Dr. Jakob Jünger (jakob.juenger@uni-greifswald.de)
Chantal Gärtner (chantal.gaertner@uni-greifswald.de)

UNIVERSITÄT GREIFSWALD
Wissen lockt. Seit 1456

Automatisierte Erhebung von Social Media Daten

Workshop Teil 3: Webscraping mit R

Webscraping

Klassisch:

aus HTML ausschneiden

Interaktiv:

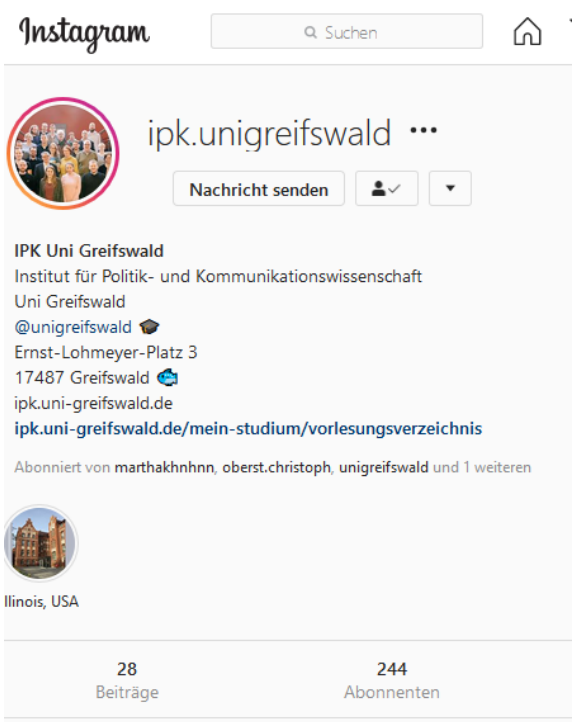
Browser fernsteuern

Kommerziell:

Daten einkaufen

Beispiel: https://www.instagram.com/ipk.unigreifswald

User Interface (Browser)



HTML (Webscraping)

```
<!DOCTYPE html>
<html lang="de" class="no-js logged-in clier
    <head>
        <meta charset="utf-8">
        <meta http-equiv="X-UA-Compatible" c

        <title>
IPK Uni Greifswald (@ipk.unigreifswald) • Ir
</title>


        <meta name="robots" content="noimage
        <meta name="apple-mobile-web-app-sta
        <meta name="mobile-web-app-capable"
        <meta name="theme-color" content="#f
        <meta id="viewport" name="viewport"
        <link rel="manifest" href="/data/mar

        <link rel="preload" href="/static/bu
<link rel="preload" href="/static/bundles/es
<link rel="preload" href="/static/bundles/es
<link rel="preload" href="/static/bundles/es
<link rel="preload" href="/static/bundles/es
<link rel="preload" href="/static/bundles/es
<link rel="preload" href="/static/bundles/es
<link rel="preload" href="/static/bundles/es
<link rel="preload" href="/static/bundles/es


        <script type="text/javascript">
        (function() {
  var docElement = document.documentElement;
```

JSON (API)

