SUMMER INCUBATOR WORKSHOP
INTRODUCTION TO WORKING WITH WEB-APIS

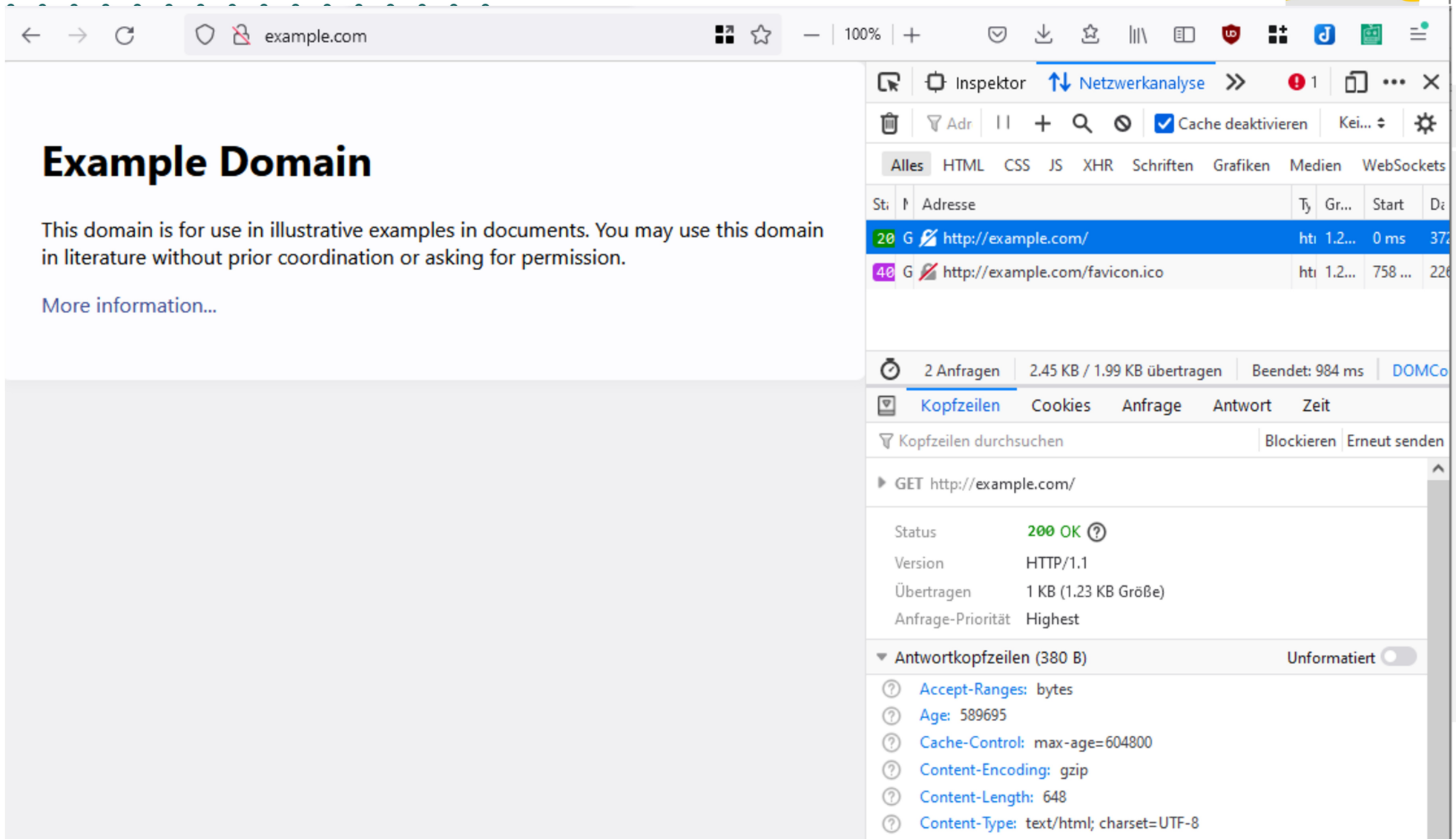
Tom Theile

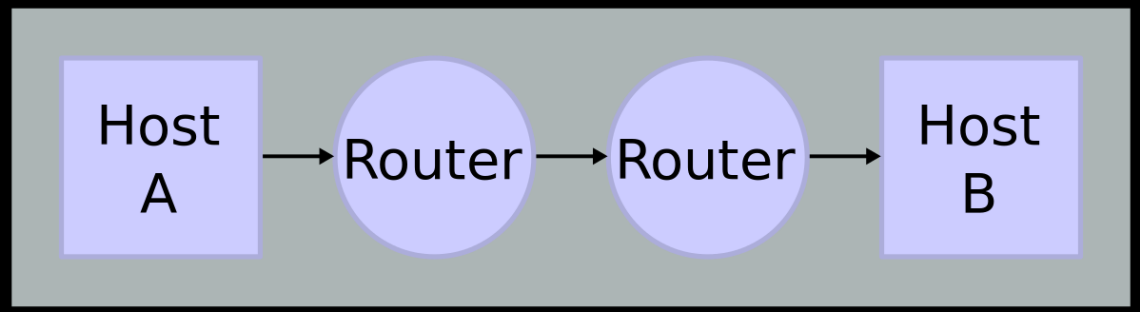
Lab of digital and computational demography

