DIGITAL DEMOGRAPHY: ANALYZING WEB AND SOCIAL MEDIA DATA

DAY 1 – INTRODUCTION, INTERNET AND SCRAPING

EDSD NOVEMBER 2022

Tom Theile

Lab of digital and computational demography,

Max-Planck-Institute for Demographic Research, Rostock

LAB OF DIGITAL AND COMPUTATIONAL DEMOGRAPHY

AGENDA AND SCHEDULE OF THE WEEK

- 9:30 – 13:00 Monday to Thursday
- Friday: assignment
- Block 1: Internet, Webscraping, APIs
- Block 2: Digital Demography
- Block 3: Social Media Data
- Extra: Digital Datasets

FINAL ASSIGNMENT

- A small number of short tasks
- To be done and submit until Friday afternoon 6pm
- Everyone has to submit the answers by email
- Email has to contain your full name!
- No group work
- More information on Thursday

WHO ARE YOU?

- 4 Things in common - related to migration, mortality, fertility and digital data

AGENDA

- What is the Internet? How do websites work?
- How to scrape data from websites?
- What to do with such data?

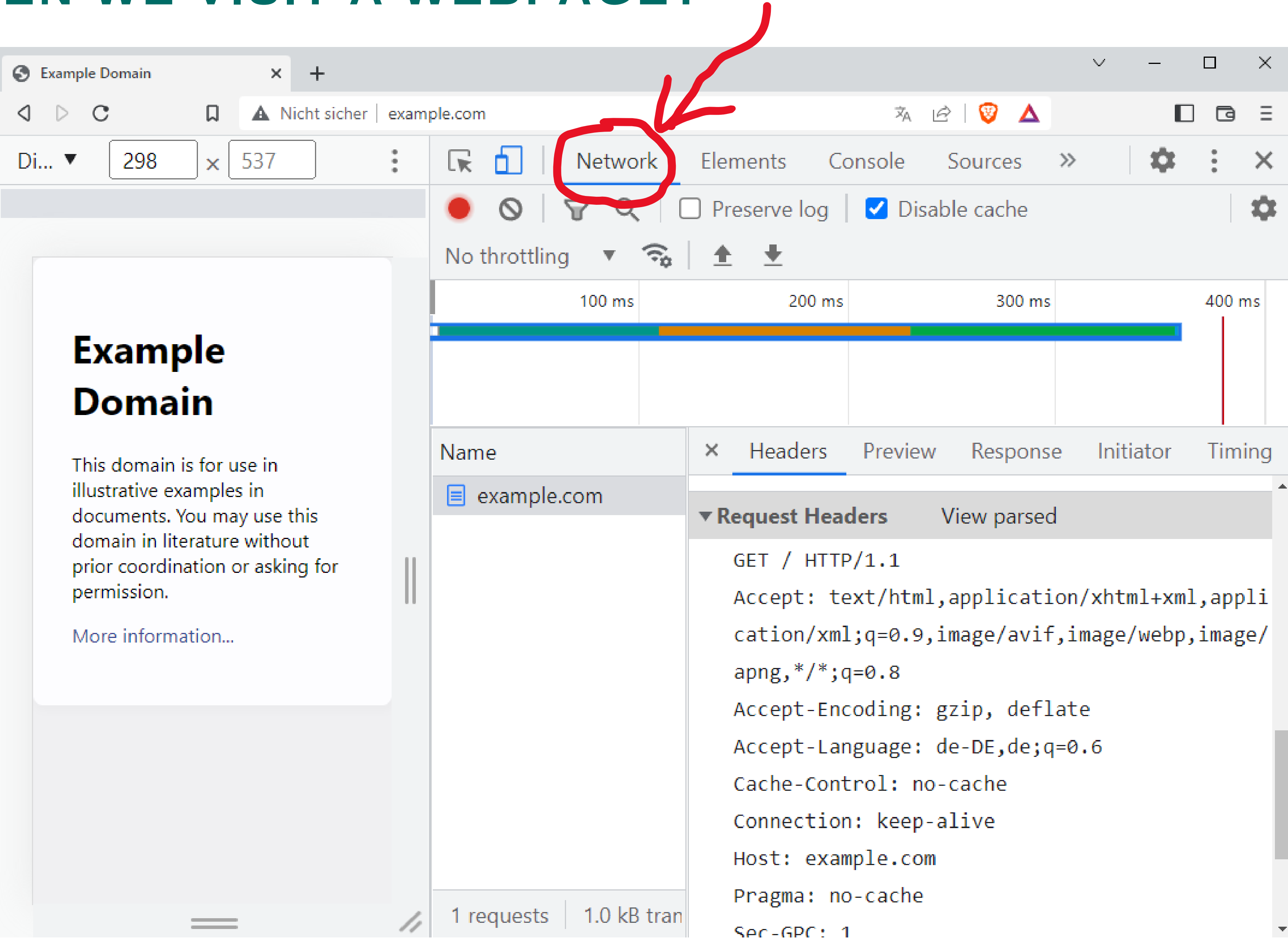
WHAT HAPPENS WHEN WE VISIT A WEBPAGE?