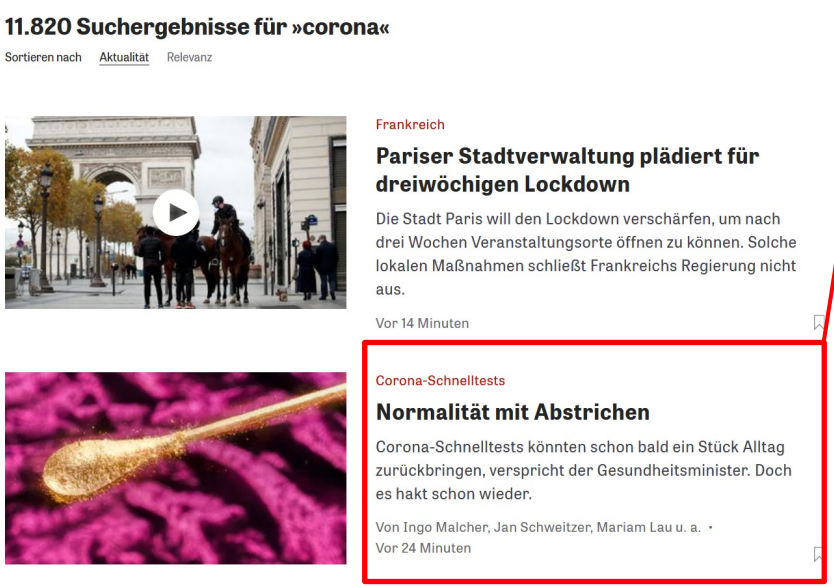

```
{
  logging_page_id: "profilePage_23191765473",
  show_suggested_profiles: false,
  show_follow_dialog: false,
▾ graphql: {
  ▾ user: {
      biography: "Institut für Politik- und
      Kommunikationswissenschaft\nUni Greifswald\n@un
      \nErnst-Lohmeyer-Platz 3\n17487 Greifswald
      greifswald.de",
      blocked_by_viewer: false,
      restricted_by_viewer: false,
      country_block: false,
      external_url: https://ipk.uni-greifswald.de/mei
      /vorlesungsverzeichnis/,
      external_url_linkshimmed: https://l.instagram.c
      %2Fipk.uni-greifswald.de%2Fmein-
      studium%2Fvorlesungsverzeichnis%2F&e=ATMEwB9fx5
      _0ypUrMLK3fmyoA7pz_vMNtCmkjBAX-
      CXX7Y0Ap0tyqvzDTxAURk404W81lX9X79sdNrW&s=1,
    ▾ edge_followed_by: {
        count: 244
      },
      followed_by_viewer: true,
    ▾ edge_follow: {
        count: 22
      },
      follows_viewer: false,
      full_name: "IPK Uni Greifswald",
      has_ar_effects: false,
      has_channel: false,
      has_blocked_viewer: false,
      highlight_reel_count: 1,
      has_requested_viewer: false,
      id: "23191765473",
      is_business_account: true,
```

HTML: Aufbau von Webseiten

HTML: Strukturierung von Inhalten	CSS: Darstellung von Inhalten	Javascript: Interaktion mit Inhalten

```
<!DOCTYPE html>
<html>
    <head>
        <script type="text/javascript" src=„actions.js“ />
        <link type="text/css" href=„styles.css" rel="stylesheet">
    </head>
    <body>
        lorem ipsum dolor
    </body>
</html>
```

HTML: Elemente + Attribute + Text




```
<div class = „teaser container“>
  <h3 class = „teaser_header“>
    <span class = „teaser_kicker“>
        Corona-Schnelltests
    </span>
    <span class = „teaser_title“>
        Normalität mit Abstrichen
    </span>
  </h3>
  <p class=„teaser_text“>
        Corona-Schnelltests könnten schon
        bald ein Stück Alltag zurückbringen,
        verspricht der Gesundheitsminister.
        Doch es hakt schon wieder.
  </p>
</div>
```

HTML Elemente (Auswahl)

Metadaten:

<head> Sammlung der Metadaten

<title> Titel des Dokuments

Abschnitte:

<body> Hauptinhalt des Dokuments

<section> Abschnitt eines Dokuments

<h1>, <h2>, … Überschriften

Gruppierte Inhalte

 allgemeiner Container

<p> Absatz

Tabellen

<table> Tabelle

<tr> Zeile einer Tabelle

<td> Zelle einer Tabelle

Listen

 Ungeordnete Liste

 Geordnete Liste

 Eintrag in einer Liste

Links

<a> Hyperlink, verweist über href-Attribut auf andere Ressource

CSS Selektoren

- Auswahl von HTML Elementen…

…by name:	`article`	selects	`<article>Mein Text</article>`
…by ID:	`#contact`	selects	`<ul id="contact">`
… by class:	`.teaser`	selects	`<p class="heavy teaser">`