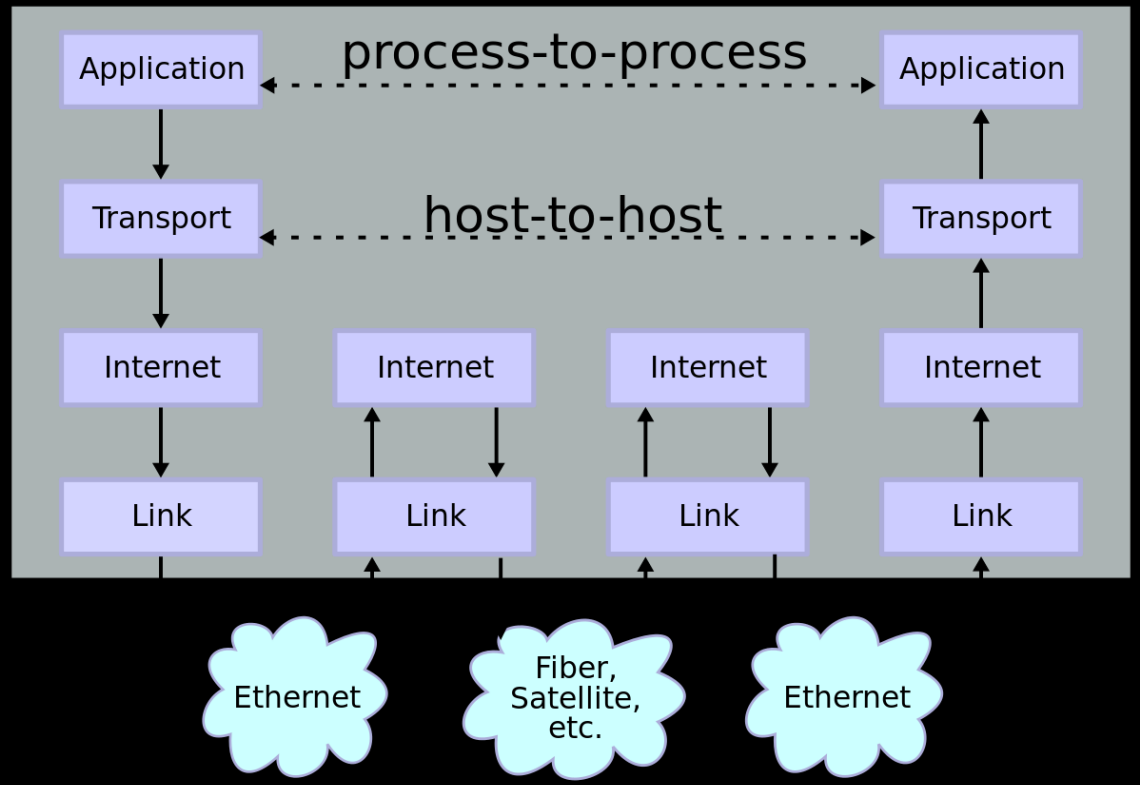
AGENDA

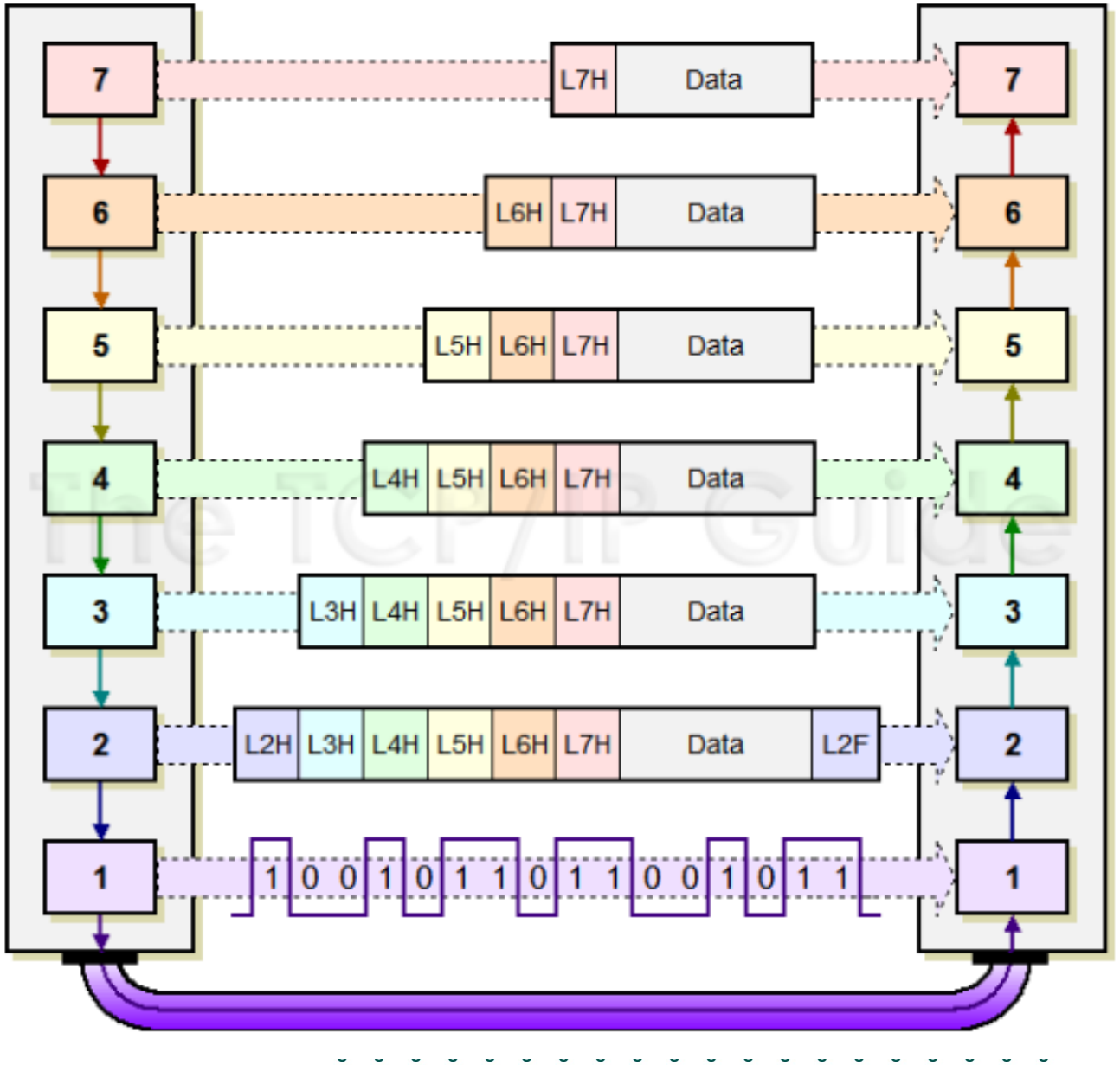
- What is the Internet? How do websites work?
- How to scrape data from websites?
- What are APIs?
- Workshop: Working with a weather-API
- Workshop: Getting Tweet-counts from the Twitter-API