Open the DevTools in your Browser:

```
Windows or Linux:
F12 or
“CTRL + shift + i”
```

```
Mac:
“Fn + F12” or
“Cmd + Option + I”
```

Open the Network panel!



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...

☐ Preserve log ☑ Disable cache

No throttling

example.com

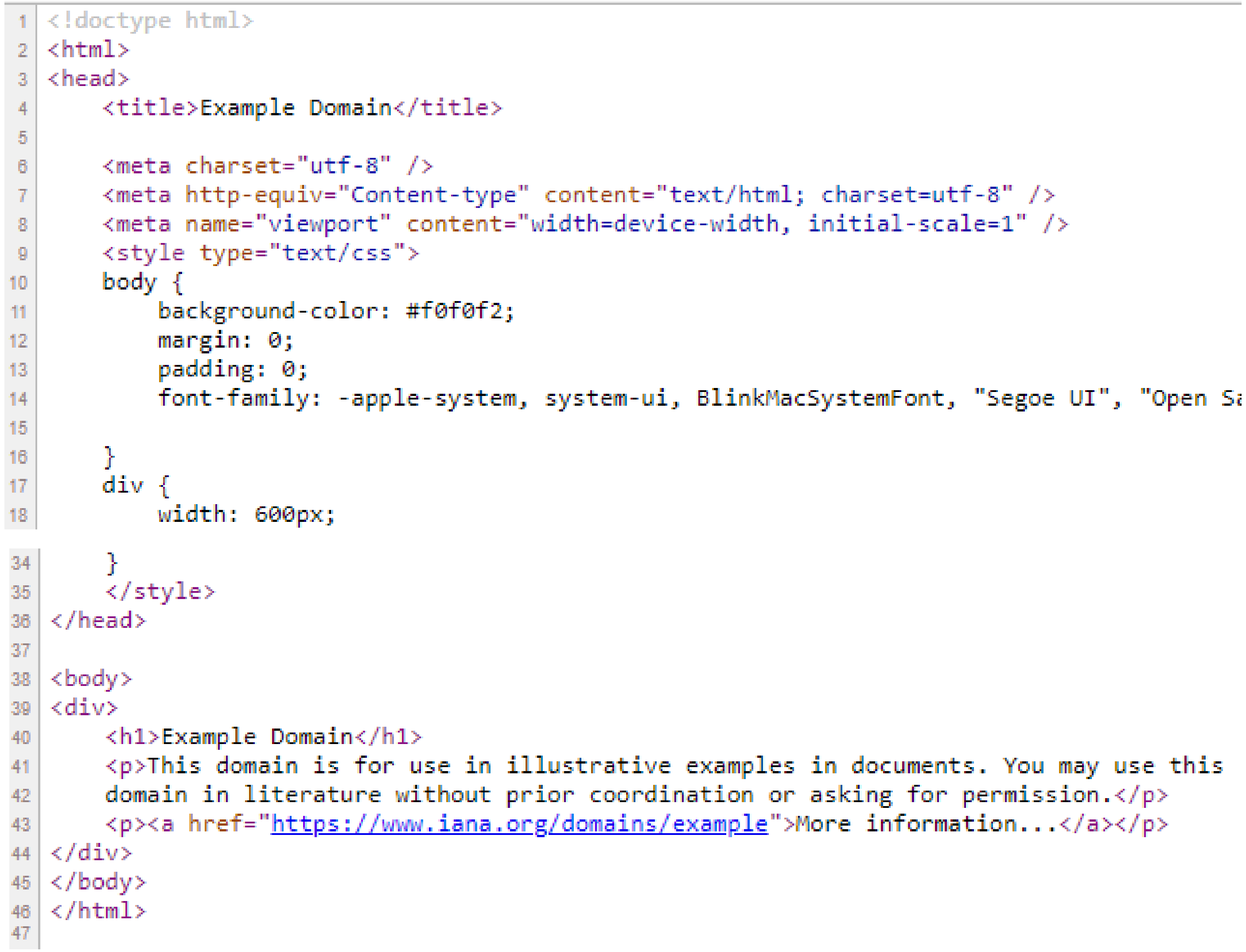
1 requests 1.0 kB tran

▾ Request Headers

```
GET / HTTP/1.1
Accept: text/html,application/xhtml+xml,appli
cation/xml;q=0.9,image/avif,image/webp,image/
apng,*/*;q=0.8
Accept-Encoding: gzip, deflate
Accept-Language: de-DE,de;q=0.6
Cache-Control: no-cache
Connection: keep-alive
Host: example.com
Pragma: no-cache
Sec-GPC: 1
```

