DIGITAL DEMOGRAPHY: ANALYZING WEB AND SOCIAL MEDIA DATA

WEB SCRAPING AND WEB-APIS WITH R

EDSD DECEMBER 2024

TOM THEILE

DEPARTEMENT OF DIGITAL AND
COMPUTATIONAL DEMOGRAPHY

DEPARTMENT OF DIGITAL AND COMPUTATIONAL DEMOGRAPHY

DEPARTEMENT OF DIGITAL AND COMPUTATIONAL DEMOGRAPHY

AGENDA FOR THE WEEK

- Block 1: Internet, Webscraping, APIs
- Block 2: Digital Demography
- Block 3: Kinship Microsimulation with rsocsim
- Block 4: Bibliometric Data for migration research
- Extra: Digital Datasets
- Thursday + Friday: assignment

AGENDA FOR TODAY

- What is the Internet? How do websites work?
  - I will give an introduction on how browsers communicate with servers
- How to scrape data from websites?
  - *We will use R to read websites and extract information from it*
- Simple sentiment analysis
  - *We will use R to analysie the scraped data*

WHAT HAPPENS WHEN WE VISIT A WEBPAGE?

Open the DevTools in your Browser:

`Windows or Linux:`

`F12 or`

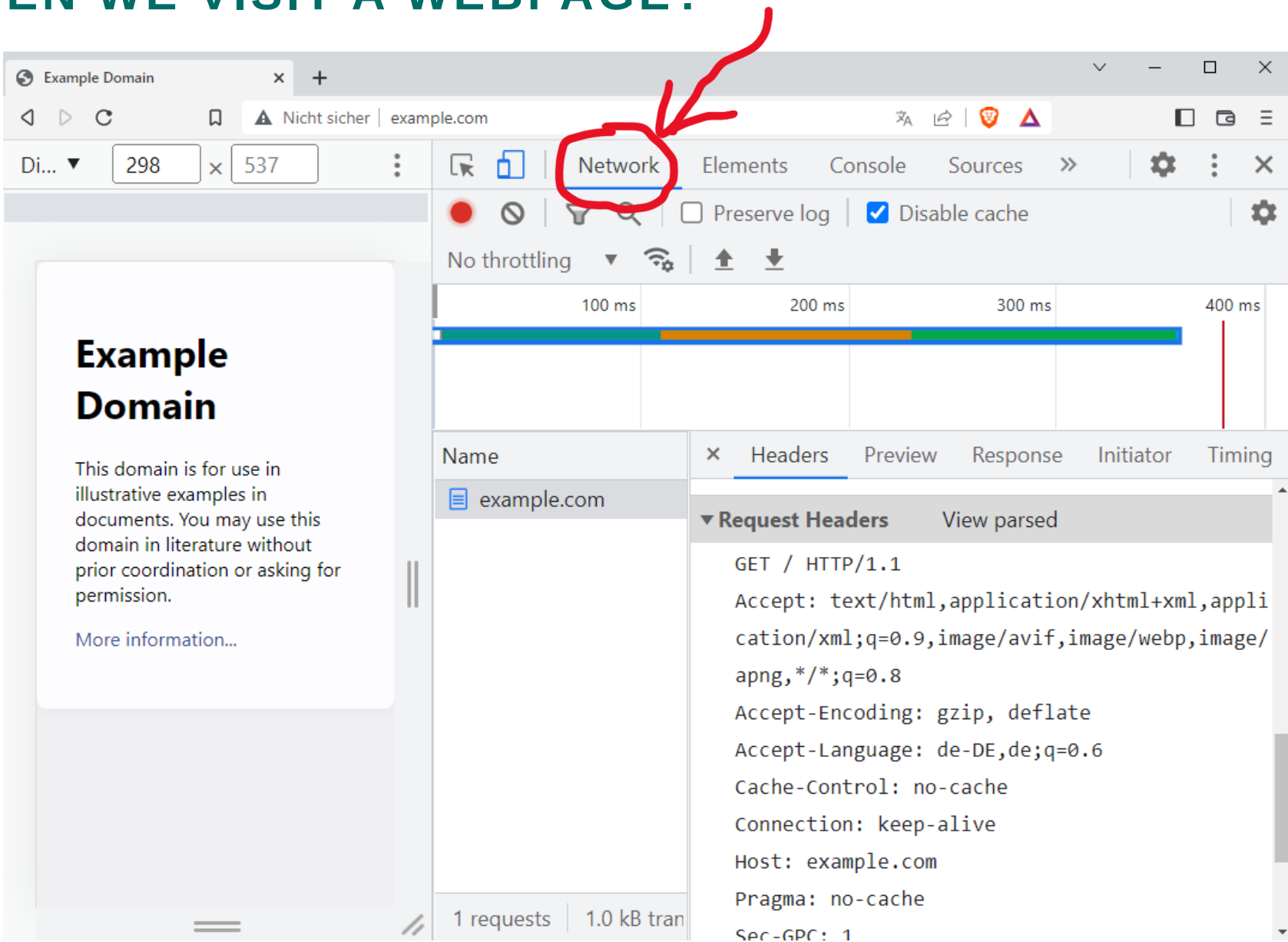
`“CTRL + shift + i”`

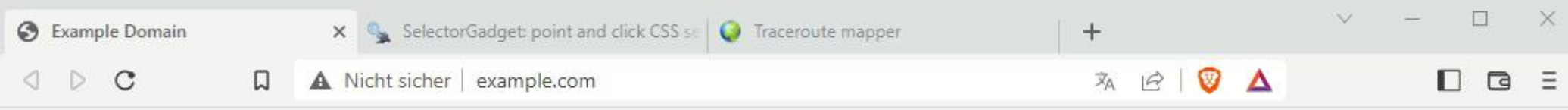
`Mac:`

`“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...