Feel free to interrupt me at any time!


example.com
100%
Example Domain
This domain is for use in illustrative examples in documents. You may use this domain in literature without prior coordination or asking for permission.
More information...
Inspektor
Netzwerkanalyse
Cache deaktivieren
Alles HTML CSS JS XHR Schriften Grafiken Medien WebSockets
Adresse
http://example.com/
http://example.com/favicon.ico
2 Anfragen
2.45 KB / 1.99 KB übertragen
Beendet: 984 ms
Kopfzeilen Cookies Anfrage Antwort Zeit
Kopfzeilen durchsuchen
Blockieren
Erneut senden
GET http://example.com/
Status 200 OK
Version HTTP/1.1
Übertragen 1 KB (1.23 KB Größe)
Anfrage-Priorität Highest
Antwortkopfzeilen (380 B)
Unformatiert
Accept-Ranges: bytes
Age: 589695
Cache-Control: max-age=604800
Content-Encoding: gzip
Content-Length: 648
Content-Type: text/html; charset=UTF-8


Host A
Router
Router
Host B


Application
process-to-process
Application
Transport
host-to-host
Transport
Internet
Internet
Internet
Internet
Link
Link
Link
Link
Ethernet
Fiber, Satellite, etc.
Ethernet


7
6
5
4
3
2
1
L7H
Data
L6H
L5H
L4H
L3H
L2H
L2F
1 0 0 1 0 1 1 0 1 1 0 0 1 0 1 1

WHAT IS THE INTERNET

- Browser sends a „GET“-request (HTTP-protocol)
- DNS-server looks up IP-address of domain name
- TCP-IP-packet is send to your internet-provider with destination IP-address
- Routers pass the packet on to other routers that are nearer to the destination
- Destination webserver reads the innermost header and the data and returns a response … the journey starts again!

- You can list all the routers on this journey with traceroute (tracert on Windows)
- Tracert example.com
- Visualize it with https://stefansundin.github.io/traceroute-mapper/

```
C:\Users\tom>tracert example.com

Routenverfolgung zu example.com [93.184.216.34]
über maximal 30 Hops:

  1     *        3 ms     1 ms  192.168.2.1
  2     *        *        *     Zeitüberschreitung der Anforderung.
  3    25 ms    24 ms    23 ms  217.237.147.45
  4    26 ms    25 ms    23 ms  195.145.92.114
  5     *        *        *     Zeitüberschreitung der Anforderung.
  6    30 ms    28 ms    27 ms  f-ed12-i.F.DE.NET.DTAG.DE [62.154.3.97]
  7    28 ms    28 ms    28 ms  ffm-b5-link.ip.twelve99.net [213.248.93.186]
  8    32 ms    27 ms    27 ms  ffm-bb1-link.ip.twelve99.net [62.115.114.88]
  9    38 ms    38 ms    38 ms  prs-bb1-link.ip.twelve99.net [62.115.123.13]
 10   155 ms   180 ms   221 ms  ash-bb2-link.ip.twelve99.net [62.115.112.242]
 11   125 ms   231 ms   198 ms  ash-b2-link.ip.twelve99.net [62.115.123.125]
 12   222 ms   199 ms   199 ms  verizon-ic-315152-ash-b1.ip.twelve99-cust.net [213
 13   125 ms   125 ms   174 ms  ae-66.core1.dcb.edgecastcdn.net [152.195.65.129]
 14   224 ms   204 ms   123 ms  93.184.216.34

Ablaufverfolgung beendet.
```

SURFING THE WEB WITH R

- With Python or R can also send and receive HTTP requests:

```
library(rvest)
simple_page <- read_html("https://user.demogr.mpg.de/theile/Files/edsd/simple.html")
print(paste(simple_page))
```

```
<!DOCTYPE html>
  <html>
  <body>

  <h1>Hello Barcelona!</h1>

  <p>This is some text.</p>

  <a href="https://www.example.com">This is a link</a>
  <a href="./simple2.html">This is also a link</a>
  </body>
  </html>
```