EXAMPLE.COM HTML

- If you are not familiar with the basics of HTML, this is a good resource to learn the basics or refresh your knowledge:
- https://developer.mozilla.org/en-US/docs/Learn/Getting_started_with_the_web/HTML_basics



```
<!doctype html>
<html>
<head>
    <title>Example Domain</title>

    <meta charset="utf-8" />
    <meta http-equiv="Content-type" content="text/html; charset=utf-8" />
    <meta name="viewport" content="width=device-width, initial-scale=1" />
    <style type="text/css">
    body {
        background-color: #f0f0f2;
        margin: 0;
        padding: 0;
        font-family: -apple-system, system-ui, BlinkMacSystemFont, "Segoe UI", "Open Sa

    }
    div {
        width: 600px;

    }
    </style>
</head>

<body>
<div>
    <h1>Example Domain</h1>
    <p>This domain is for use in illustrative examples in documents. You may use this
    domain in literature without prior coordination or asking for permission.</p>
    <p><a href="https://www.iana.org/domains/example">More information...</a></p>
</div>
</body>
</html>
```

CSS - CASCADING STYLE SHEETS

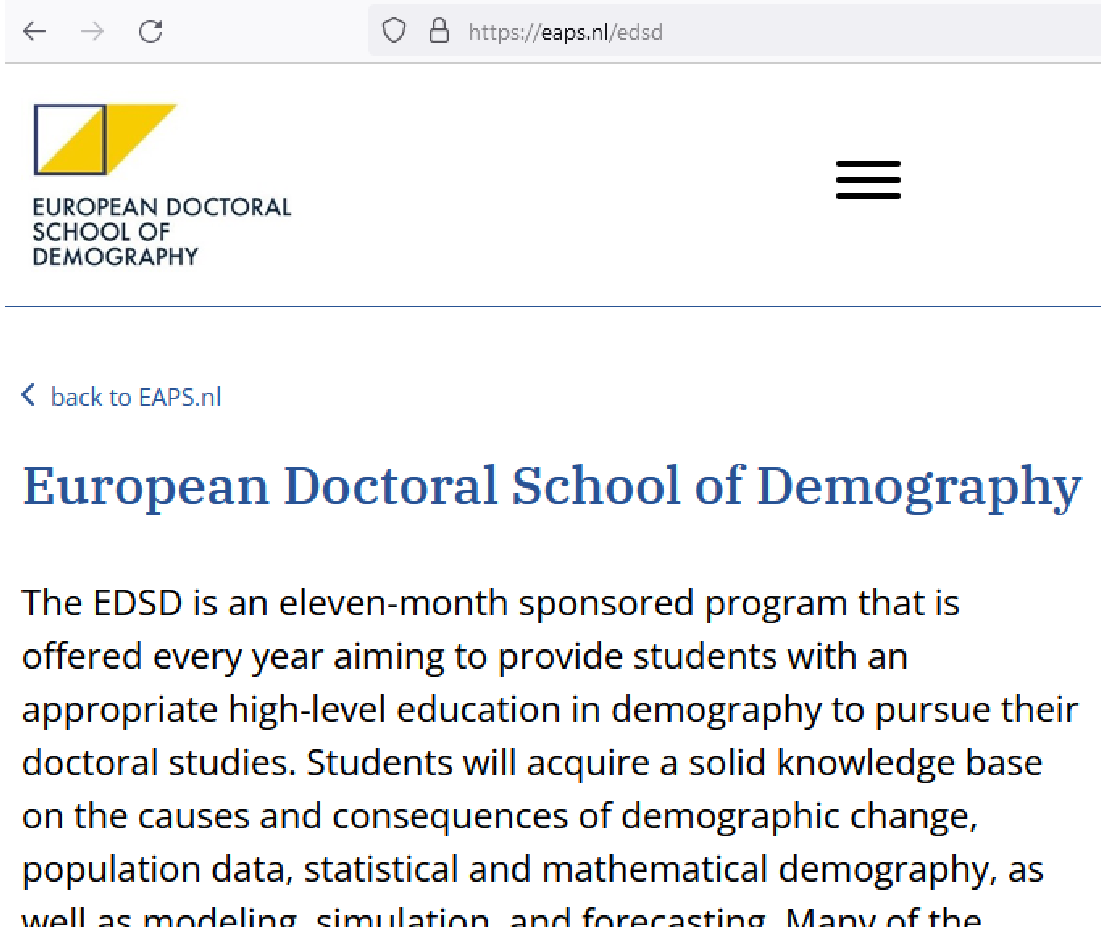
CSS - CASCADING STYLE SHEETS

Selector	Example	Example description
#id	#firstname	Selects the element with id="firstname"
.class	.intro	Selects all elements with class="intro"
element.class	p.intro	Selects only <p> elements with class="intro"
*	*	Selects all elements
Element	p	Selects all <p> elements
element,element,..	div, p	Selects all <div> elements and all <p> elements
element element	div p	Selects all <p> elements inside <div> elements
element>element	div > p	Selects all <p> elements where the parent is a <div> element

- **https://www.w3schools.com/cssref/css_selectors.php**
- **https://developer.mozilla.org/en-US/docs/Web/CSS/CSS_Selectors**

