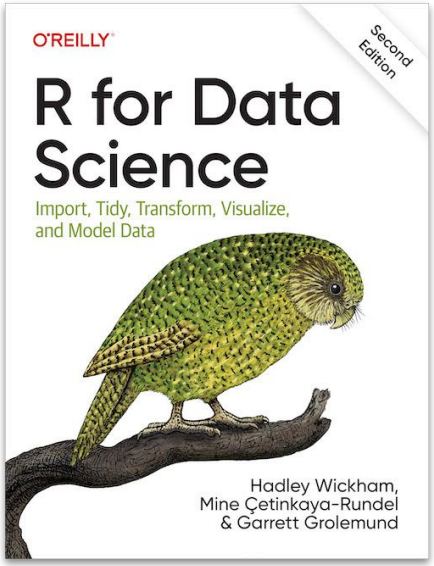
SEARCHING TEXT

R & stringr

CONTENT

- Why is text different?
- Working with text data using `{stringr}`
  - Searches in text
  - Extract matches
  - Replace matches
  - Split text

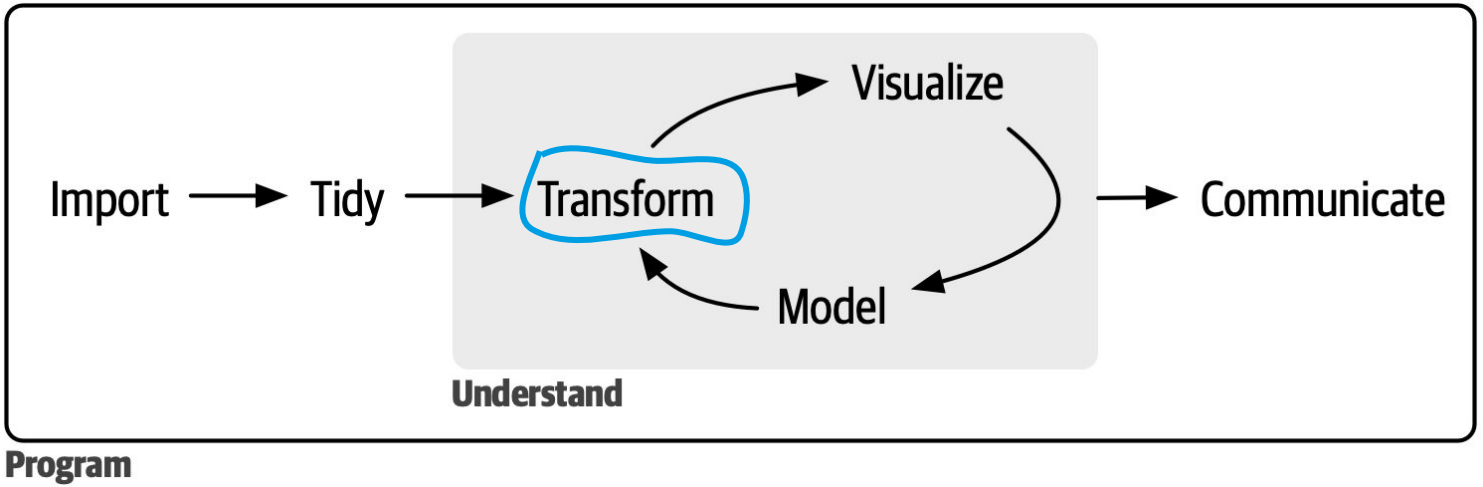
RECOMMENDED LITERATURE



Wickham, Hadley, and Garrett Grolemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. 2nd edition, O'Reilly, 2023. Available online: https://r4ds.hadley.nz/

→ Chapter 14 - Strings
→ Chapter 15 - Regular Expressions

WHERE ARE WE?

