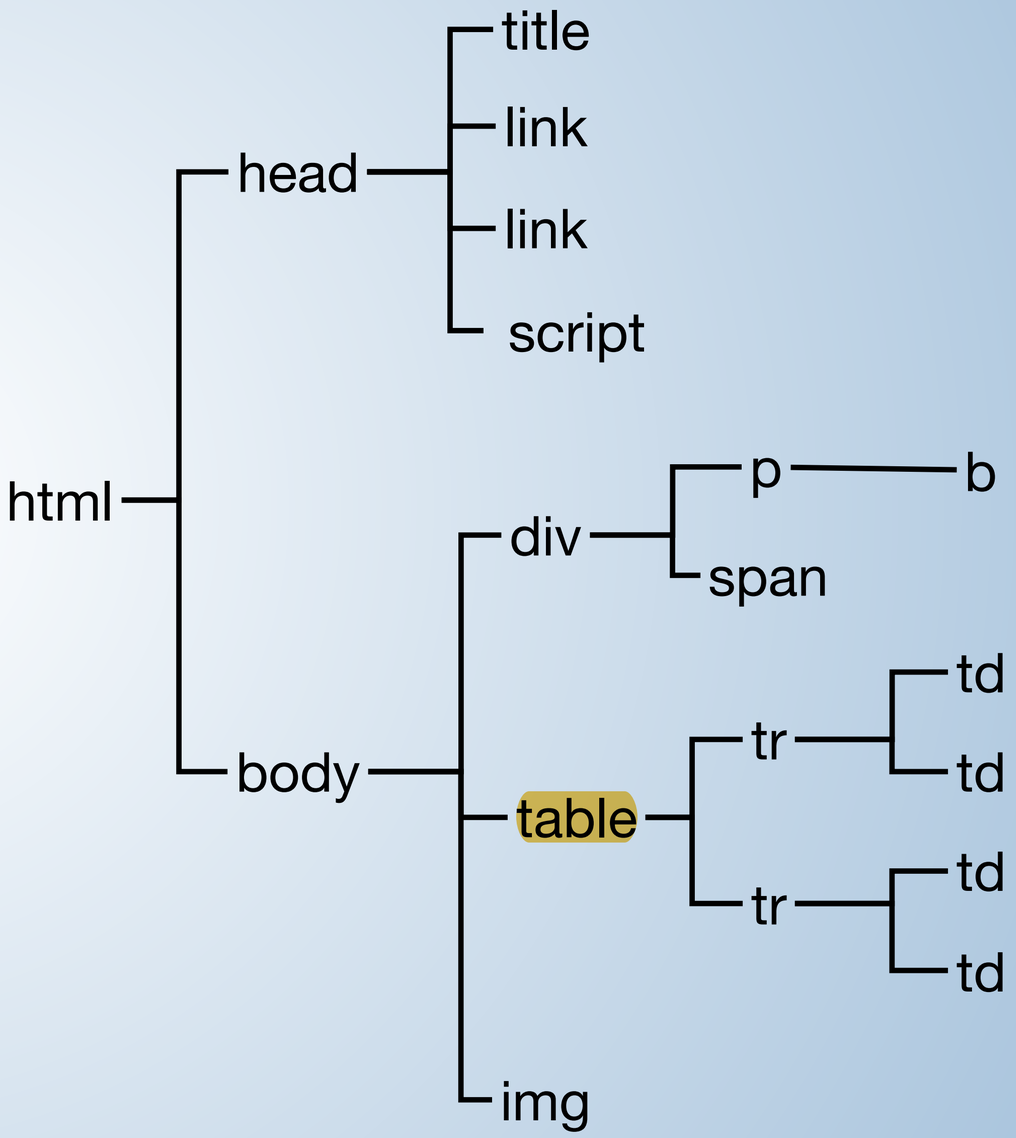
1. Browser sends request to server that hosts the website.

3. Browser uses instructions in HTML to render the website

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>
    <img src="http://ia.media-imdb.com/images.png"/>
  </body>
</html>
```

HTML Dokument (vereinfacht)

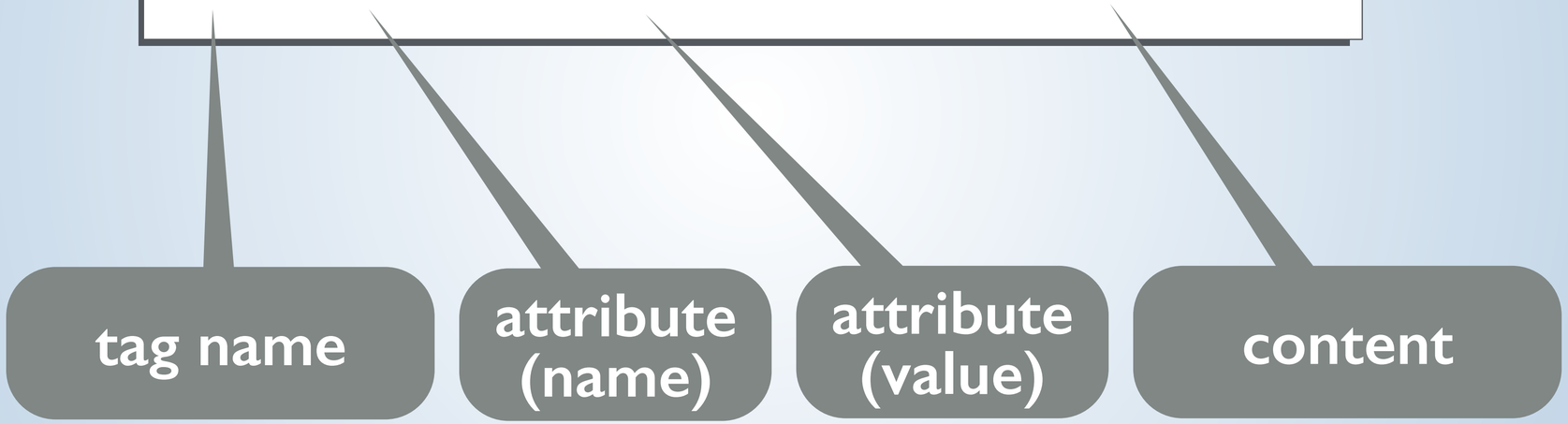


Baumstruktur der HTML Elemente

HTML (Review)

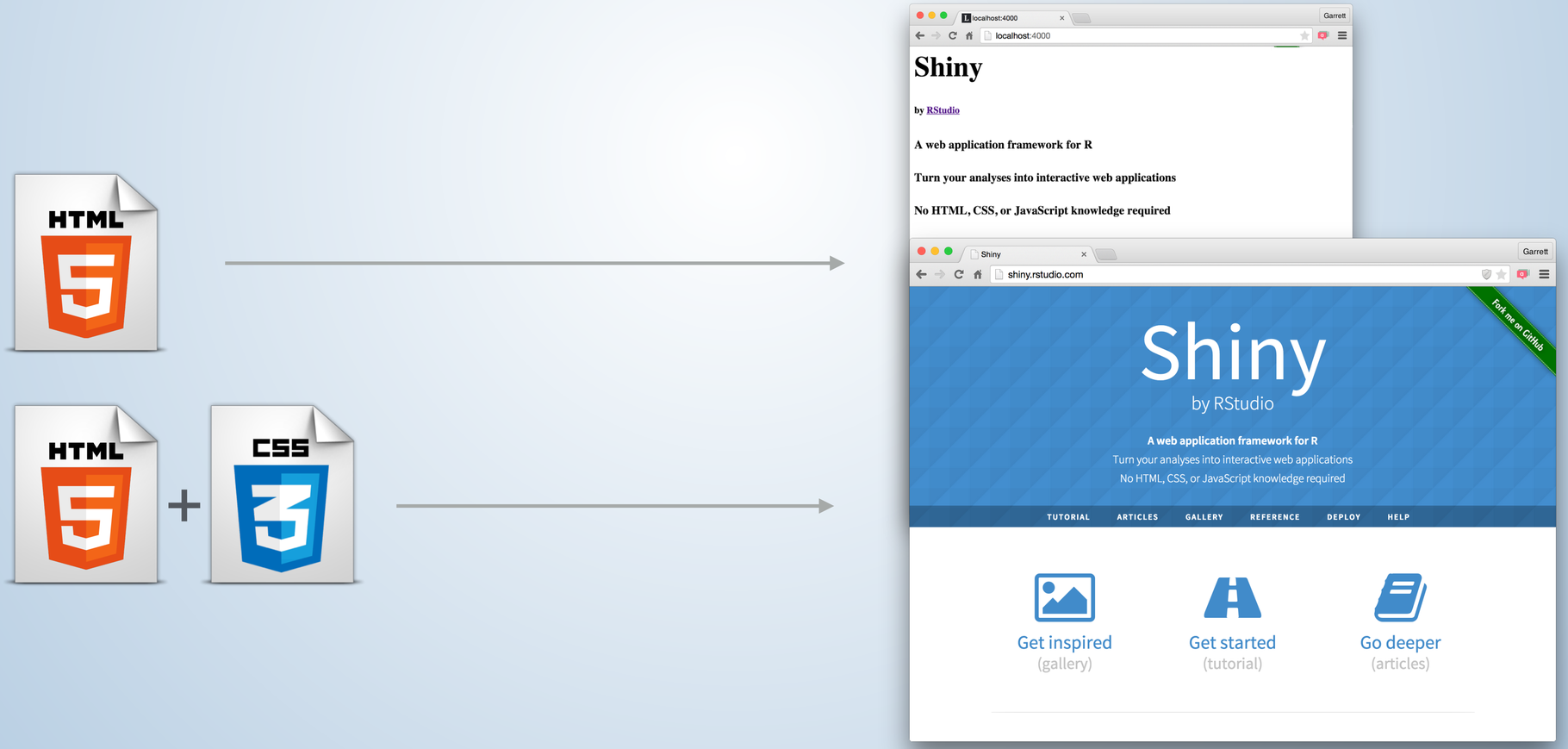
Each element in the page is created by a tag.