A MORE COMPLICATED WEBSITE

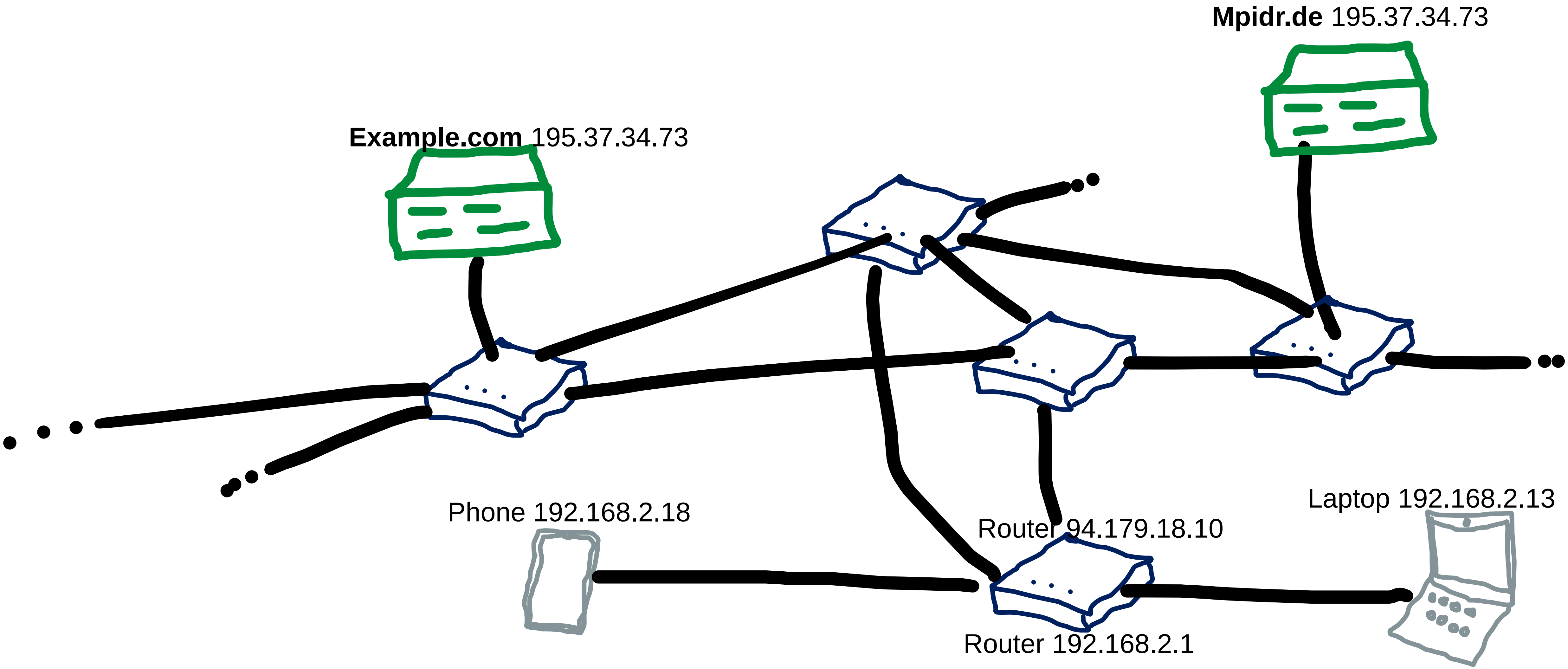
open https://eaps.nl/edsd



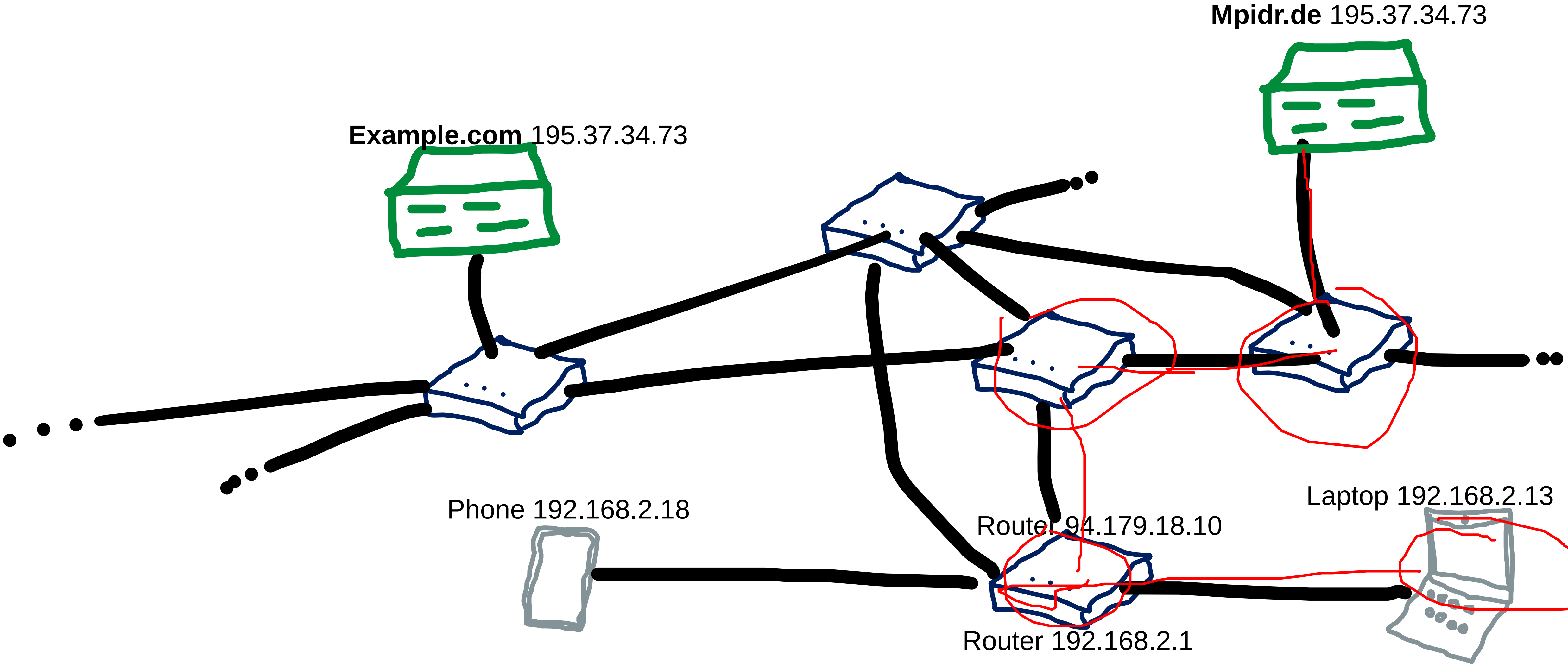
European Doctoral School of Demography

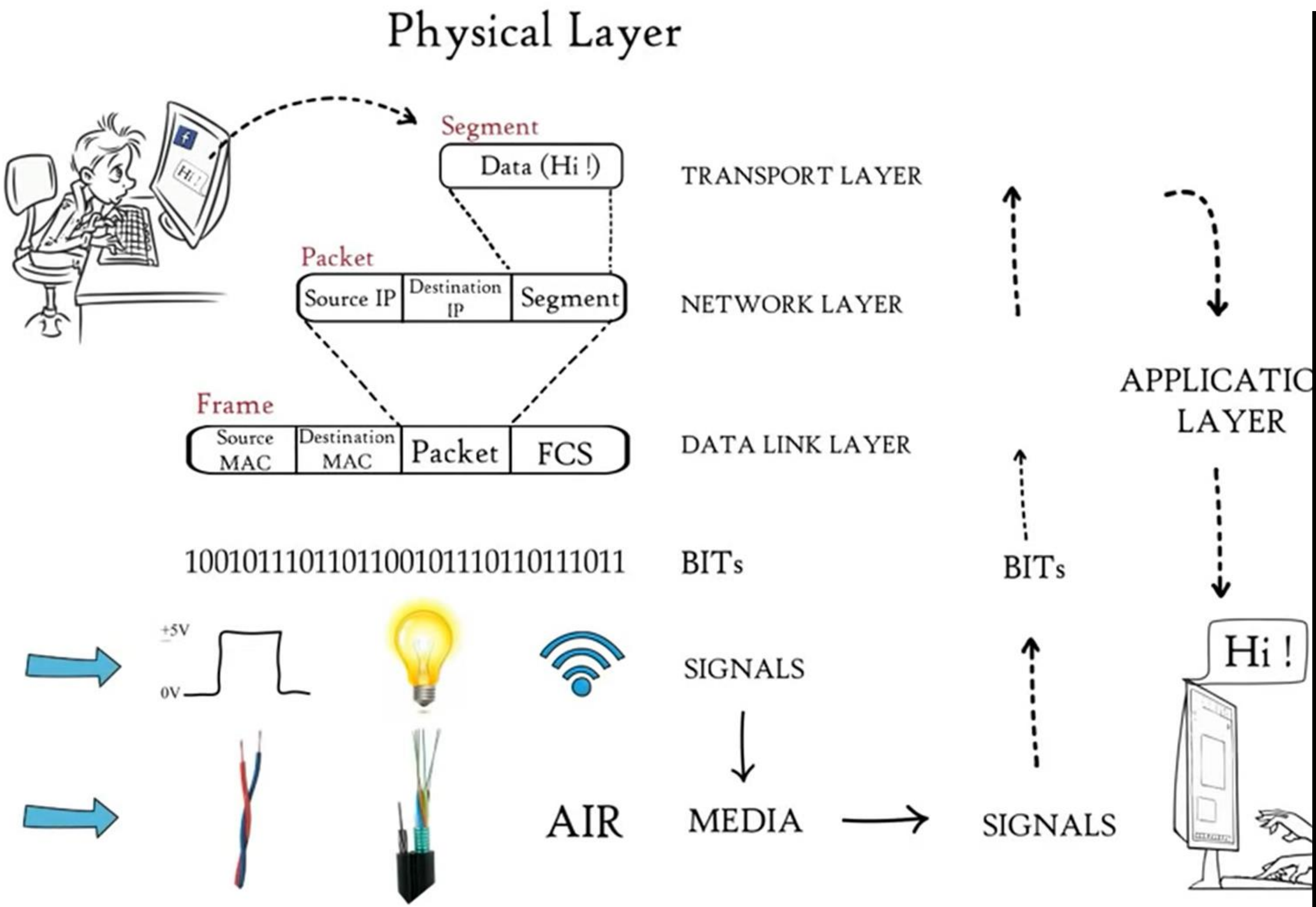
The EDSD is an eleven-month sponsored program that is offered every year aiming to provide students with an appropriate high-level education in demography to pursue their doctoral studies. Students will acquire a solid knowledge base on the causes and consequences of demographic change, population data, statistical and mathematical demography, as well as modeling, simulation, and forecasting. Many of the