SURFING THE WEB WITH R

- The library rvest includes helpful tools to extract information from HTML-pages:

```
css_selector <-".current-w-temp"

page_weather <- read_html("https://kachelmannwetter.com/de/wetter/2844588-rostock")

temperature <- page_weather %>%
    html_nodes(css_selector) %>%   # This line selects all the html-elements with that css-selector
    html_text()   # This extracts only the Text of the selected html-elements

print(temperature)
```

```
[1] "27°"
```

DIFFERENT APIS

- There are different classes of APIs
  - Operating System APIs (e.g.: write a file to a folder)
  - R-package APIs (e.g.: read a dataframe from a csv-file with read.csv())
  - ...
  - Web-APIs
- We only care about Web-APIs in this course

WEB APIS

- Web APIs are "websites for machines". They provide machine-readable access to to data and services.
- While websites are designed to be read by humans, APIs are designed to be read by machines.
- That means that the answers of APIs don't need to be formatted in a fancy style (like a flashy website), but have to follow predictable rules.
- Test it: https://dog-facts-api.herokuapp.com/api/v1/resources/dogs?number=3

FROM SCRAPING TO APIS

- Webpages are for Humans, APIs are for computers
- Webpages change over time and all the scraping scripts won't work anymore…
- APIs are structured and machine readable
- RESTful APIs - most common nowadays
- GraphQL - can make complex queries, newer, not as common
- APIs can response with any data type, but most often: JSON
- → let's use an API!

FROM SCRAPING TO APIS

- https://dog-facts-api.herokuapp.com/api/v1/resources/dogs?number=3

- https://docs.openalex.org/about-the-data/work
  https://api.openalex.org/authors?filter=display_name.search:tom+theile
  https://api.openalex.org/works?filter=author.id:A773951722 as webpage

- Please open 02_using_a_simple_API.Rmd

THE TWITTER API

- https://developer.twitter.com/en https://developer.twitter.com/en/docs/twitter-api
- Different access levels (Commercial, free, academic access)
- extensively documented (but not always easy to find the right information)
- Access to tweets, tweetcounts, users, trends and more
- „Over 11% of Twitter's revenue in FY 2021, or $571.8 million, was from data licensing and other sources.“ (https://www.investopedia.com/ask/answers/120114/how-does-twitter-twtr-make-money.asp)

V2 Access Levels

Essential

With Essential access, you can now get access to Twitter API v2 quickly and for free!

- Retrieve 500,000 Tweets per month
- 1 Project per account
- 1 App environment per Project
- Limited access to standard v1.1 (**only media endpoints**)
- No access to premium v1.1, or enterprise

Elevated

With Elevated access, you can get free, additional access to endpoints and data, as well as additional App environments.

- Retrieve 2 million Tweets per month
- 1 Project per account
- 3 App environments per Project
- Access to standard v1.1, premium v1.1, and enterprise

Academic Research

If you qualify for our Academic Research access level, you can get access to even more data and advanced search endpoints.

- Retrieve 10 million Tweets per month
- Access to full-archive search and full-archive Tweet counts
- Access to advanced search operators

Enterprise: Gnip 2.0

Our enterprise APIs offer the highest level of access and reliability to those who depend on Twitter data.

Learn more >

Premium v1.1

The premium v1.1 endpoints offer scalable access to Twitter data for those looking to grow, experiment, and innovate by using historical search and subscribing to user activities.

Learn more >

Standard v1.1

The standard v1.1 endpoints were launched in 2012 and enables you to post, interact, and retrieve data for resources such as Tweets, Users, Direct Messages, Lists, Trends, Media, and Places.

Learn more >

TWITTER API

-

THANK YOU FOR YOUR ATTENTION!

Tom Theile

Software developer at the lab of digital and computational demography

theile@demogr.mpg.de

THANK YOU!