Food

Every Day except next Wednesday:

Coupon = 1 Meal 1 Drink for max. 9.90€

Next Wednesday

Coupon = 1 Meal 1 Drink for max. 7.00€

The difference between the actual cost and the coupon maximum cannot be used!

What is digital trace data and how do we collect it?

Game plan

1. What is digital trace data?
2. Intro API
3. Intro Web scraping
4. API vs Web scraping
5. Group Exercises
   - API
   - Web scraping (with `RSelenium` and `rvest`)

What is digital trace data?

https://commons.wikimedia.org/wiki/File:Duchamp_Fountaine.jpg
https://commons.wikimedia.org/wiki/File:%27David%27_by_Michelangelo_JBU0001.JPG

What is digital trace data?

For our purposes:

Data that is not created for the purpose of being analyzed by social science researchers, but is a byproduct of everyday online activity.

E.g.,
Mobile phone location data
Social media conversations and friend networks
Google search data
Newspaper articles
Parliamentary protocols
And so much more

Benefits and issues of digital trace data

Work in groups to discuss the potential benefits and problems associated with the use of digital trace data for research purposes.

10 min

Benefits and issues of digital trace data

- Big (enables analysis of small differences/prevelance)
- Always on (enable capturing of rare and surprising events)
- Non-Reactive
- Captures Social Relationships

- Big (difficult to handle)
- Non-Representative
  - Biases depending on platform
- Drifting
- Algorithmic Confounding
- Unstructured and noisy
- Sensitive
- Incomplete (e.g., demographic info)

- Accessibility

Collecting digital trace data

Our focus: Textual data

Two main ways of collecting text data online:

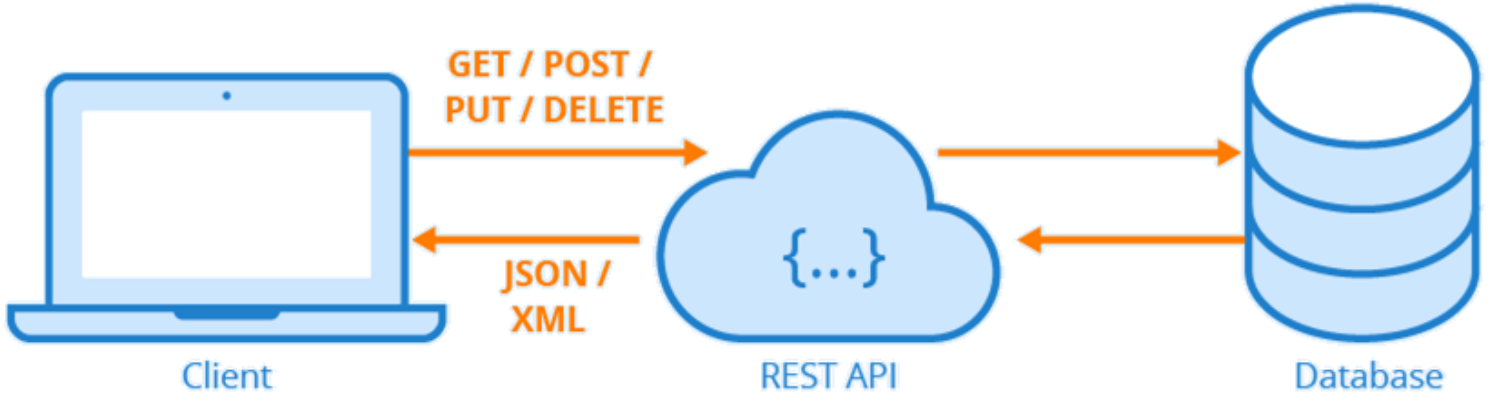
API

Web scraping

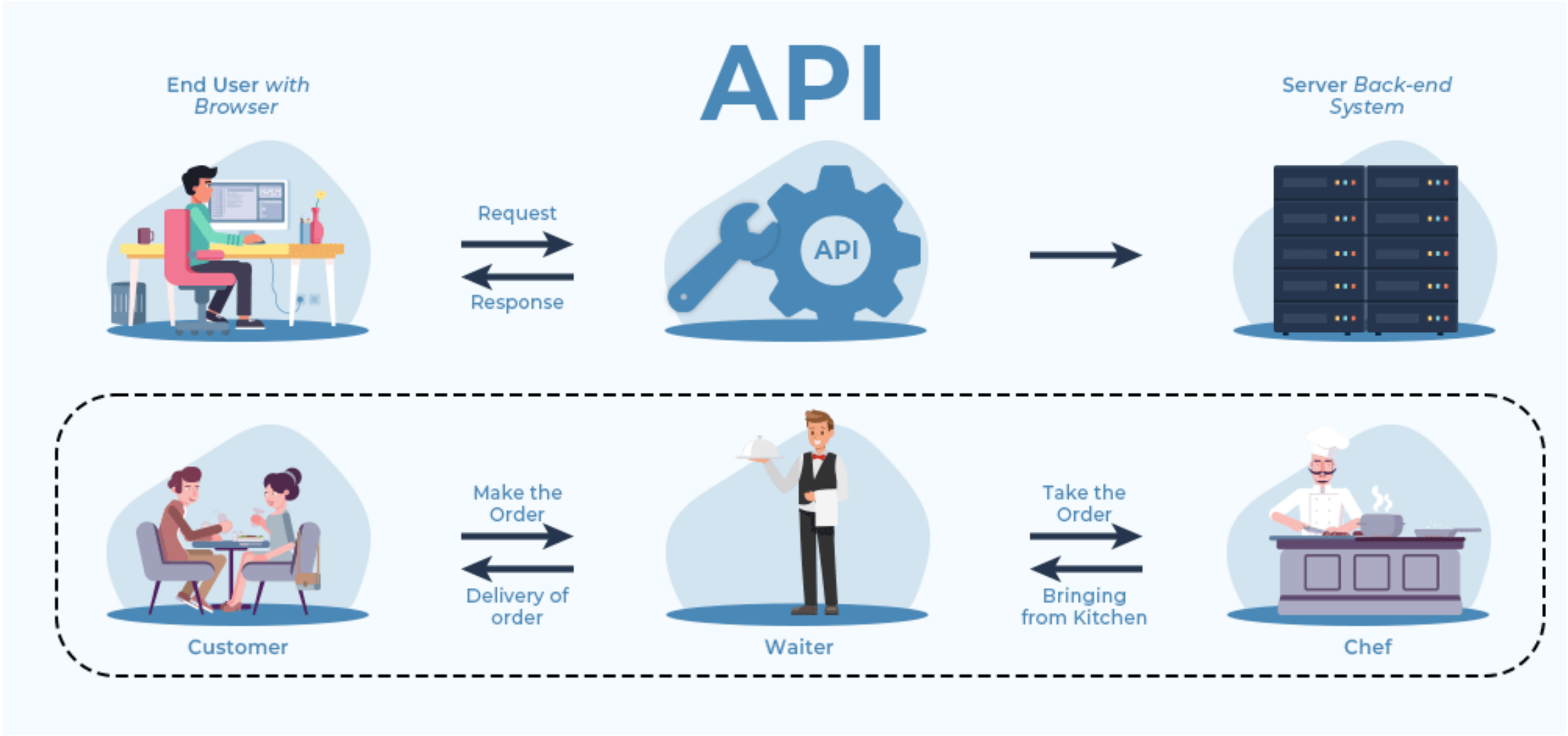
API Intro

What is an API

- Aplication Programming Interface
- An interface provided by the data base owner which enables you to access data on their server conveniently



What is an API



How do we make an order?

https://api.genderize.io?name=anna&country_id=DE

What is an URL

APIs are always accessed via an URL, therefore it is important to know how an URL is actually structured.

Protocol/ scheme — Domain — Path — Query

https://www.website.com/api/cheese/cheesecake?color=yellow&form=circular

? : 1

& : n

What is an URL

Protocol Domain Query

https://api.genderize.io?name=anna&country_id=DE

How do we know which queries to use?

Documentation!

Example: https://api.congress.gov/

API Authentification

Different forms of authentication.

- None
- API key (fully open, registration)
- client key + secret key (mostly for sensitive or paid data)
- OAuth2 (most secure, involves separate authentication server)

API Authentification

Do not save your key directly in your script.

Instead you can use environment variables:

Run: `savefile.edit("~/.Renviron")`

Write: `"key = [your key]“`

Save and restart R

Run: `viaSys.getenv("key")`

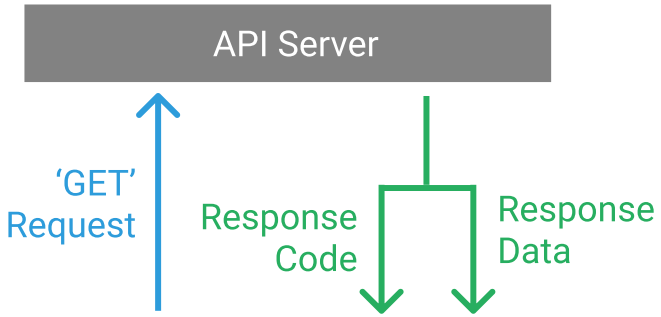
API call, example with `httr`

https://www.website.com/api/cheese/cheesecake?color=yellow&form=circular

```
httr_rec <- GET(
   "https://www.website.com/api/cheese/cheesecake",
   query = list(color = yellow,
                form = circular,
                api_key = viaSys.getenv("key")))
```

Note, that in some cases you may want to also supply a header (mostly for authentication), see `?httr::add_headers`

API response



- An HTTP status code (200 is what you want)
- Headers
- A body typically consisting of XML, JSON, plain text, HTML, or some kind of binary representation.

    Extract body using `content()` from the `httr` package

Tasks

API

Use the API provided by https://api.congress.gov/ to

1. Names and other information on **members** of congress. Save the resulting data on your hard drive.
2. **Congressional records** from 2020 to 2023 (only includes URLs to records). Save the resulting data on your hard drive.

https://github.com/StefanMunnes/SICSS_2023

It does not matter how the data looks in the end. We will learn how to clean and prepare web/text data tomorrow!