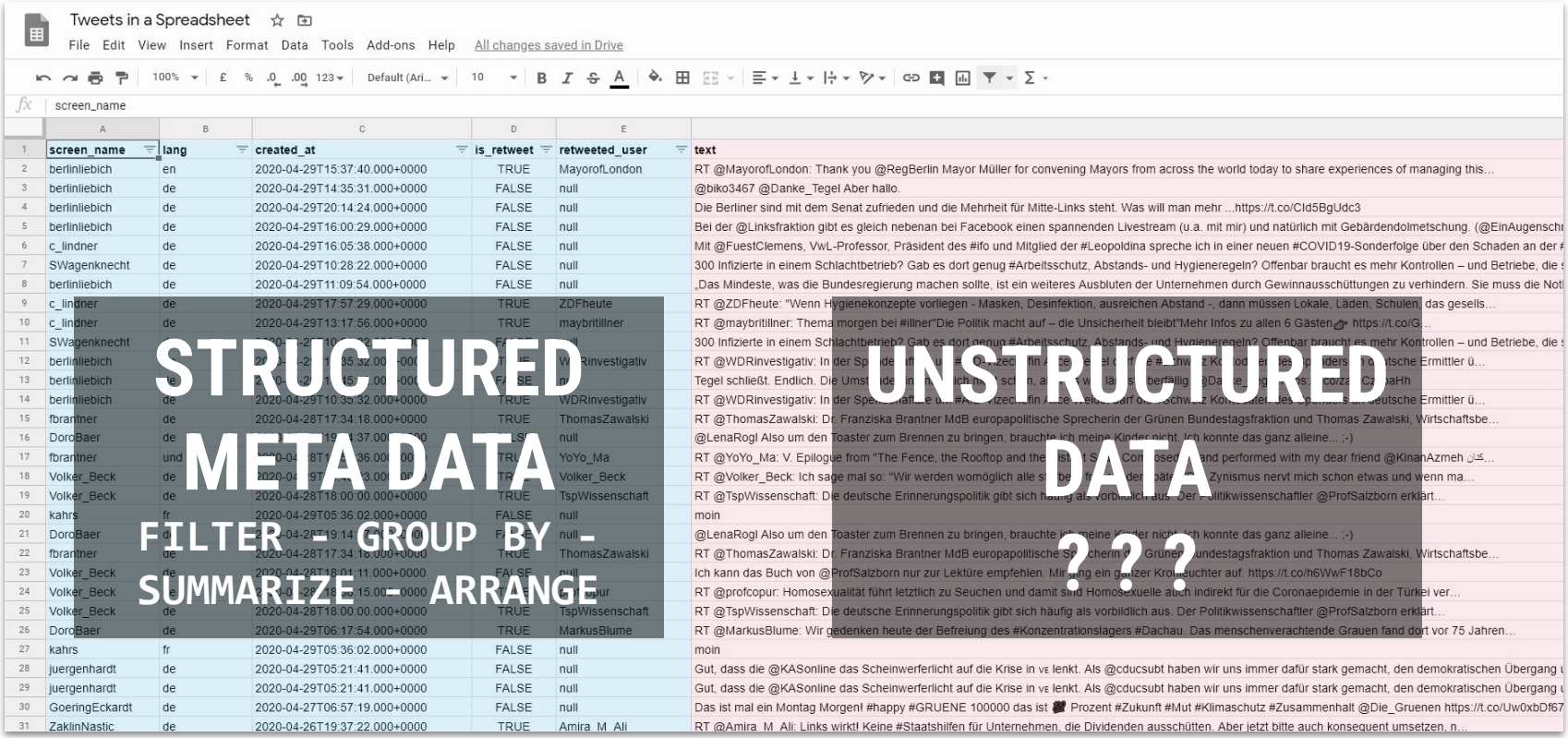


Source: Wickham, Hadley, and Garrett Grolemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. First edition, O'Reilly, 2016. URL: https://r4ds.hadley.nz/diagrams/data-science/base.png

WHY IS TEXT DIFFERENT?

WHY IS TEXT DIFFERENT?

