



Automatisierte Erhebung von Social Media Daten

Workshop Teil I: Datenzugänge, Datenformate
und Erhebungsverfahren

Ziele

Praktische Anwendung



Grundverständnis



Ziele



Welche Posts auf Facebook fahren **den meisten** Ärger ein?

Wie können **Kommentare** auf Facebook erhoben werden?



Über welche wissenschaftliche Disziplin wird **am wenigsten** auf Zeit.de berichtet?

Wie können **Kommentare** auf Nachrichtenseiten erhoben werden?



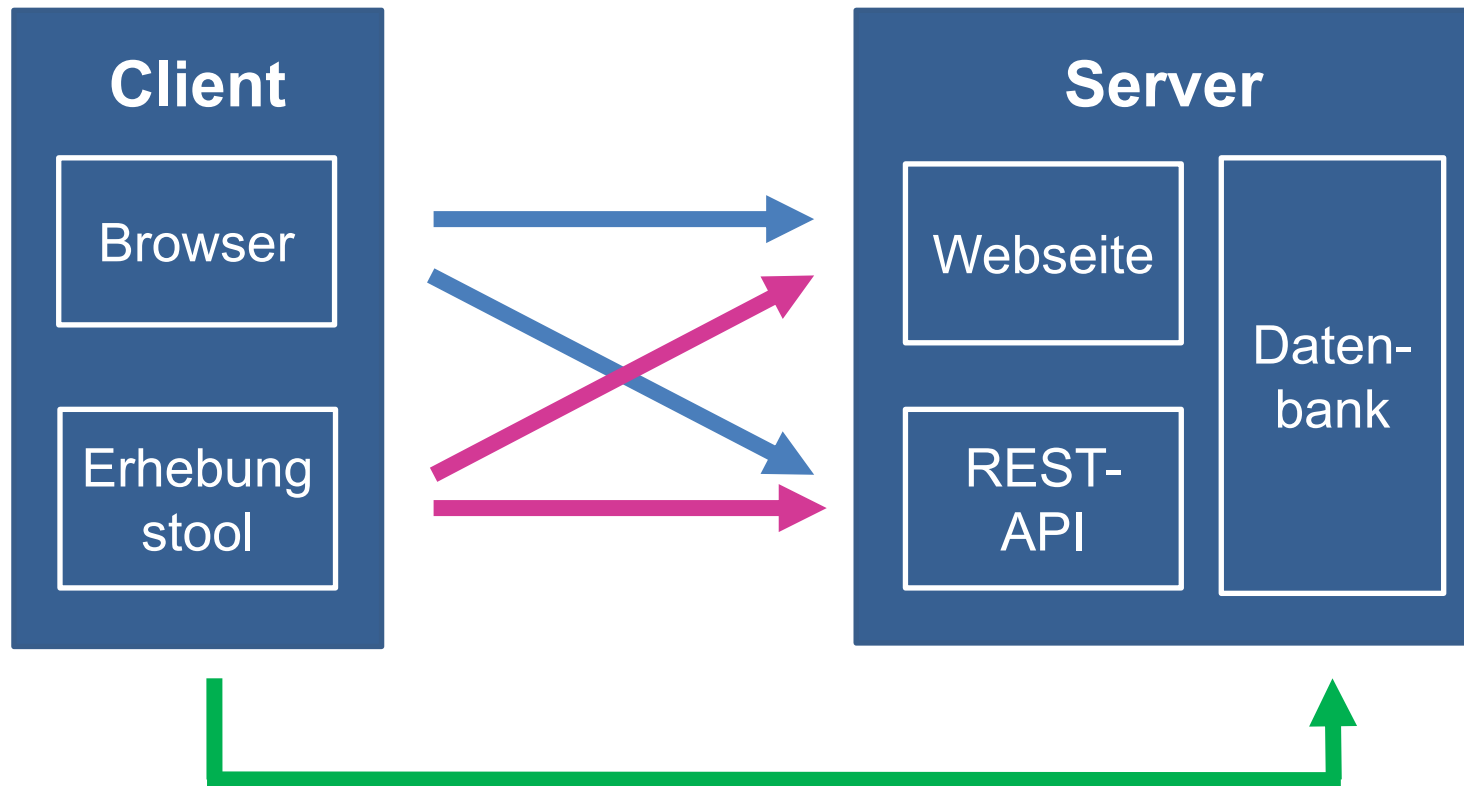
Welche ist die **beliebteste** Online-Plattform?