EXAMPLE.COM HTML

```html
<!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>

```

HOW DOES THE REQUEST FIND THE WAY TO THE SERVER?

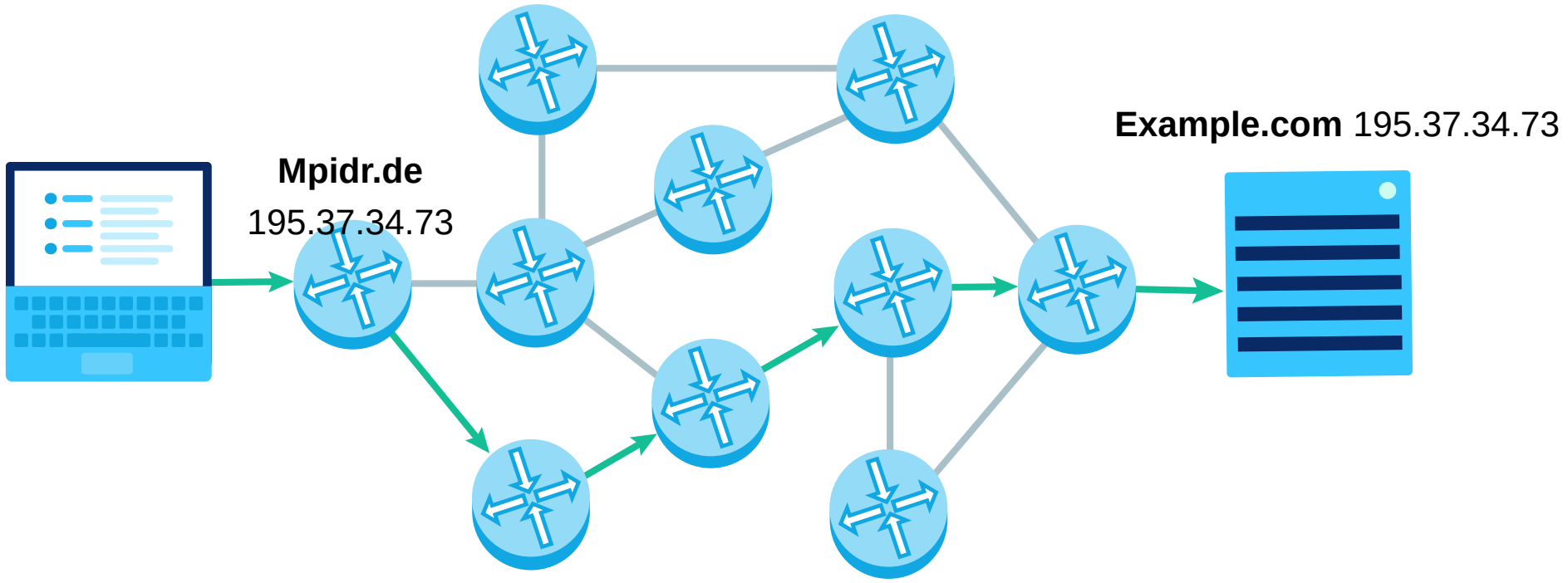


IMAGE SOURCE: HTTPS://WWW.KHANACADEMY.ORG/COMPUTING/COMPUTERS-AND-INTERNET/

HOW DOES THE REQUEST FIND THE WAY TO THE SERVER?

TO:
91.198.174.192

FROM:
216.3.192.1

Mpidr.de
195.37.34.73

Router

IMAGE SOURCE: HTTPS://WWW.KHANACADEMY.ORG/COMPUTING/COMPUTERS-AND-INTERNET/

HOW DOES THE REQUEST FIND THE WAY TO THE SERVER?