STRUCTURED VS UNSTRUCTURED



	A	B	C	D	E	
1	screen_name	lang	created_at	is_retweet	retweeted_user	text
2	berlinliebich	en	2020-04-29T15:37:40.000+0000	TRUE	MayorofLondon	RT @MayorofLondon: Thank you @RegBerlin Mayor Müller for convening Mayors from across the world today to share experiences of managing this…
3	berlinliebich	de	2020-04-29T14:35:31.000+0000	FALSE	null	@biko3467 @Danke_Tegel Aber hallo.
4	berlinliebich	de	2020-04-29T20:14:24.000+0000	FALSE	null	Die Berliner sind mit dem Senat zufrieden und die Mehrheit für Mitte-Links steht. Was will man mehr ...https://t.co/CId5BgUdc3
5	berlinliebich	de	2020-04-29T16:00:29.000+0000	FALSE	null	Bei der @Linksfraktion gibt es gleich nebenan bei Facebook einen spannenden Livestream (u.a. mit mir) und natürlich mit Gebärdendolmetschung. (@EinAugenschr
6	c_lindner	de	2020-04-29T16:05:38.000+0000	FALSE	null	Mit @FuestClemens, VwL-Professor, Präsident des #ifo und Mitglied der #Leopoldina spreche ich in einer neuen #COVID19-Sonderfolge über den Schaden an der #
7	SWagenknecht	de	2020-04-29T10:28:22.000+0000	FALSE	null	300 Infizierte in einem Schlachtbetrieb? Gab es dort genug #Arbeitsschutz, Abstands- und Hygieneregeln? Offenbar braucht es mehr Kontrollen – und Betriebe, die s
8	berlinliebich	de	2020-04-29T11:09:54.000+0000	FALSE	null	„Das Mindeste, was die Bundesregierung machen sollte, ist ein weiteres Ausbluten der Unternehmen durch Gewinnausschüttungen zu verhindern. Sie muss die Not
9	c_lindner	de	2020-04-29T17:57:29.000+0000	TRUE	ZDFheute	RT @ZDFheute: "Wenn Hygienekonzepte vorliegen - Masken, Desinfektion, ausreichen Abstand -, dann müssen Lokale, Läden, Schulen, das gesells…
10	c_lindner	de	2020-04-29T13:17:56.000+0000	TRUE	maybritillner	RT @maybritillner: Thema morgen bei #illner"Die Politik macht auf – die Unsicherheit bleibt"Mehr Infos zu allen 6 Gästen https://t.co/G…
11	SWagenknecht	de	2020-04-29T10:28:22.000+0000	FALSE	null	300 Infizierte in einem Schlachtbetrieb? Gab es dort genug #Arbeitsschutz, Abstands- und Hygieneregeln? Offenbar braucht es mehr Kontrollen – und Betriebe, die s
12	berlinliebich	de	2020-04-29T10:35:32.000+0000	TRUE	WDRinvestigativ	RT @WDRinvestigativ: In der Sp… …tsche Ermittler ü…
13	berlinliebich	de	2020-04-29T10:35:32.000+0000	FALSE	null	Tegel schließt. Endlich. Die Umst… …baHh
14	berlinliebich	de	2020-04-29T10:35:32.000+0000	TRUE	WDRinvestigativ	RT @WDRinvestigativ: In der Spe… …tsche Ermittler ü…
15	fbrantner	de	2020-04-28T17:34:18.000+0000	TRUE	ThomasZawalski	RT @ThomasZawalski: Dr. Franziska Brantner MdB europapolitische Sprecherin der Grünen Bundestagsfraktion und Thomas Zawalski, Wirtschaftsbe…
16	DoroBaer	de	2020-04-29T19:41:37.000+0000	FALSE	null	@LenaRogl Also um den Toaster zum Brennen zu bringen, brauchte ich meine Kinder nicht. Ich konnte das ganz alleine... ;-)