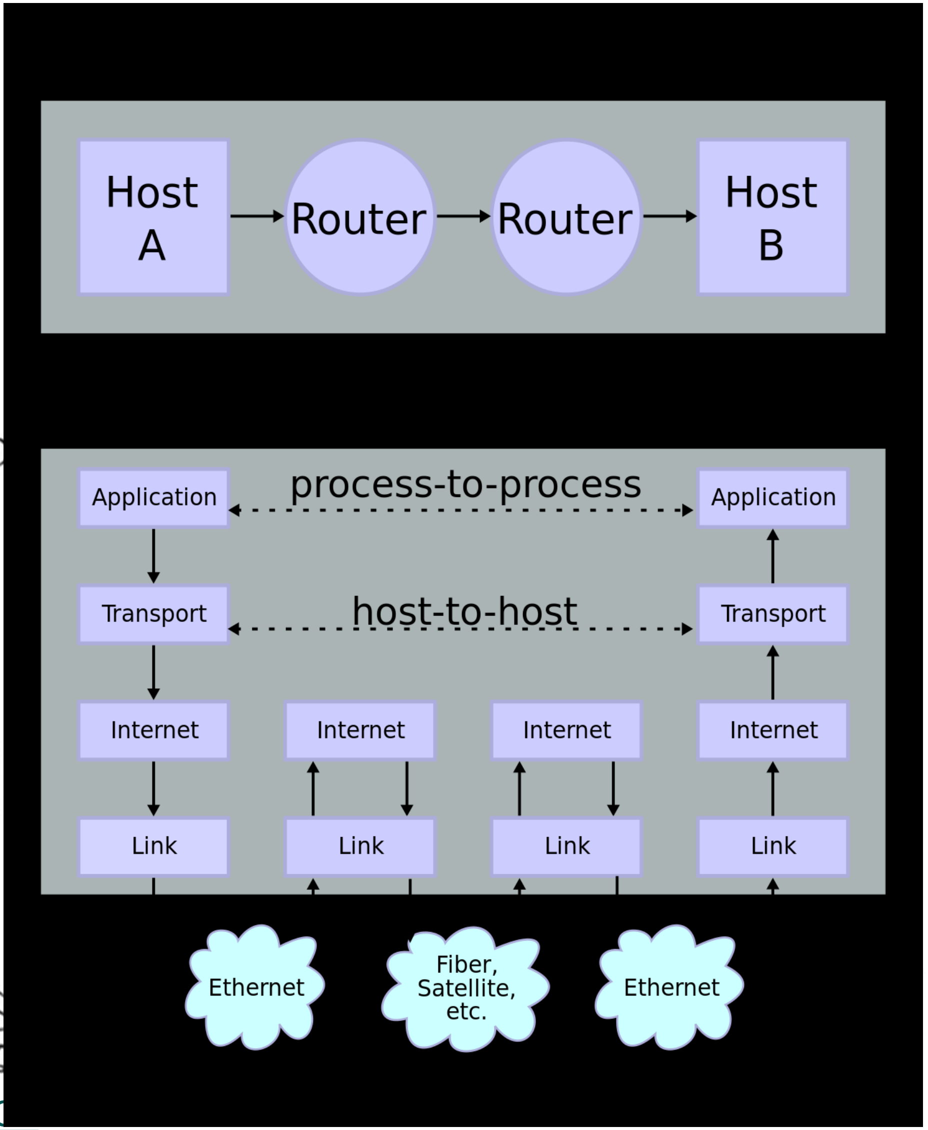
HOW DOES THE REQUEST FIND THE WAY TO THE SERVER?



HOW DOES THE REQUEST FIND THE WAY TO THE SERVER?