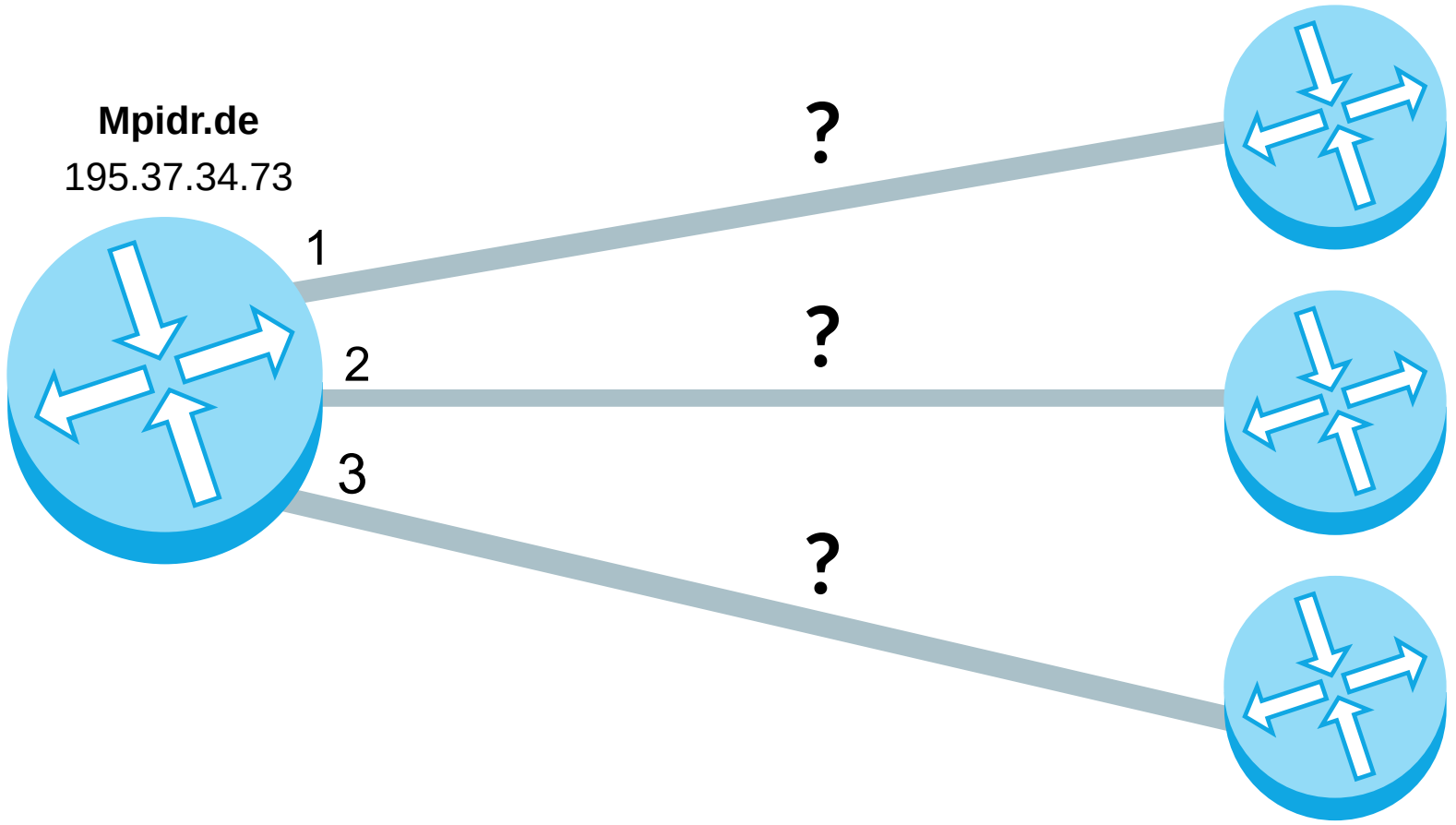
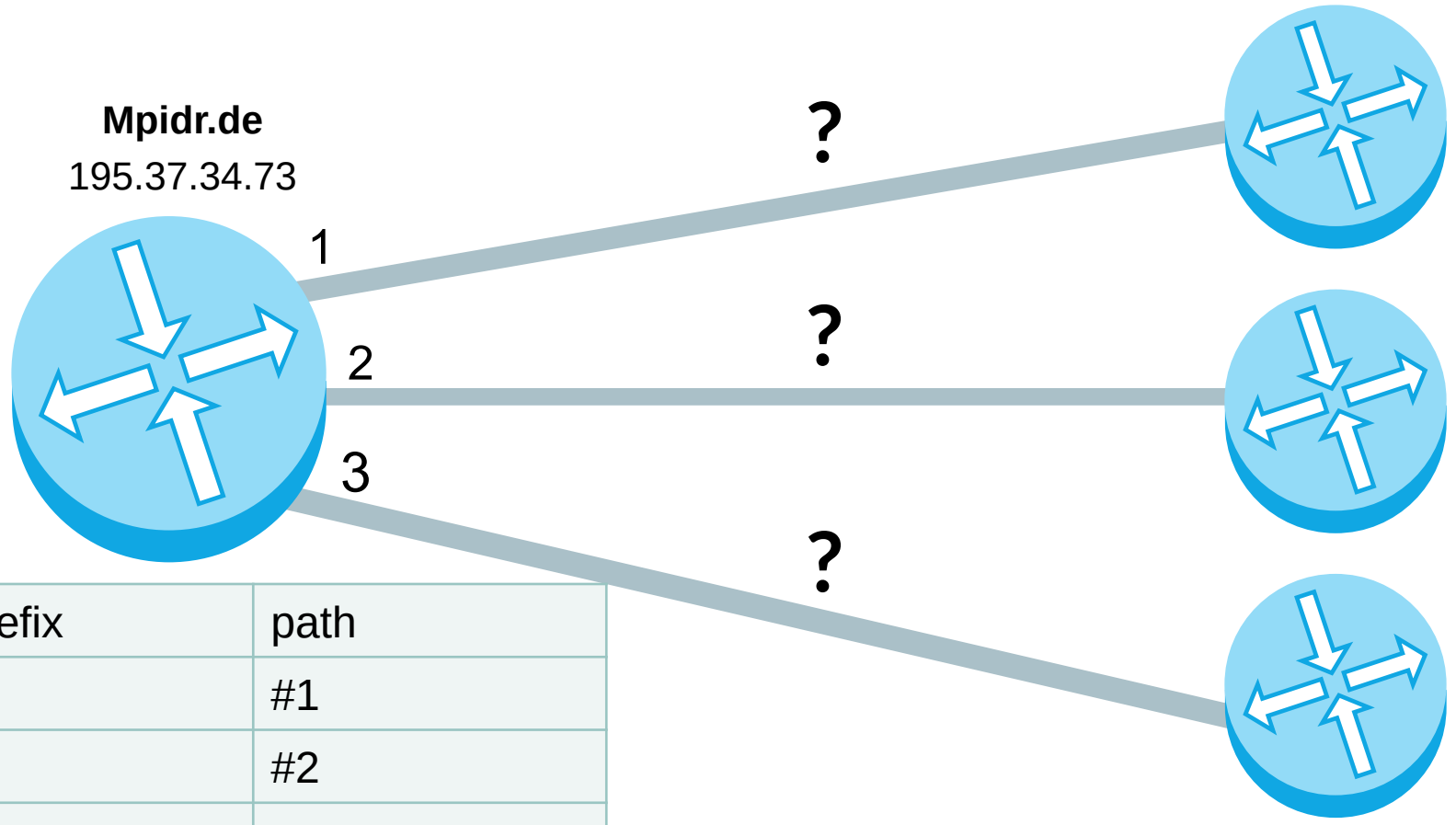