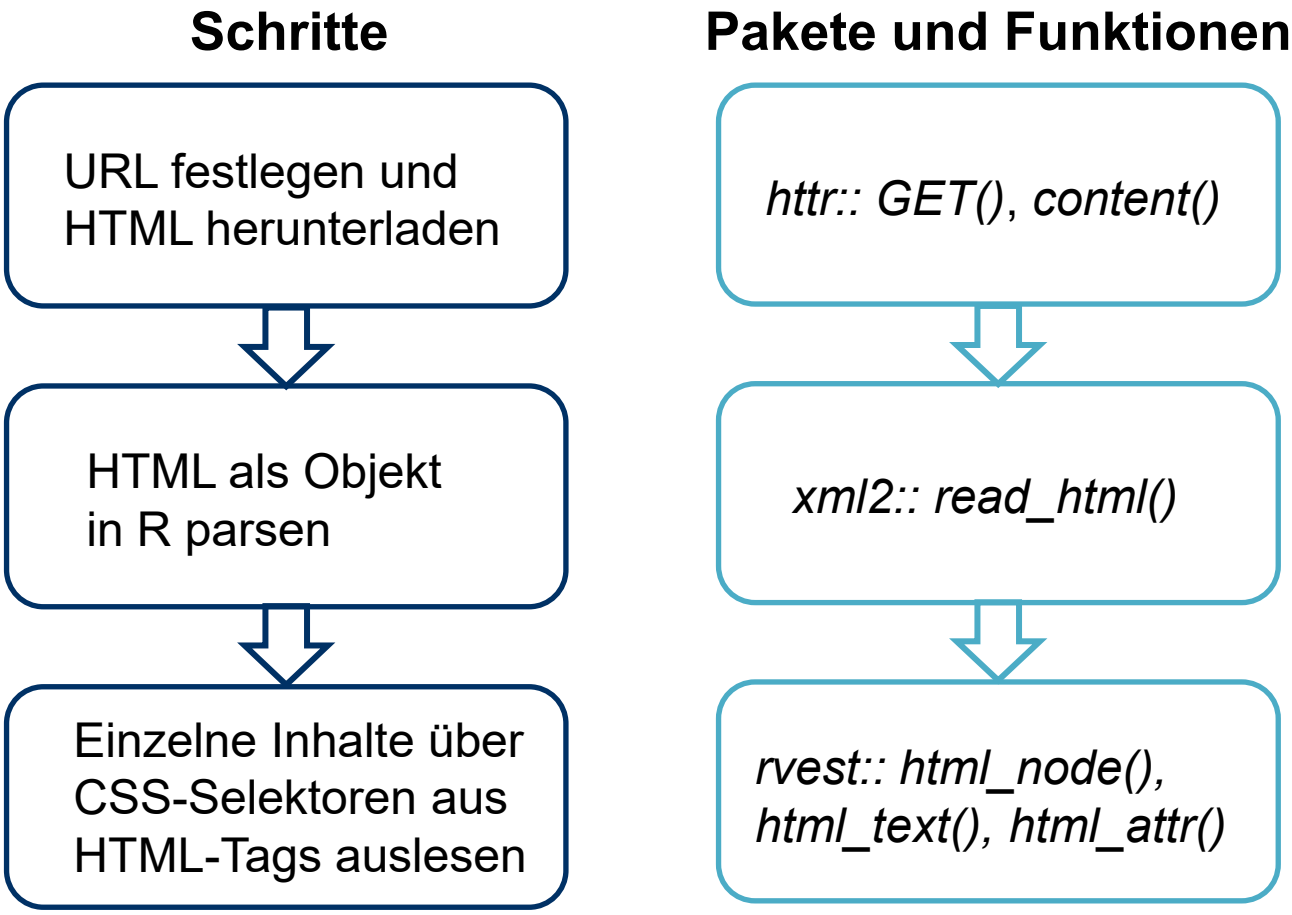
- Selektoren können aneinander gehängt werden, um verschachtelte Elemente auszuwählen:

```
ul#contact li.ceo.vip > a
```

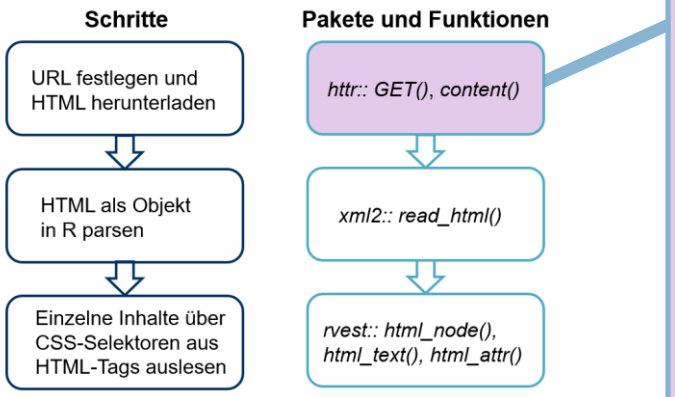
- Empfehlung: Elementname und eine Klasse verwenden

```
h2.search-counter__hits
```