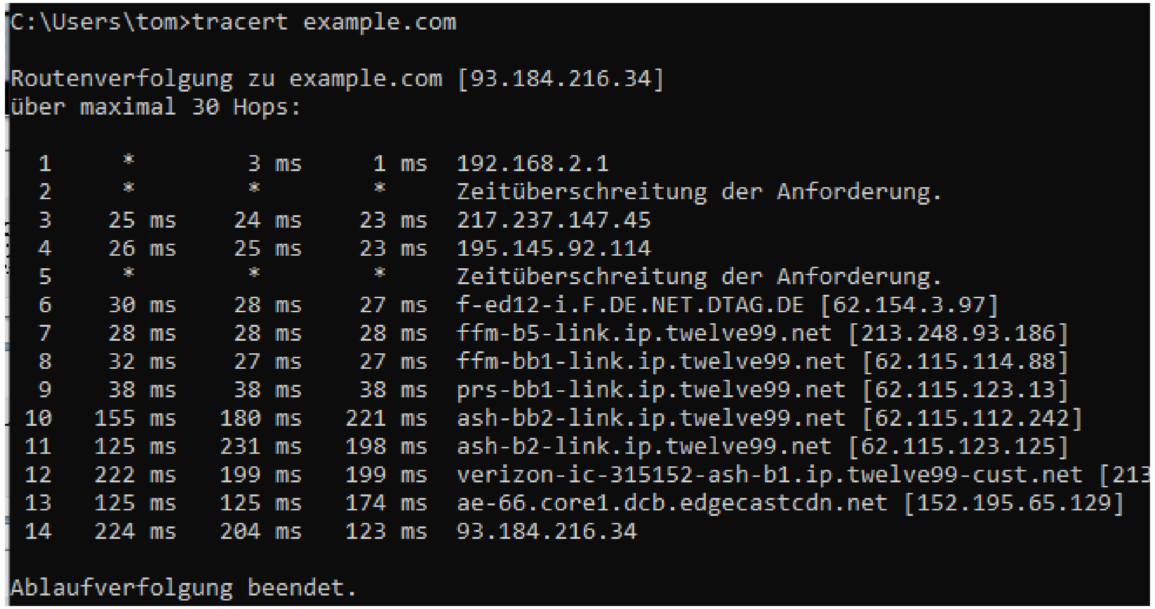




Source: https://www.youtube.com/watch?v=vv4y_uOneC

TRACEROUTE

- 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/

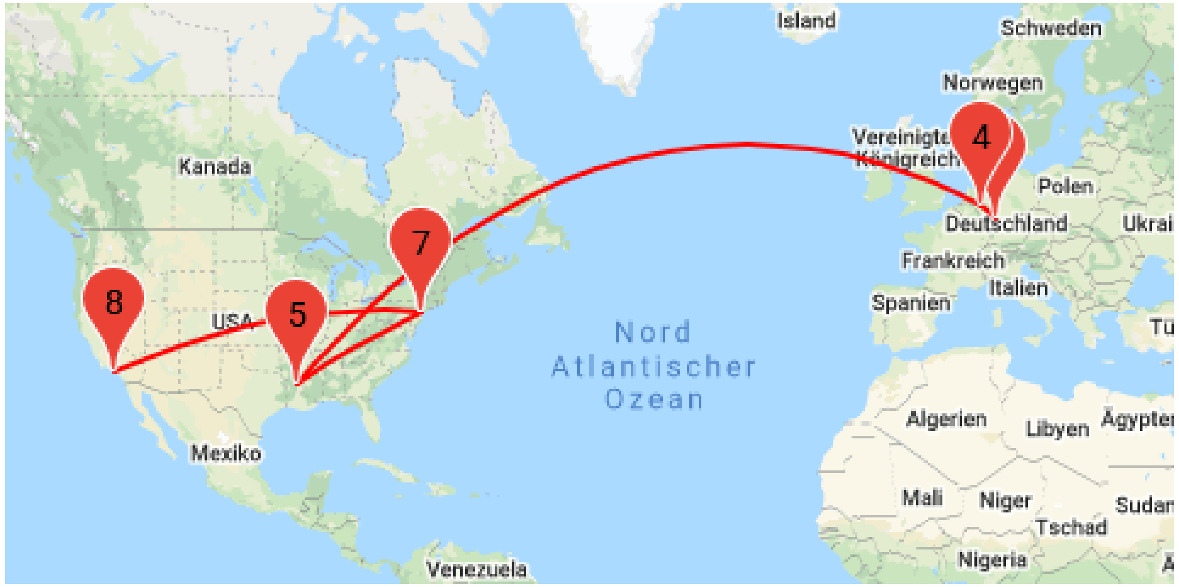


```
C:\Users\tom>tracert example.com

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.
```



PART 2 – SURFING THE WEB WITH R - WEBSCRAPING

Install Selector Gadget from https://selectorgadget.com/

- Show/enable the bookmark-toolbar in your browser
- Drag link to the bookmark toolbar

Alternatively you can install the chrome extension

Open a website, open SelectorGadget and click on the text you want to select.

PART 2 – SURFING THE WEB WITH R - WEBSCRAPING

Please open the script 01_webscraping.R with Rstudio

https://nextcloud.demogr.mpg.de/s/iDQaEGNDqsd5XtQ

https://www.tidyverse.org/

https://rvest.tidyverse.org/

PART 2 – SURFING THE WEB WITH R - WEBSCRAPING

Discussion: Is web scraping legal?

For most owners of webpages, it is fine to scrape their site, as long as you “behave”

“Behaving” means:

- You don’t induce interruptions or unreasonable costs to their service by scraping too fast or too much (good rule of thumb: only scrape one domain every 2 seconds)
- You don’t use the data to the disadvantage of the scraped site. Since content is often copyrighted, you are mostly not allowed to share your scraped data publicly. (Exemption: you alter or aggregate the data enough)
- You respect their robots.txt

