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 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?

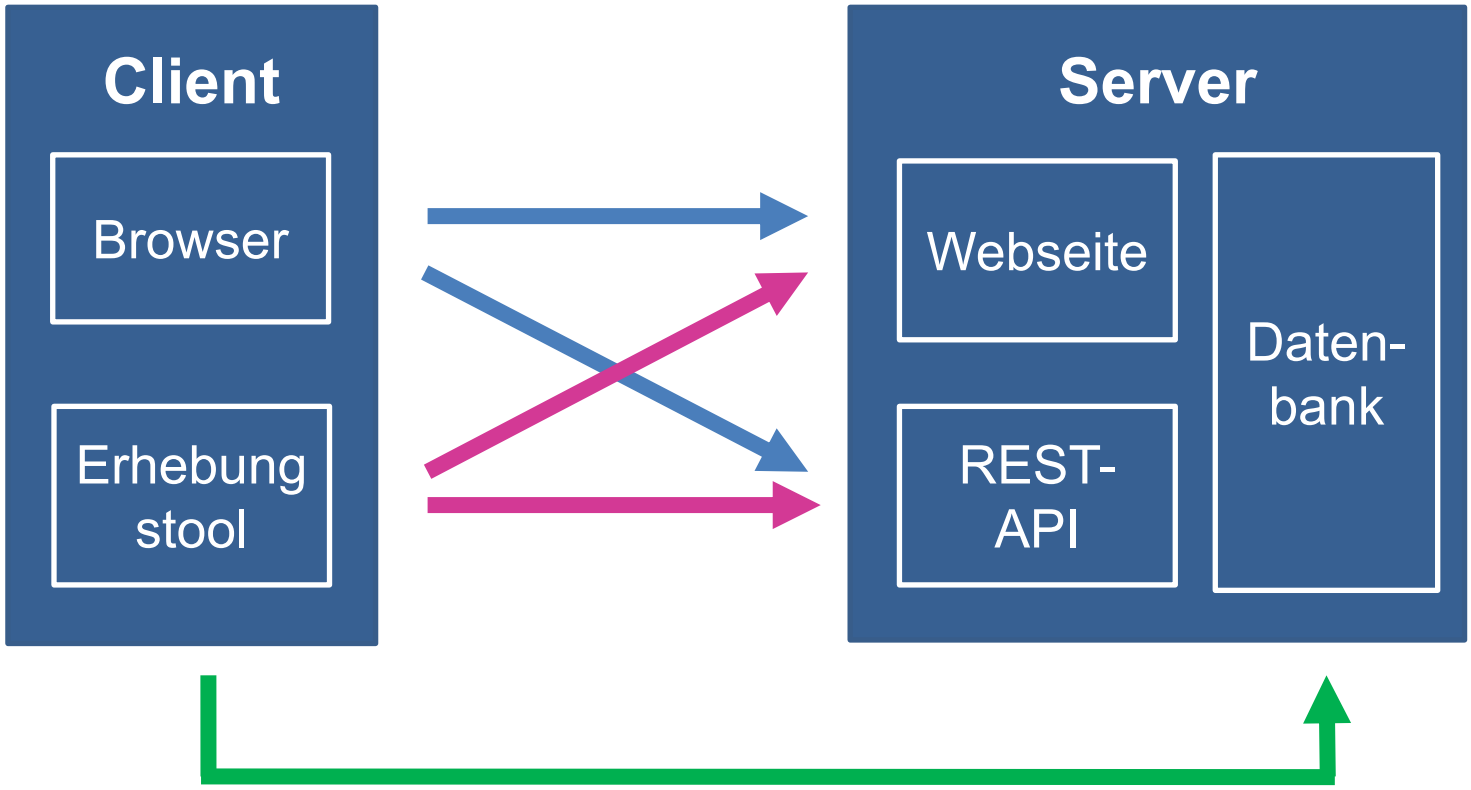
Welche ist die **beliebteste** Online-Platform?