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



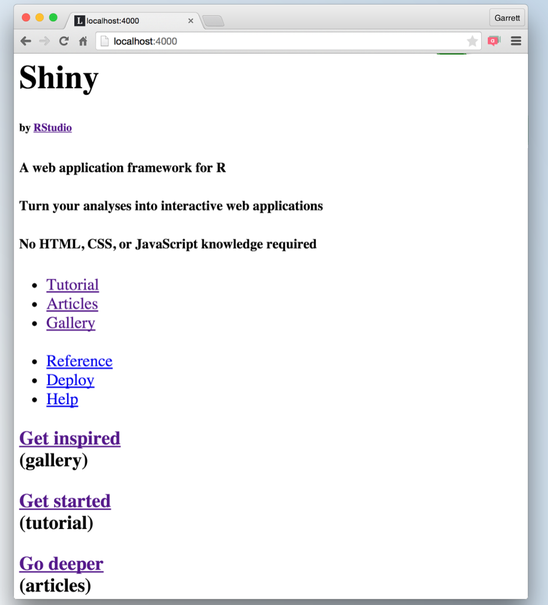
CSS (Review)

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

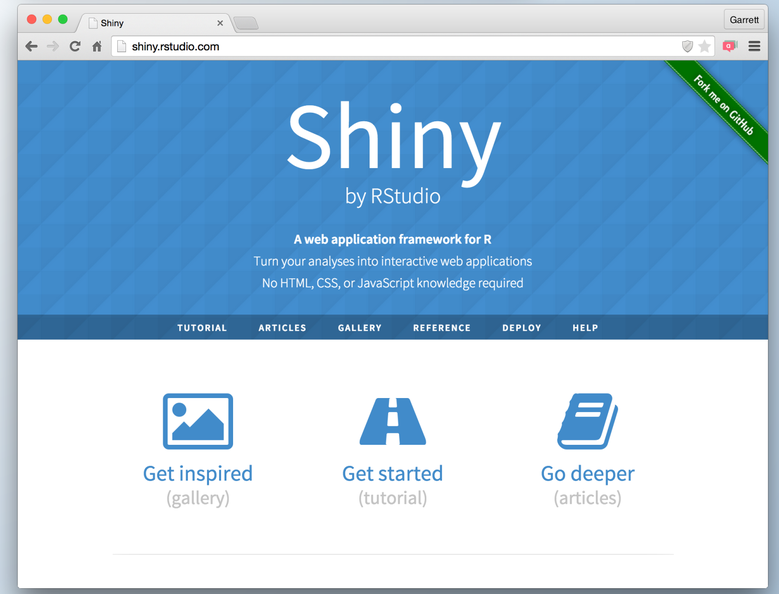


CSS (Review)


HTML


localhost:4000
Garrett
localhost:4000
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)


HTML
CSS


Shiny
Garrett
shiny.rstudio.com
Fork me on GitHub
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)

CSS (Review)


```
span {
  color: #ffffff;
}

.num {
  color: #a8660d;
}

table.data {
  width: auto;
