SICSS Berlin – Day 3

Problem solving

But!

But!

RSelenium?

Preprocessing

Cleaning, preparing, ..

Game plan

- From list to data frame
- Regular Expressions
- String operations

- Exercise: Preparing data for tomorrow

Data formats

API response normally in JSON or XML format

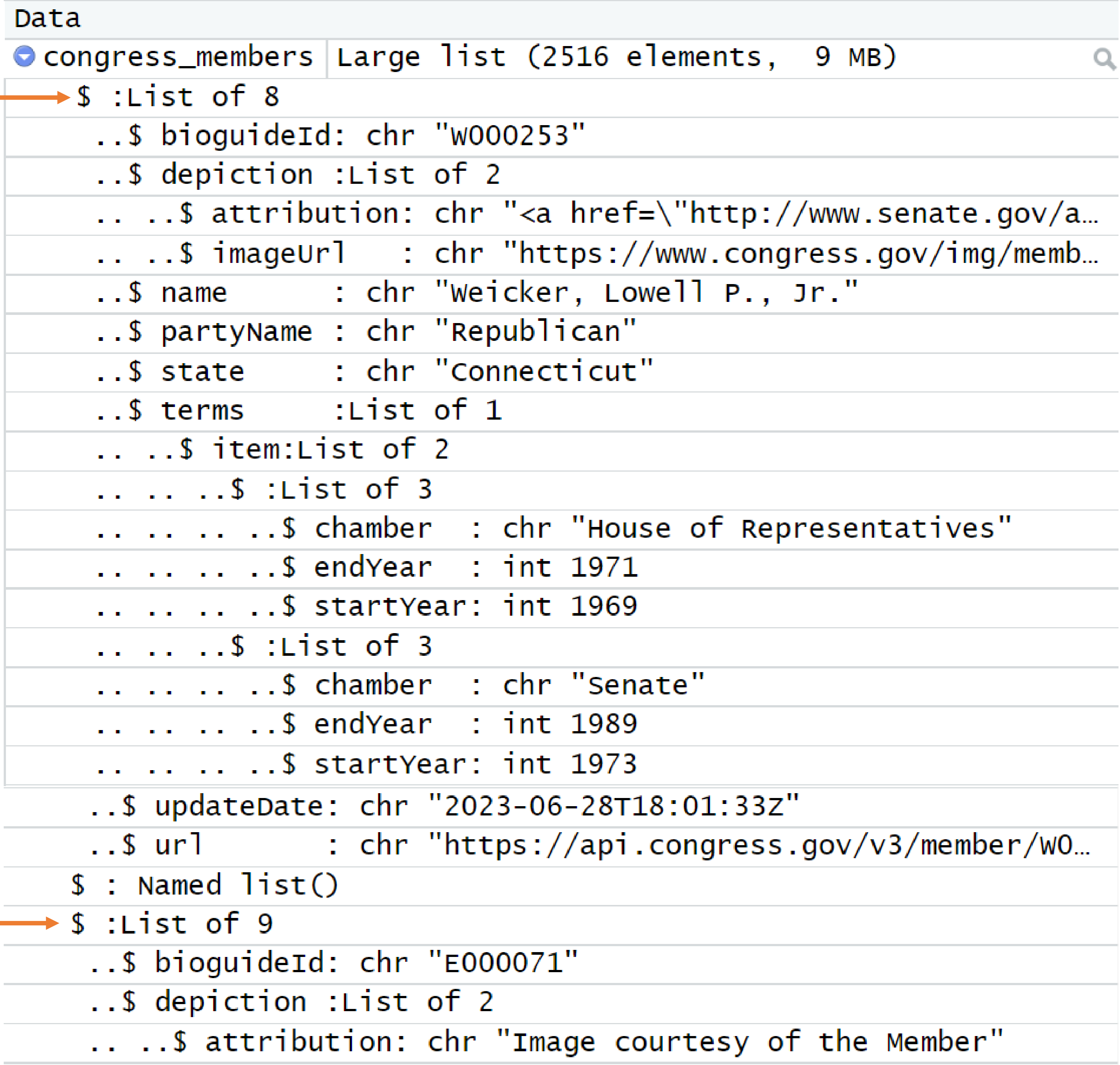
R functions automatically translate these into lists (if not, you can use packages like `jsonlite, xml2`)

We will not deal with either format directly.