IMAGE SOURCE: HTTPS://WWW.KHANACADEMY.ORG/COMPUTING/COMPUTERS-AND-INTERNET/

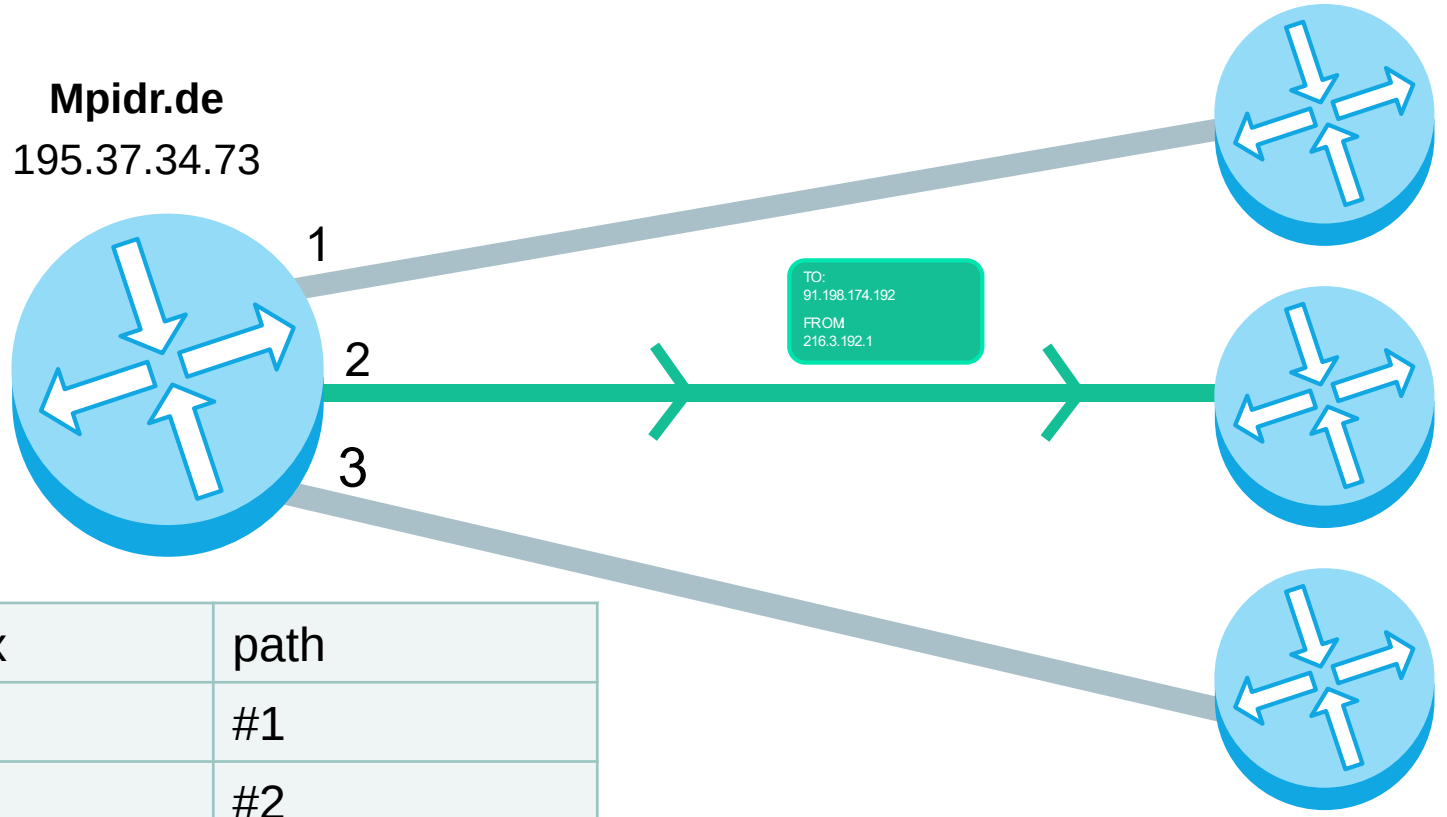
HOW DOES THE REQUEST FIND THE WAY TO THE SERVER?