Klassisches Webscraping in R



Webscraping mit RSelenium



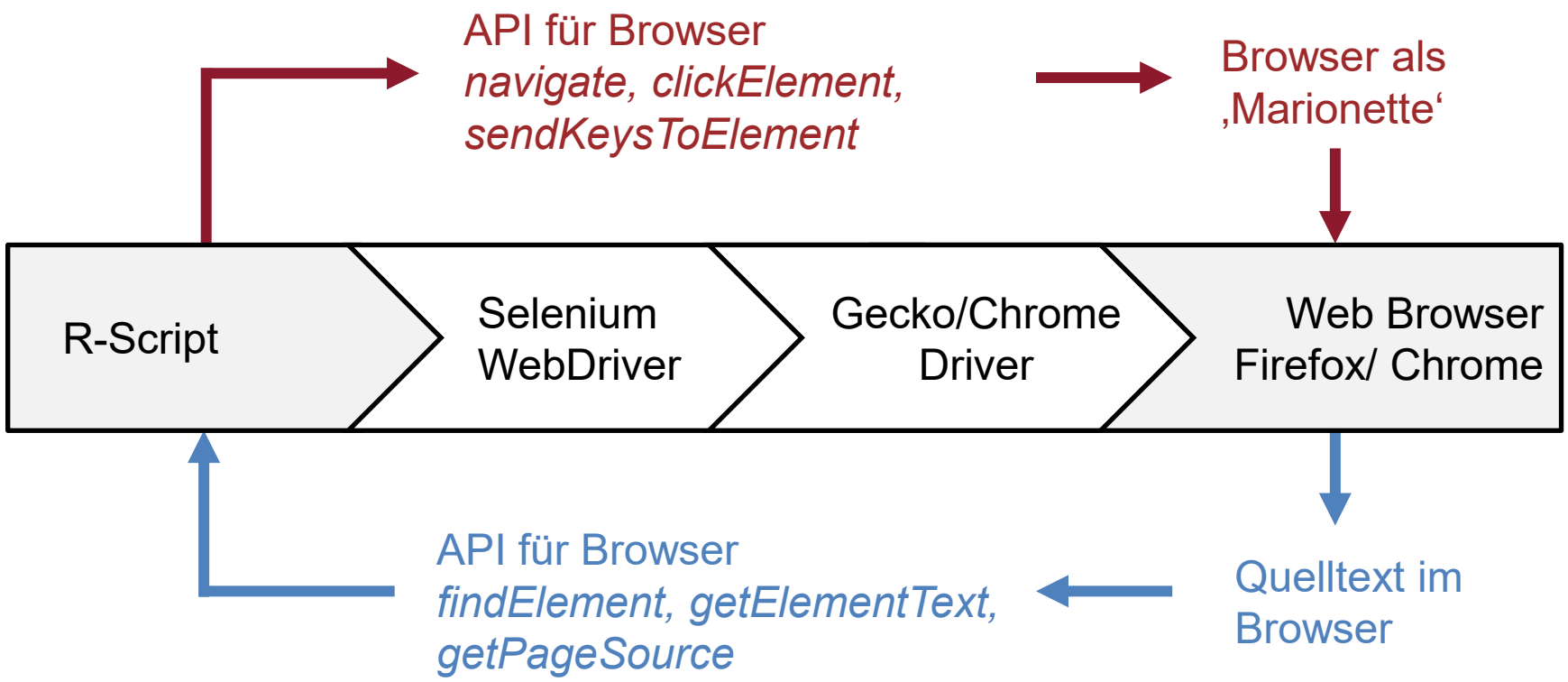
RSelenium::
remoteDriver$**navigate**()

RSelenium::
remoteDriver$**findElements**()

RSelenium::
webElement$**clickElement**()

RSelenium::
remoteDriver$**getPageSource** ()