Instead, transform into data frames.

From list to data frame

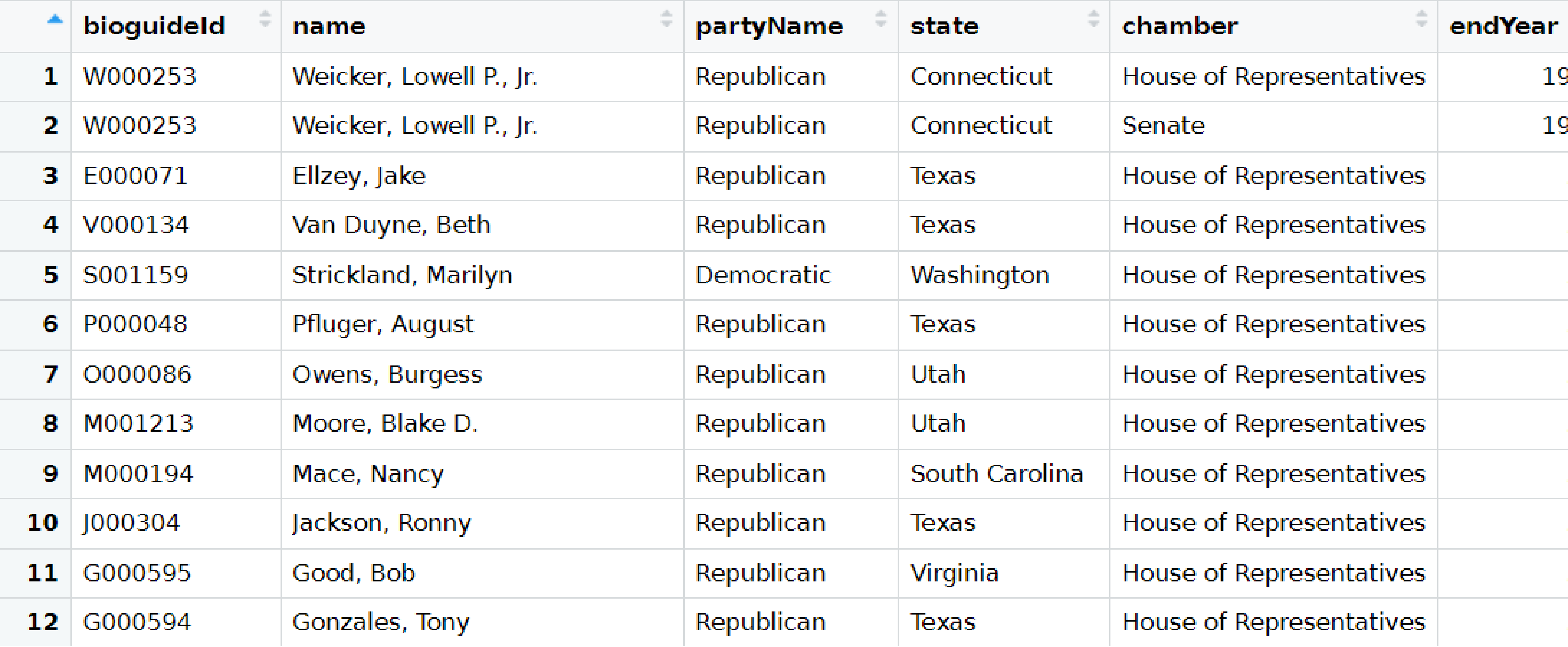
- Yesterday: Extracted lists of US congress members via API
- Instead use list that I provide
- Large list of 2516 elements



```
Data
congress_members  Large list (2516 elements,  9 MB)
  $ :List of 8
   ..$ bioguideId: chr "W000253"
   ..$ depiction :List of 2
   .. ..$ attribution: chr "<a href=\"http://www.senate.gov/a…
   .. ..$ imageUrl   : chr "https://www.congress.gov/img/memb…
   ..$ name       : chr "Weicker, Lowell P., Jr."
   ..$ partyName  : chr "Republican"
   ..$ state      : chr "Connecticut"
   ..$ terms      :List of 1
   .. ..$ item:List of 2
   .. .. ..$ :List of 3
   .. .. .. ..$ chamber  : chr "House of Representatives"
   .. .. .. ..$ endYear  : int 1971
   .. .. .. ..$ startYear: int 1969
   .. .. ..$ :List of 3
   .. .. .. ..$ chamber  : chr "Senate"
   .. .. .. ..$ endYear  : int 1989
   .. .. .. ..$ startYear: int 1973
   ..$ updateDate: chr "2023-06-28T18:01:33Z"
   ..$ url       : chr "https://api.congress.gov/v3/member/W0…
  $ : Named list()
  $ :List of 9
   ..$ bioguideId: chr "E000071"
   ..$ depiction :List of 2
   .. ..$ attribution: chr "Image courtesy of the Member"
```