}

#firstname {
  background-color: yellow;
}
```

selector
styling

selector
styling

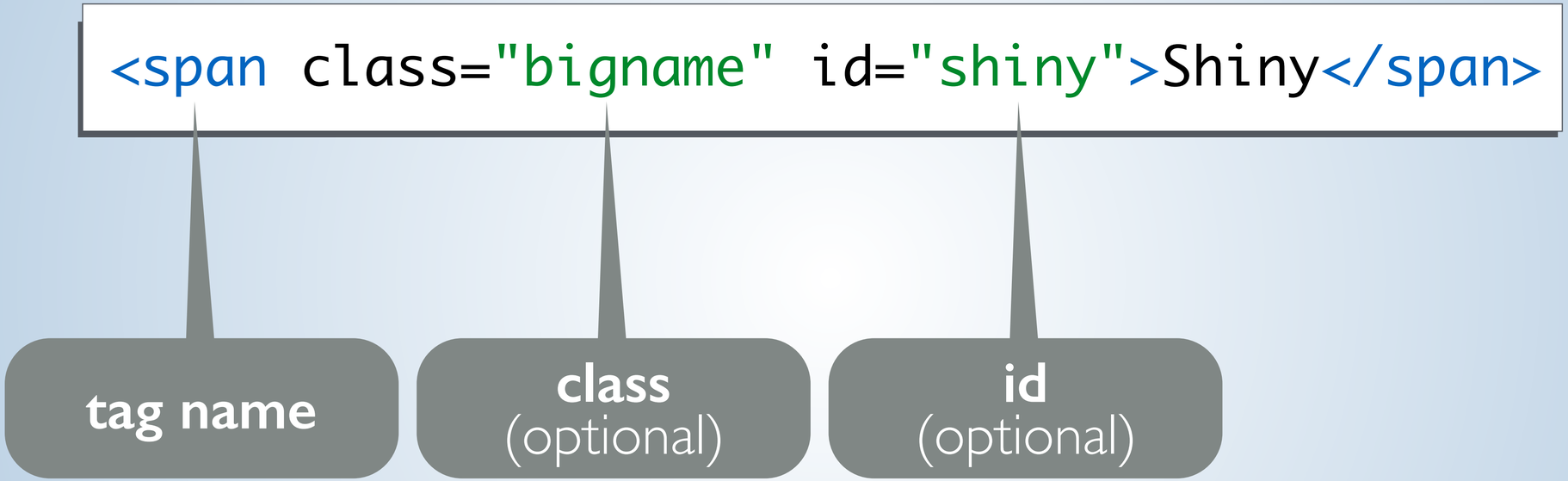
selector
styling

selector
styling

Auszug aus einer CSS Datei (vereinfacht)

CSS (Review)

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



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**

R Package: rvest

rvest

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

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

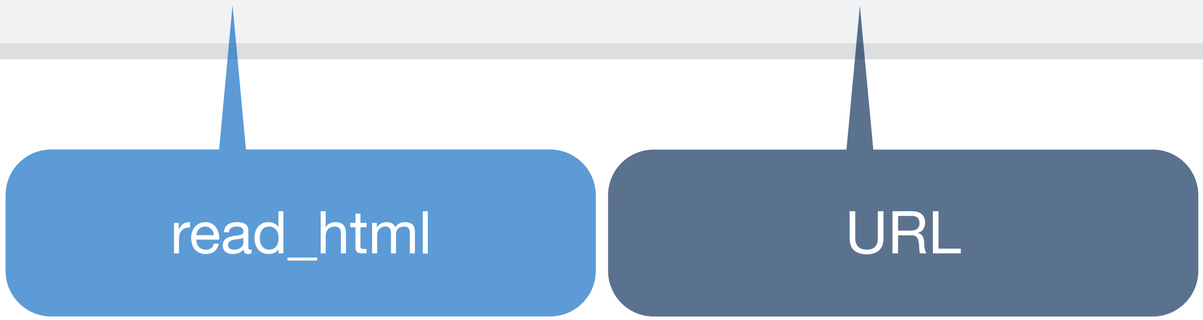