17	fbrantner	und	2020-04-28T17:34:36.000+0000	TRUE	YoYo_Ma	RT @YoYo_Ma: V. Epilogue from "The Fence, the Rooftop and the… …and performed with my dear friend @KinanAzmeh …
18	Volker_Beck	de	2020-04-29T… …3.000+0000	TRUE	Volker_Beck	RT @Volker_Beck: Ich sage mal so: "Wir werden womöglich alle … …Zynismus nervt mich schon etwas und wenn ma…
19	Volker_Beck	de	2020-04-28T18:00:00.000+0000	TRUE	TspWissenschaft	RT @TspWissenschaft: Die deutsche Erinnerungspolitik gibt sich … …Politikwissenschaftler @ProfSalzborn erklärt…
20	kahrs	fr	2020-04-29T05:36:02.000+0000	FALSE	null	moin
21	DoroBaer	de	2020-04-29T19:14:… …37.000+0000	FALSE	null	@LenaRogl Also um den Toaster zum Brennen zu bringen, brauchte ich meine Kinder nicht. Ich konnte das ganz alleine... ;-)
22	fbrantner	de	2020-04-28T17:34:18.000+0000	TRUE	ThomasZawalski	RT @ThomasZawalski: Dr. Franziska Brantner MdB europapolitische Sprecherin … Grünen Bundestagsfraktion und Thomas Zawalski, Wirtschaftsbe…
23	Volker_Beck	de	2020-04-28T18:01:11.000+0000	FALSE	null	Ich kann das Buch von @ProfSalzborn nur zur Lektüre empfehlen. Mit … ein ganzer Kro…buchter auf. https://t.co/h6WwF18bCo
24	Volker_Beck	de	2020-04-28T18:… …15.000+0000	TRUE	…pur	RT @profcopur: Homosexualität führt letztlich zu Seuchen und damit sind Homosexuelle auch indirekt für die Coronaepidemie in der Türkei ver…
25	Volker_Beck	de	2020-04-28T18:00:00.000+0000	TRUE	TspWissenschaft	RT @TspWissenschaft: Die deutsche Erinnerungspolitik gibt sich häufig als vorbildlich aus. Der Politikwissenschaftler @ProfSalzborn erklärt…
26	DoroBaer	de	2020-04-29T06:17:54.000+0000	TRUE	MarkusBlume	RT @MarkusBlume: Wir gedenken heute der Befreiung des #Konzentrationslagers #Dachau. Das menschenverachtende Grauen fand dort vor 75 Jahren…
27	kahrs	fr	2020-04-29T05:36:02.000+0000	FALSE	null	moin
28	juergenhardt	de	2020-04-29T05:21:41.000+0000	FALSE	null	Gut, dass die @KASonline das Scheinwerferlicht auf die Krise in … lenkt. Als @cducsubt haben wir uns immer dafür stark gemacht, den demokratischen Übergang …
29	juergenhardt	de	2020-04-29T05:21:41.000+0000	FALSE	null	Gut, dass die @KASonline das Scheinwerferlicht auf die Krise in … lenkt. Als @cducsubt haben wir uns immer dafür stark gemacht, den demokratischen Übergang …
30	GoeringEckardt	de	2020-04-27T06:57:19.000+0000	FALSE	null	Das ist mal ein Montag Morgen! #happy #GRUENE 100000 das ist … Prozent #Zukunft #Mut #Klimaschutz #Zusammenhalt @Die_Gruenen https://t.co/Uw0xbDf67
31	ZaklinNastic	de	2020-04-26T19:37:22.000+0000	TRUE	Amira_M_Ali	RT @Amira_M_Ali: Links wirkt! Keine #Staatshilfen für Unternehmen, die Dividenden ausschütten. Aber jetzt bitte auch konsequent umsetzen, n…