Webscraping mit RSelenium



Zusammenfassung

Klassisch: HTML-Scraping

- Einfacher: kein JavaScript, keine Interaktivität, nur Downloads
- Schneller: Nur eine Abfrage
- Komplexer: Authentifizierung über Cookies, User-Agents

Interaktiv: Selenium

- Komplexer: Installation eines WebDrivers, interaktive Seite
- Langsamer: Komplette Seite wird geladen inkl. JavaScript
- Hybrid: Manuelle und automatisierte Interaktion
- Natürlicher: Darstellung in einem normalen Browser

Ende

CSS-Selektoren in R

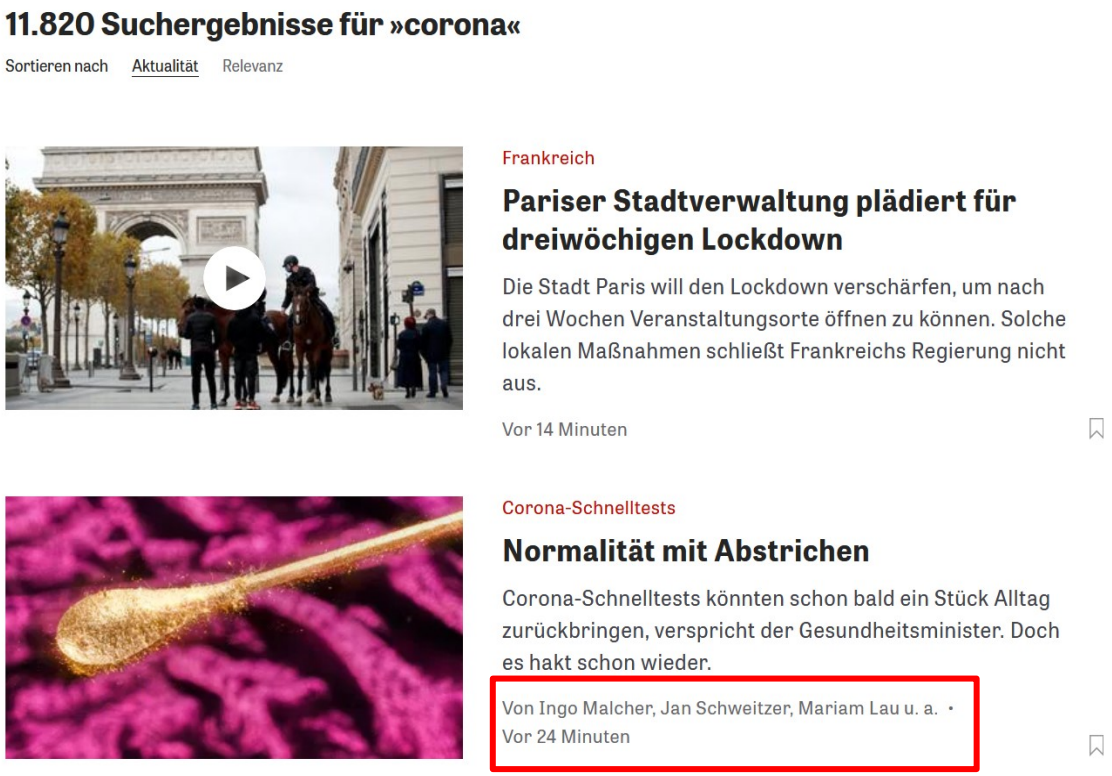
Zum **Ansteuern** einzelner Inhalte eines HTMLs: *html_node()*

- Gesamtes Element auswählen: Elementname, z.B. *html_node("time")*
- Nur eine Klasse (=Art von Attribut) innerhalb eines Elements auswählen: Elementname.Klassenname, z.B. *html_node("time.standard_datetime")*

Zum **Auswählen** einzelner Inhalte: u.a. *html_text()*, *html_attr()*

```
<time class="standard_datetime"
datetime="2021-03-01T09:05:37+01:00">
Vor 20 Minuten
</time>
```

Übung: Inhalte aus HTML auslesen



- Link öffnen in Firefox: https://www.zeit.de/suche/index?q=corona
- HTML-Quellcode anzeigen lassen (Element Untersuchen)
- Frage: Über welche Elemente und Attribute können Autoren und Datumsangaben ausgelesen werden?