tidyverse Package

** 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()**

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



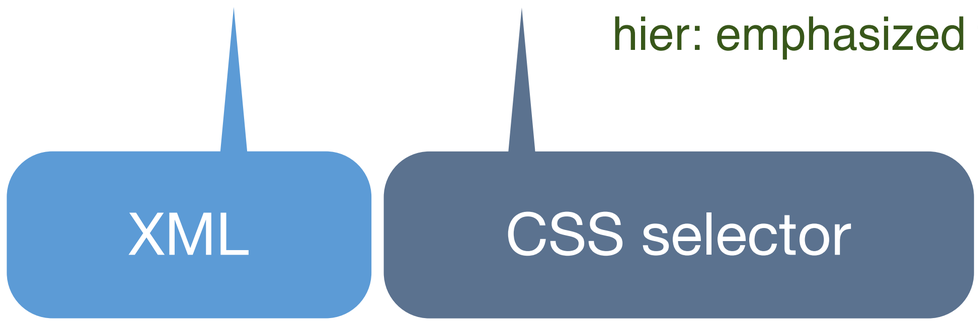
** read_html() comes in the xml2 package*

Basic Workflow

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

2. Extract specific nodes with **html_nodes()**

```
itals <- html_nodes(frozen, "em")
```



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()**

Tables

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

```