First member

Second member



	bioguideId	name	partyName	state	chamber	endYear
1	W000253	Weicker, Lowell P., Jr.	Republican	Connecticut	House of Representatives	19
2	W000253	Weicker, Lowell P., Jr.	Republican	Connecticut	Senate	19
3	E000071	Ellzey, Jake	Republican	Texas	House of Representatives	
4	V000134	Van Duyne, Beth	Republican	Texas	House of Representatives	
5	S001159	Strickland, Marilyn	Democratic	Washington	House of Representatives	
6	P000048	Pfluger, August	Republican	Texas	House of Representatives	
7	O000086	Owens, Burgess	Republican	Utah	House of Representatives	
8	M001213	Moore, Blake D.	Republican	Utah	House of Representatives	
9	M000194	Mace, Nancy	Republican	South Carolina	House of Representatives	
10	J000304	Jackson, Ronny	Republican	Texas	House of Representatives	
11	G000595	Good, Bob	Republican	Virginia	House of Representatives	
12	G000594	Gonzales, Tony	Republican	Texas	House of Representatives	

From list to data frame – Exercise

Use congress_members.Rds

Goal:

- Data should be in data frame format and
- should not *not* include any nested lists and
- should not include *redundant* rows

Information it should hold in the end:

ID, name, party name, state, district, URL, update date and start and end year of the members term.

Regular expression

- "pattern-matching notation"[1]
- "is a sequence of characters that specifies a match pattern in text"[2]
- Very important for string operations
- But intimidating on first sight:

  ```
  "^(.[A-Za-Z]+(\\s[A-Za-Z]\\.)?(?!,))\\s(.+(-.+)?)"
  ```
- Easier example:

```
"^\w+\.\d+@\w+\.\w{2}"
sicss.2023@wzb.eu
peter.92@gmx.de
houses.123@hotmail.de
```

Regular expression – Exercise

https://regexone.com/

Regular expression

Groups

```
"\s([A-C]).{2}([Ff].)\."
```

Group \1 and group \2

Look arounds

Take into account what appears before or after the pattern of interest.

`a(?=c)`	a followed by c
`a(?!c)`	a not followed by c
`(?<=b)a`	a preceded by b
`(?<!b)a`	a not preceded by b

Regular expression – Notes on R

- Instead of one slash to escape a character, you need two \\s instead of \s
- Some functions do not recognize line breaks as "Any character", you have to add \\n explicitly.

String operations – Exercise

Complete the tasks in 2_string_operations.R

The `stringr` cheat sheet will help you figure out what to do.

Cleaning scraped data – Exercise

Now for the real challenge.

Solve the tasks in the **3_cleaning.R** script using the **CNN_complete.Rds** dataset I provided.

Cleaning scraped data – Exercise

Notes:

If you laptop has trouble with these operations, reduce the number of articles with `cnn_data <- sample_n(cnn_data, 400)`

If you finish early, I have a bonus task. Just ask me.