Some companies think it is not okay to scrape their webpage - but you are probably still legally allowed to scrape it!

- A US court ruled, that LinkedIn has to remove technical measures that prevented a startup from scraping public profile information of LinkedIn users source
- Sometimes you have to have an account and accept terms and contents which may forbid scraping.

You still have to respect local laws like the GDPR in Europe, which limits the collection of PII

PART 3 - WEB-APIS

- Websites are finicky and prone to change
- CSS-selectors might not work tomorrow
- Webpages are made to be read by humans, not R-scripts

WEB-APIS

APIs!

→ Application Programming Interfaces

→ Interfaces that are meant to be used by machines

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-API
- == webpage for machines

WEB-APIS

Simple APIs, test them in your browser:

- 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

WEB-APIS

WEB-APIS

Please open and follow the second script:

`02_using_web_APIs.R`

WEB-APIS

Please open and follow the second script:

`02_using_web_APIs.R`

DISCUSSION

You have read the paper “Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona’s SB 1070Using Twitter Data” for yesterdays course

Now that you know how to scrape headlines: Imagine that you have dataset with a lot of headlines from various sources of the same timeframe that was handled in the Flores paper. What would be better or worse compared to twitter data?

THANK YOU FOR YOUR ATTENTION!

Tom Theile

Research Software Engineer

theile@demogr.mpg.de