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

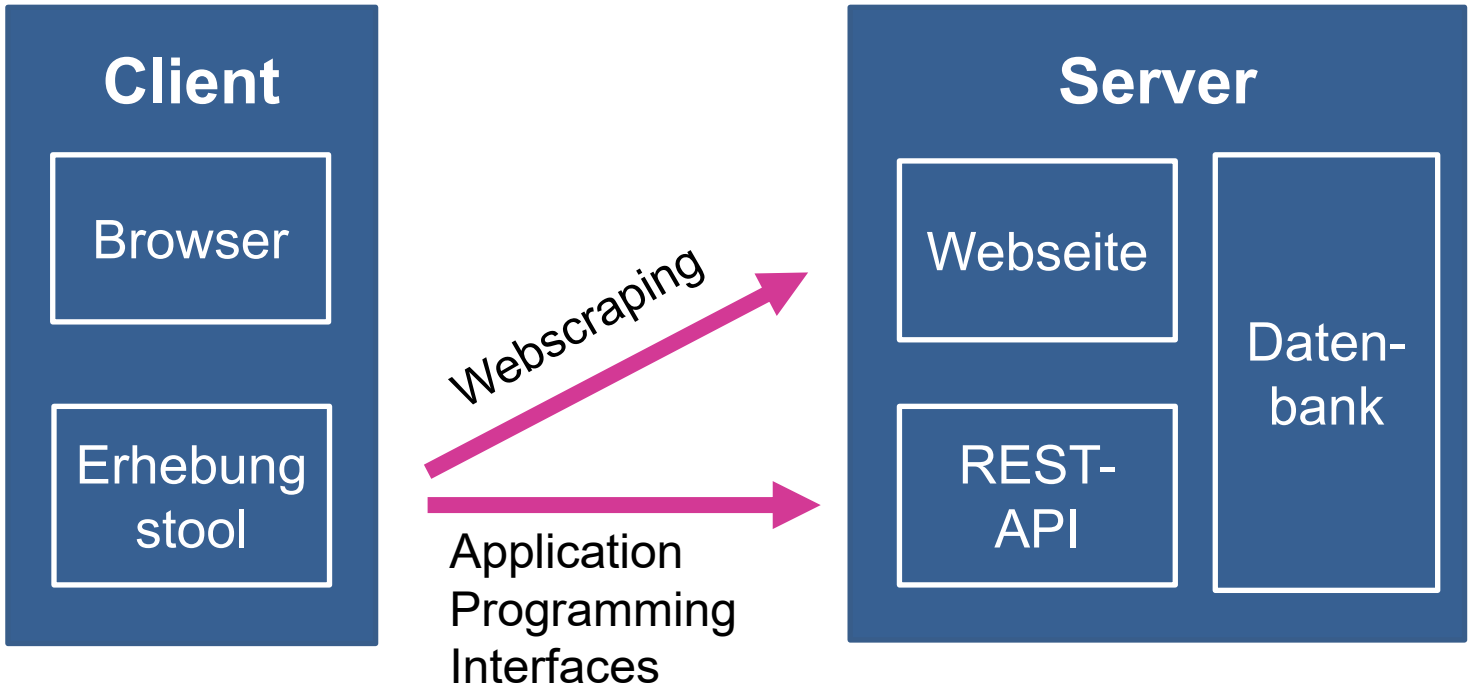
Zeitplan

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	*Mittagspause*
12:45 – 14:15 Uhr	APIs: Einlesen und aufbereiten von Daten in R
14:15 – 14:30 Uhr	*Kaffee- und Teepause*
Teil 3	
14:30 – 16:00 Uhr	Klassisches Webscraping in R
16:00 – 16:15 Uhr	*Kaffee- und Teepause*
16:15 – 17:15 Uhr	Webscraping mit Selenium
17:15 – 18:00 Uhr	Offene Fragen und Ausblick

Automatisierte Datenerhebung im Web

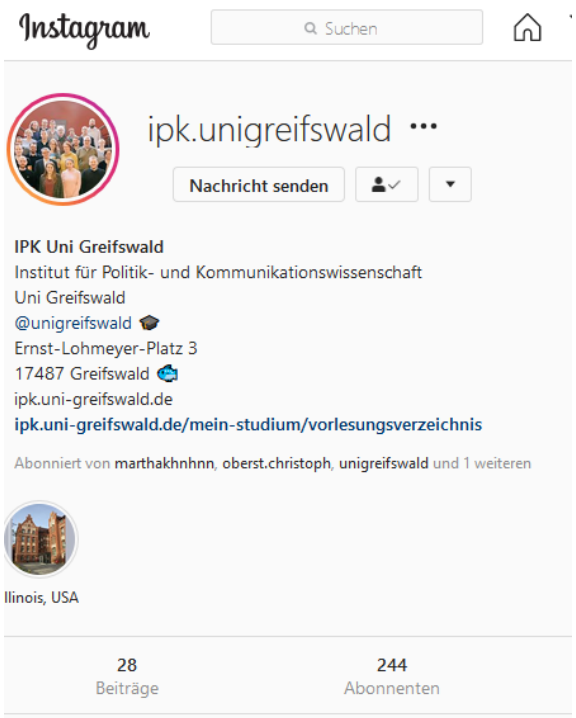


Automatisierte Datenerhebung im Web



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

User Interface (Browser)



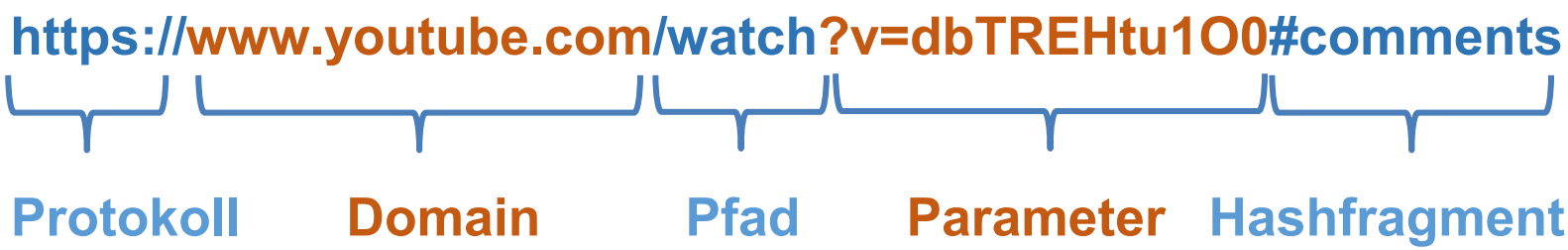
HTML (Webscraping)

JSON (API)

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;Gegruendet;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