IP address prefix	path
91.112	#1
91.198	#2
192.92	#3

IMAGE SOURCE: HTTPS://WWW.KHANACADEMY.ORG/COMPUTING/COMPUTERS-AND-INTERNET/

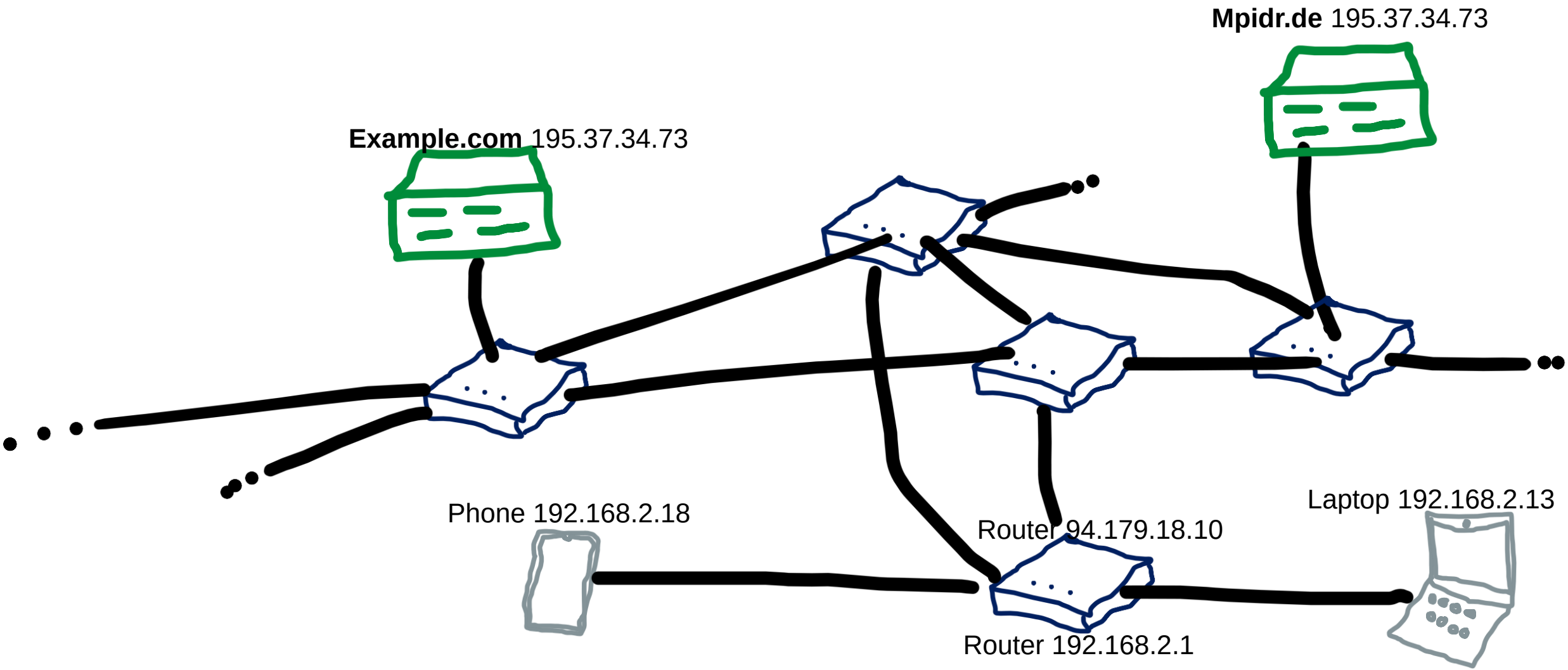
HOW DOES THE REQUEST FIND THE WAY TO THE SERVER?