tables <- html_nodes(kw, css = "table")
html_table(tables, header = TRUE)[[2]]
```

Recap

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()

Werkzeug: selectorGadget

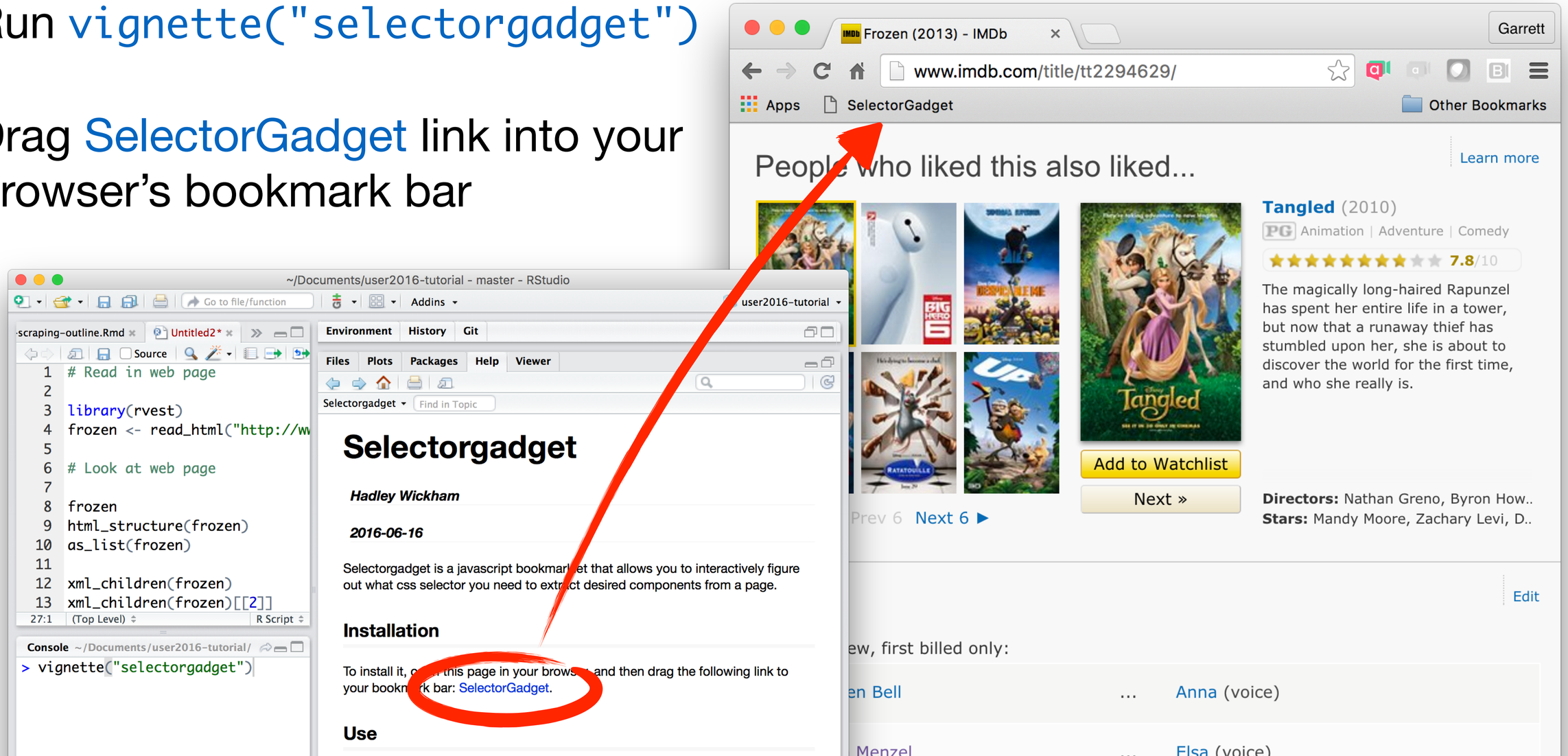
selectorGadget

A GUI tool to identify CSS selector combinations

To Install

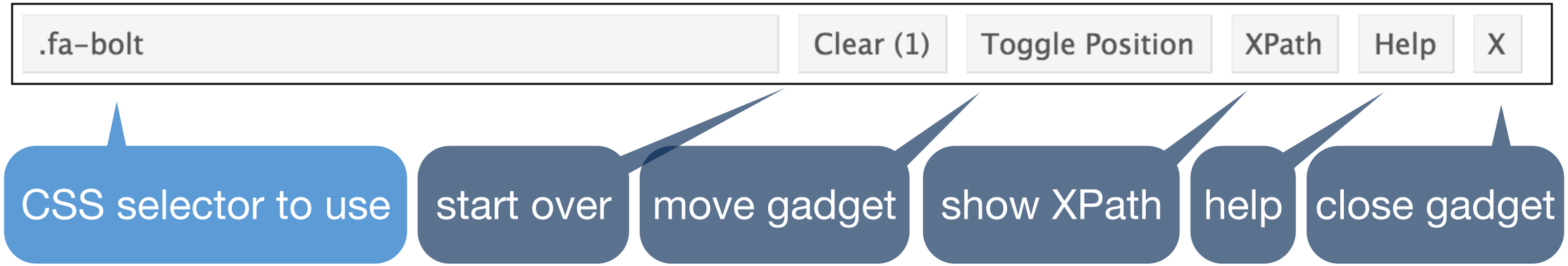
1. Run `vignette("selectorgadget")`

2. Drag SelectorGadget link into your browser's bookmark bar



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

Rechtliche Bedenken

Rechtliche Bedenken

- Urheberrechte
- Nutzungsbedingungen (Verbot automatischen Auslesens?)
- ggf. Inhalt der robots.txt Datei prüfen

Zusammenfassung

Zusammenfassung

- Prinzip: Informationen aus HTML extrahieren; dabei HTML/CSS Strukturen nutzen um die interessante Information auf umfangreichen Webseiten zu identifizieren
- im R Package **rvest** stehen die dazu erforderlichen Funktionen bereit
- mit dem Bookmarklet **selectorGadget** lassen sich leicht geeignete CSS Selektoren finden
- immer erst die rechtliche Lage überprüfen

Vielen Dank!