What can we do with text without changing the structure?

Apply `filter`, but with different operators from `{stringr}`:

- Search for keywords with `str_detect`
- Search for patterns with regular expressions

Apply `mutate` and extract matches:

- Extract whether a keyword has been matched
- Extract the matched keyword(s)
- Extract multiple matches and explode to rows

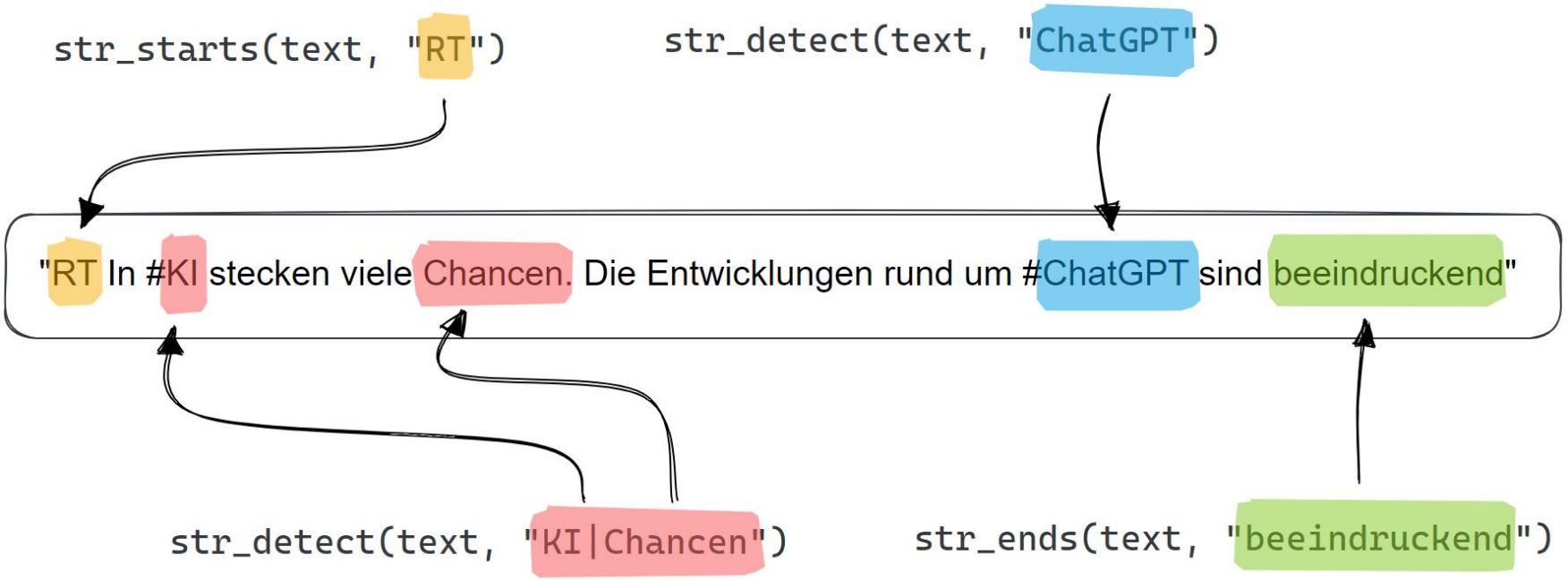
SEARCHES IN TEXT WITH {stringr}

- The `{stringr}` package contains functions for working with text

- We can use `{stringr}` to search in text:
  - `str_detect` for simple keyword matches
  - `str_starts` and `str_ends` as special cases
  - `str_to_lower` to improve matches

SEARCHES IN TEXT

SIMPLE KEYWORD MATCHES



Simple keyword matches anywhere in the text:

```
tweets |>
  filter(str_detect(text, "ChatGPT")) |>
  select(screen_name, text)
```

Searching at the start of a text:

```
tweets |>
  filter(str_starts(text, "RT", negate = TRUE)) |>
  select(screen_name, text)
```

Searching at the end of a text:

```
tweets |>
  filter(str_ends(text, "CL")) |>
  select(screen_name, text)
```

SEARCHES IN TEXT WITH
IMPROVED KEYWORD MATCHES

Improved keyword matches with `str_to_lower`:

```
tweets |>
  mutate(text = str_to_lower(text)) |>
  filter(str_detect(text, "chatgpt")) |>
  select(screen_name, text)
```