Wie können **Kommentare** auf TikTok erhoben werden?

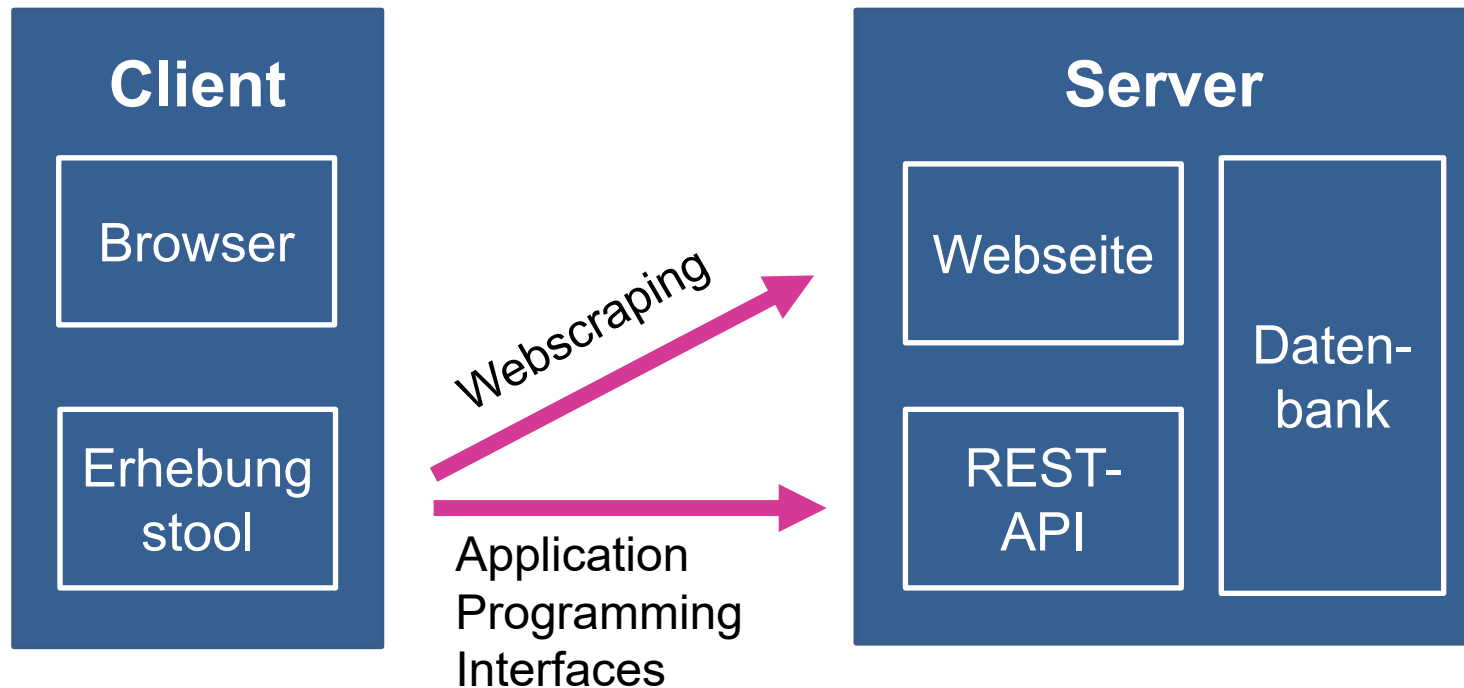
Zeitplan

Teil 1	
9:00 – 10:00 Uhr	Einführung Datenzugänge und Datenformate
Teil 2	
10:00 – 12:00 Uhr	APIs: Erhebung mit Facepager
12:00 – 12:45 Uhr	<i>Mittagspause</i>
12:45 – 14:15 Uhr	APIs: Einlesen und aufbereiten von Daten in R
14:15 – 14:30 Uhr	<i>Kaffee- und Teepause</i>
Teil 3	
14:30 – 16:00 Uhr	Klassisches Webscraping in R
16:00 – 16:15 Uhr	<i>Kaffee- und Teepause</i>
16:15 – 17:15 Uhr	Webscraping mit Selenium
17:15 – 18:00 Uhr	Offene Fragen und Ausblick

Automatisierte Datenerhebung im Web

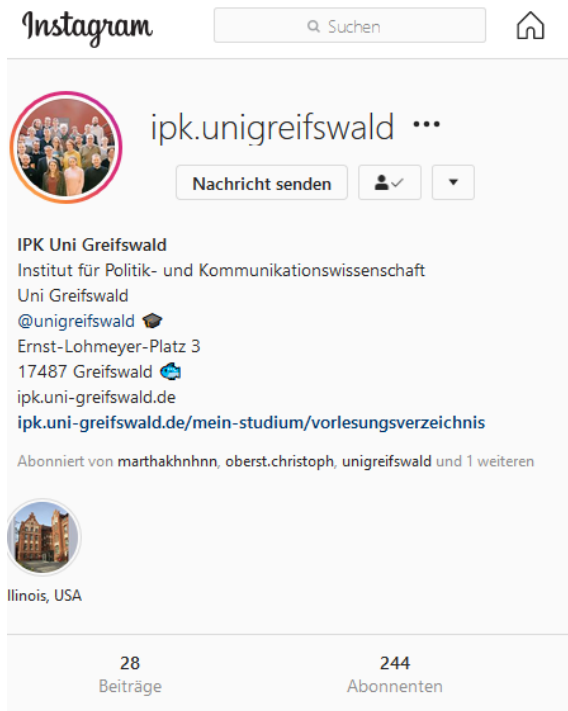


Automatisierte Datenerhebung im Web



https://www.instagram.com/ipk.unigreifswald/?__a=1

User Interface (Browser)



HTML (Webscraping)

[illegible]

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\nun  
☛\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://1.instagram.c%2Fipk.uni-greifswald.de%2Fmein-studium%2Fvorlesungsverzeichnis%2F%e=ATMEwB9fx5\_OypUrMLK3fmycA7pz\_vMvNcmkjBAX-CXX7Y0Ap0tyqvzDTxAURk404W8l1X9X79sdnNrW=s1,