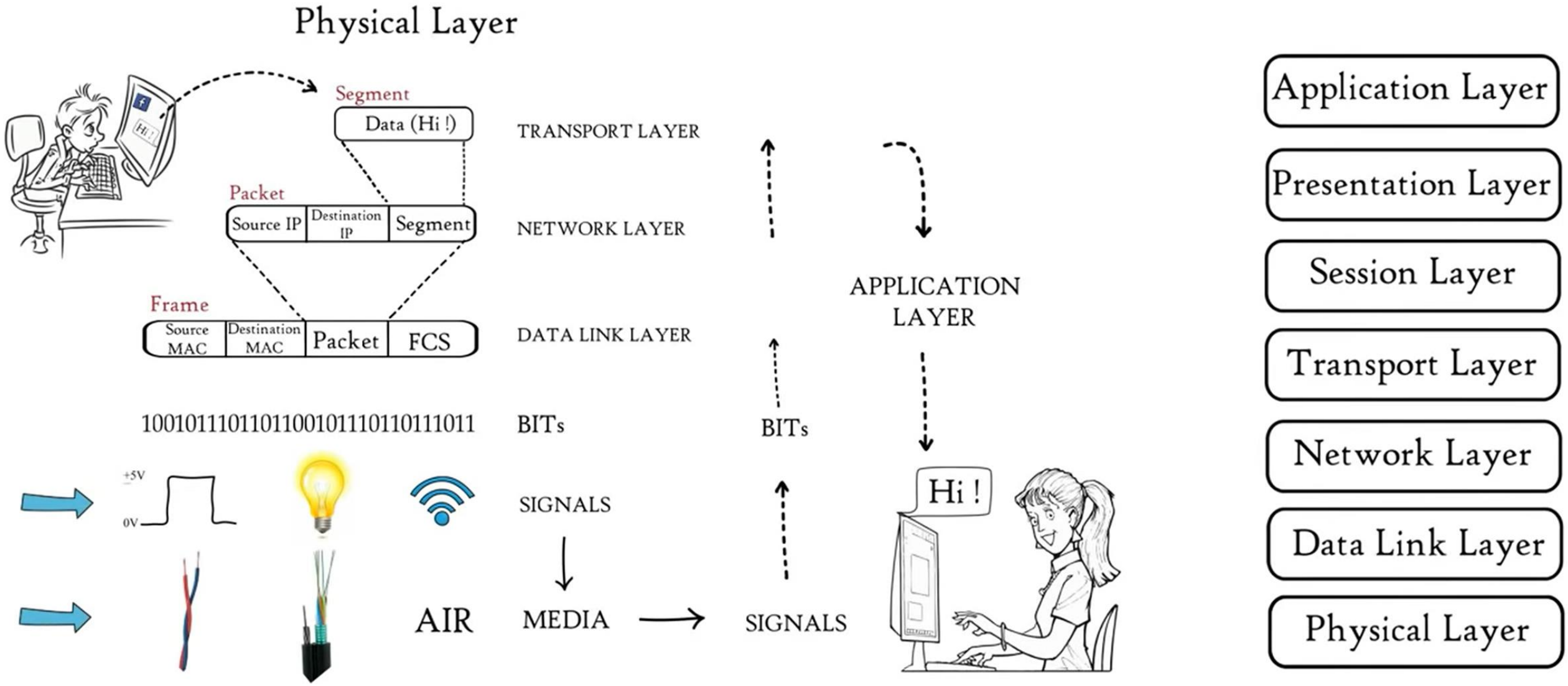


IP address prefix	path
91.112	#1
91.198	#2
192.92	#3

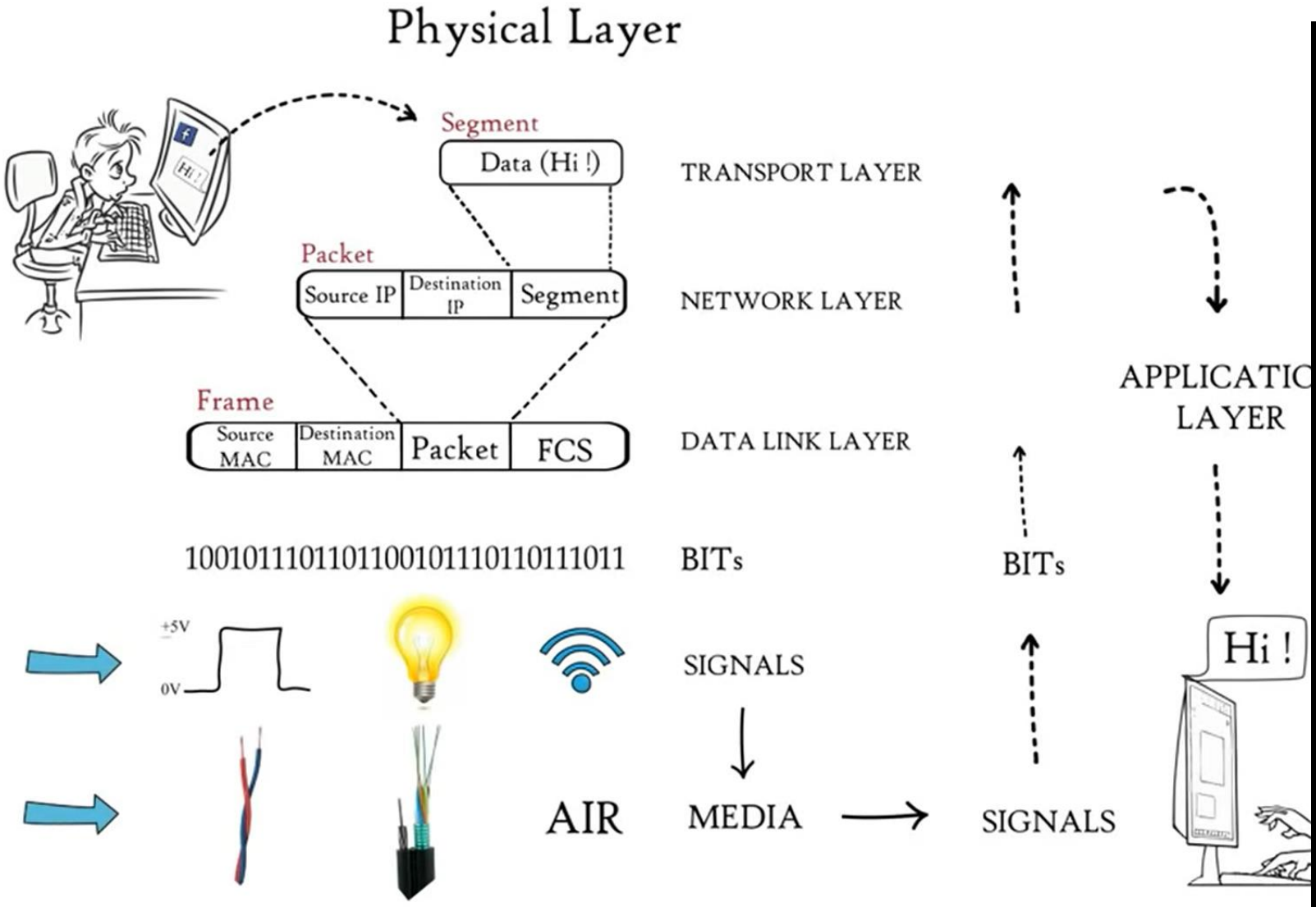
IMAGE SOURCE: HTTPS://WWW.KHANACADEMY.ORG/COMPUTING/COMPUTERS-AND-INTERNET/

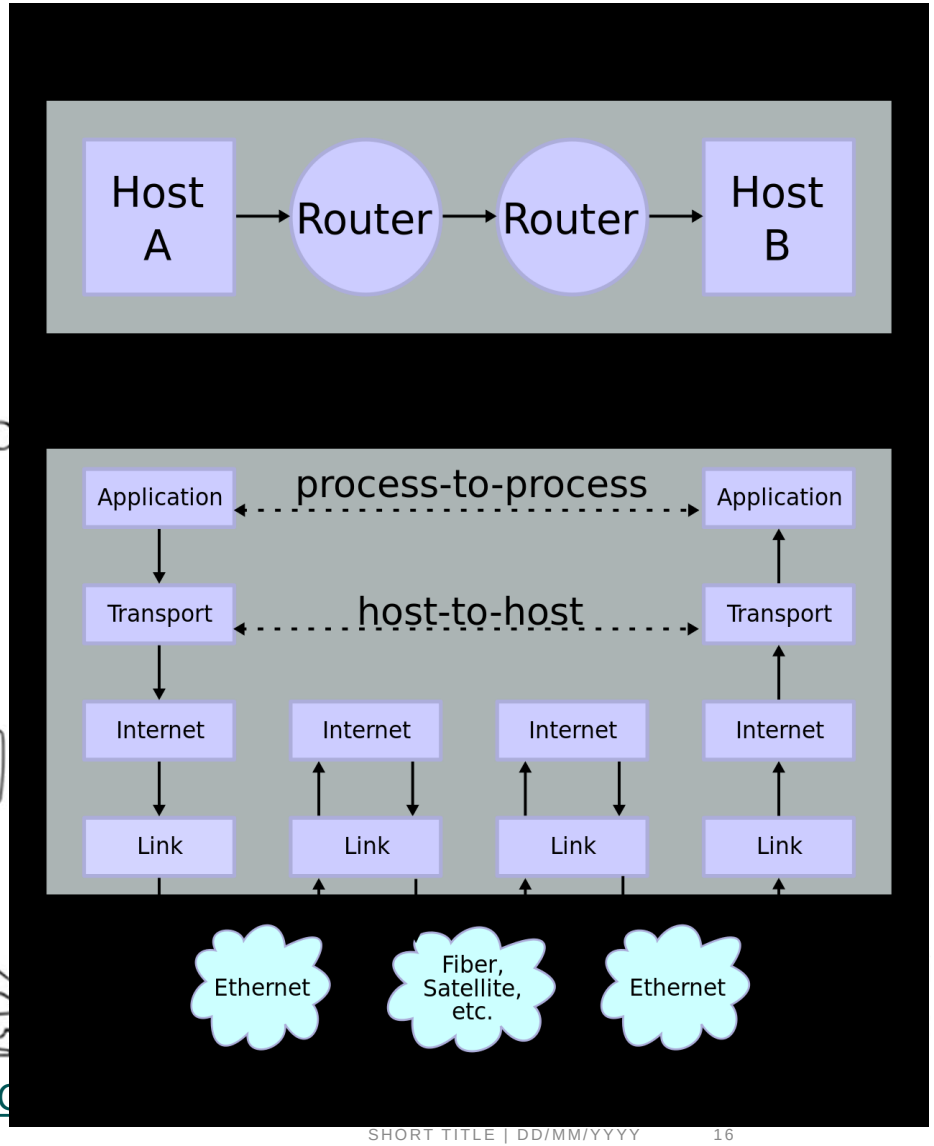
HOW DOES THE REQUEST FIND THE WAY TO THE SERVER?





Source: https://www.youtube.com/watch?v=vv4y_uOneC0 TechTerms Channel





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/ (sometimes out of order)
- Alternative: Geo Traceroute https://geotraceroute.com/ (will do traceroute for you)

```
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.
```

TRACEROUTE

- You can list all the routers on this journey with traceroute (tracert on Windows)
- `Tracert mpidr.de`
- Geo Traceroute https://geotraceroute.com/
- Geo Traceroute