SEARCHES IN TEXT WITH
REGULAR EXPRESSIONS

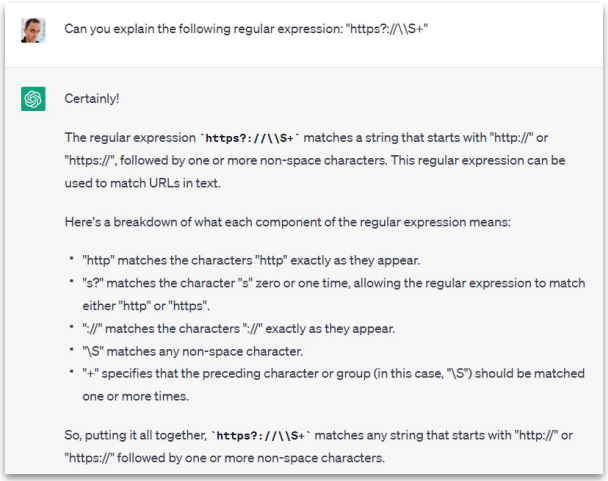
Regular expressions are powerful, yet can be complex to write.

Some examples:

- `str_detect(text, "chatgpt|gpt3|gpt4")`
- `str_detect(text, "@\\w+")`
- `str_detect(text, "https?://\\S+")`
- `str_detect(text, "\\s{2,}")`

All `{stringr}` functions actually take regular expressions!

ChatGPT to the rescue!



EXTRACT MATCHES WITH {stringr}

EXTRACT MATCHES

It is good to know whether and which keyword matched in a text:

- `str_extract` to pull a match from the text into a new column
- `str_extract_all` to pull all matches into a new column as a list
- `str_extract_all` with `str_c` and `unlist` to concatenate all matches into a single string
- `str_extract_all` with `unnest_longer` from `{tidyr}` to extract all matches into separate rows

Extract only the first match:

```
tweets |>
  mutate(text = str_to_lower(text)) |>
  mutate(first_match = str_extract(text, "chatgpt|gpt3|gpt4")) |>
  filter(!is.na(first_match)) |>
  select(first_match, text)
```

EXTRACT MATCHES

ALL MATCHES AS LIST

Extract all matches as a list:

```
tweets |>
  mutate(text = str_to_lower(text)) |>
  mutate(matches = str_extract_all(text, "chatgpt|gpt3|gpt4")) |>
  filter(!is.na(matches)) |>
  select(matches, text)
```

EXTRACT MATCHES

ALL MATCHES AS STRING

With `str_c` and `unlist`, we can transform a list of strings into a concatenated string:

```
tweets |>
  mutate(text = str_to_lower(text)) |>
  mutate(matches = str_extract_all(text, "chatgpt|gpt3|gpt4")) |>
  mutate(matches_flat = str_c(unlist(matches), collapse = ",")) |>
  select(matches_flat, text)
```

EXTRACT MATCHES

ALL MATCHES AS ROWS

With `unnest_longer`, we can explode a list of character strings into separate rows:

```
tweets |>
  mutate(extracted_urls = str_extract_all(text, "https?://\\S+")) |>
  unnest_longer(extracted_urls, keep_empty = TRUE) |>
  select(id, screen_name, extracted_urls, text) |>
  arrange(id)
```

REPLACE MATCHES WITH

{stringr}

REPLACE MATCHES

Occasionally, we want to remove things from text:

- `str_replace` to replace the first match with a new string
- `str_replace_all` to replace all matches with a new string
- `str_remove` and `str_remove_all` to delete occurrences
- `str_trim` to remove leading and trailing white spaces from text

All code examples are on GitHub: https://github.com/winf-hsos/data-analytics-code

SPLIT TEXT WITH {stringr}

SPLIT TEXT

ALL MATCHES AS ROWS

Splitting a string into multiple pieces can be helpful at times:

- `str_split` to break a long string into a list of strings using a separator character or pattern
- `str_split_i` to break a long string using a separator and keeping only the i-th result