      * 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,
    }
  }
}
```

Wichtige Datenformate

- XML/HTML: `<p class="wichtig">Inhalt eines Absatzes</p>`
Auszeichnungssprache,
Strukturierung von Text mit Elementen („Tags“),
Hierarchische Struktur
- JSON: `{"Name": "Greifswald"}`
Name-Wert-Paare und Listen
Hierarchische Struktur
- CSV: `Name;Gegrundet;Mitarbeiterinnen\n`
Tabellenformat, jede Zeile ist ein Datensatz, erste Zeile ist
Überschrift, Felder durch Komma oder Semikolon getrennt.

URLs

<https://www.youtube.com/watch?v=dbTREHtu1O0#comments>

Protokoll Domain Pfad Parameter Hashfragment

Tools & Co

- Variante 1: Kommerzielle **Dienste**
- Variante 2: **Tools** mit Benutzeroberflächen
 - Lokal: Facepager, Rapidminer
 - Server: DMI Tools
- Variante 3: **Packages** für R oder Python
 - Beispiel: twitterR
- Variante 4: **Frameworks** benutzen
 - Beispiel: Scrapy
- Variante 5: **Skripte** entwickeln
 - R: Rvest, RSelenium
 - Python: requests, beautifulsoup, selenium

Einfach &
schnell



Flexibel &
transparent