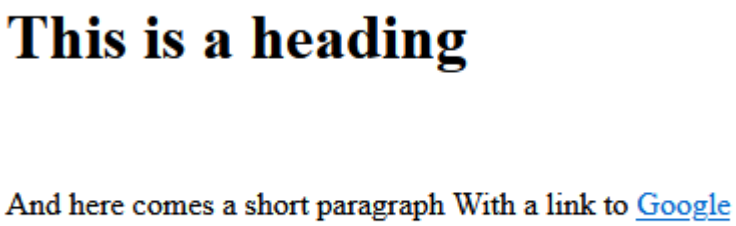
EXAMPLE.COM HTML

- 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>

```

EXAMPLE HTML

```
<!DOCTYPE html>
<html>
<head>
  <title>Page Title</title>
</head>
<body>

  <h1>This is a heading</h1>
  <p>And here comes a short paragraph

With a link to

    <a href="https://google.com">Google</a>

  </p>
</body>
</html>
```

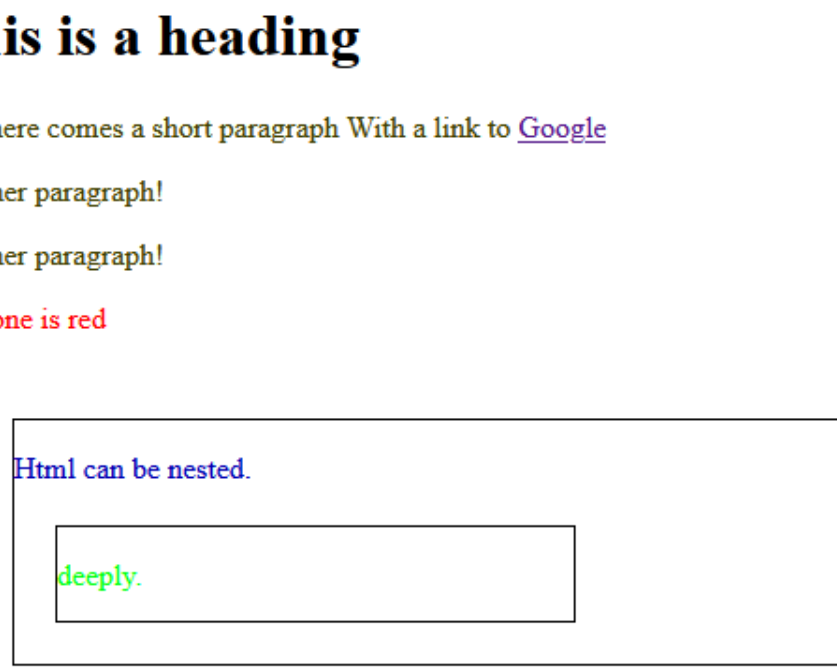


EXAMPLE HTML

```
<!DOCTYPE html>
<html>
<head>
   <title>Page Title</title>

<style type="text/css">

    body {

        background-color: #f0f0f2;

        margin: 0;

        padding: 0;

        font-family: -apple-system, system-ui,
BlinkMacSystemFont, "Segoe UI", "Open Sans",
"Helvetica Neue", Helvetica, Arial, sans-serif;

    }

    div {

        width: 600px;

        margin: 5em auto;

        padding: 2em;

        background-color: #fdfdff;

        border-radius: 0.5em;

        box-shadow: 2px 3px 7px 2px rgba(0,0,0,0.02);

    }

    a:link, a:visited {

        color: #38488f;

        text-decoration: none;

    }

    @media (max-width: 700px) {

        div {

            margin: 0 auto;

            width: auto;

        }

    }

    </style>
</head>
<body>

   <h1>This is a heading</h1>
   <p>And here comes a short paragraph

With a link to

      <a href="https://google.com">Google</a>

   </p>
</body>
</html>
```

EXAMPLE HTML

```
<!DOCTYPE html>
<html>
<head>
   <title>Page Title</title>

<style type="text/css">

    div {border:  1px solid;

         width: 50%;

         margin: 4%;}

    p {color: #404000; }

    #theredone {color: #FF0000; }

    div > p {color: #0000B0; }

    .allgreen {color: #00FF0F;}

</style>

</head>

<body>

 <h1>This is a heading</h1>

 <p>And here comes a short paragraph

 With a link to <a
href="https://google.com">Google</a>

 </p>

 <p>Another paragraph!</p>

 <p>Another paragraph!</p>

 <p id="theredone">This one is red</p>

 <div>

  <p>Html can be nested.</p>

  <div>

    <p class="allgreen">deeply.</p>

  </div>

 </div>

</body>

</html>
```



CSS SELECTORS

Selector	Example	Example description
.class	.intro	Selects all elements with class="intro"
.class1.class2	.name1.name2	Selects all elements with both *name1* and *name2* set within its class attribute
.class1 .class2	.name1 .name2	Selects all elements with *name2* that is a descendant of an element with *name1*
#id	#firstname	Selects the element with id="firstname"
*	*	Selects all elements
element	p	Selects all <p> elements
element.class	p.intro	Selects all <p> elements with class="intro"
element,element	div, p	Selects all elements and all <p> elements
element element	div p	Selects all <p> elements inside elements
element>element	div > p	Selects all <p> elements where the parent is a element
element+element	div + p	Selects the first <p> element that is placed immediately after elements
element1~element2	p ~ ul	Selects every element that is preceded by a <p> element
[*attribute*]	[target]	Selects all elements with a target attribute
[*attribute=value*]	[target="_blank"]	Selects all elements with target="_blank"

https://www.w3schools.com/cssref/css_selectors.php

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. For more information you can watch the video on https://selectorgadget.com/

PART 2 – SURFING THE WEB WITH R - WEBSCRAPING

Please open the script 01_webscraping.R with Rstudio

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 copy-righted, you are mostly not allowed to share your scraped data publicly. (Exemption: you alter or aggregate the data enough; the data is not copy-rightable)
- 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

WEB SCRAPING WITH SELENIUM

Some websites heavily rely on Javascript to fetch content and show it to the user.

These websites are often hard to scrape with rvest (which can not execute Javascript)

Selenium lets you control a Browser (Firefox or Chrome) from R (and other languages)

The script 06_scraping_with_selenium.R explains the installation and basic usage of Selenium with the R package rselenium.

I provided this only for those who are interested in selenium. It is not necessary for this course

REVERSE ENGINEERING HIDDEN APIS

Some websites heavily rely on Javascript to fetch content and show it to the user.

We will learn another method to scrape these sites after we learned about APIs.

THANK YOU FOR YOUR ATTENTION!

Tom Theile

Research Software Engineer

theile@demogr.mpg.de