Tidying and manipulating data using the tidyverse

Dr. Calum Webb
Sheffield Methods Institute, the University of Sheffield
c.j.webb@sheffield.ac.uk

This training course is designed to be hands-on. We'll be spending most of our time working through real applications of the data tidying tools that make up the tidyverse.

Data tidying requires the use of multiple tools

- The idea of today is to introduce you to all the tools that will allow you to tidy any untidy dataset.
- You won't be able to use all of them perfectly right from the start.
- But if you invest time into learning to use them, you can become very proficient in data tidying.

Why spend time learning how to tidy data?

- Tidying data and preparing it for analysis or visualisation is often the most time consuming part of any quantitative research project.
- Tidying data is not often reproducable *unless it has been tidied programatically*.

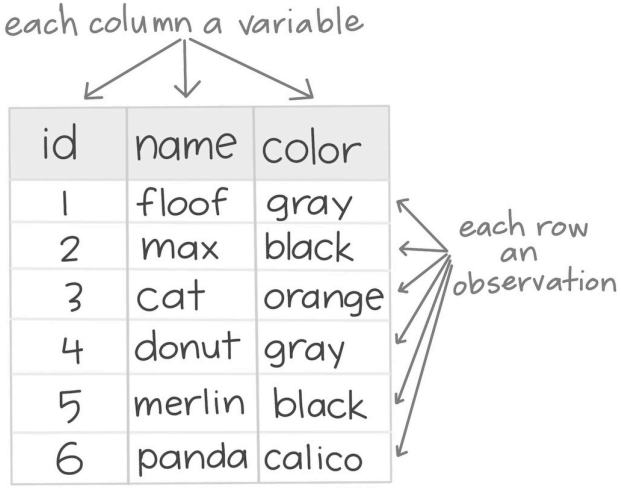
Introduction: What is tidy data?

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

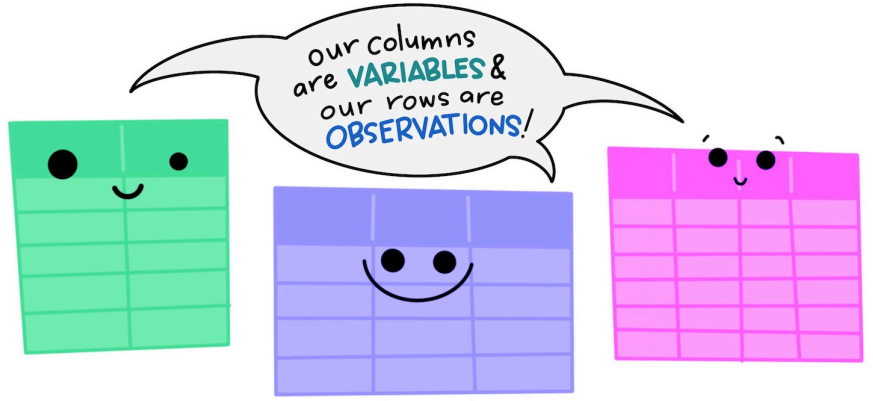
- each variable forms a column
- each observation forms a row
- each cell is a single measurement



Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

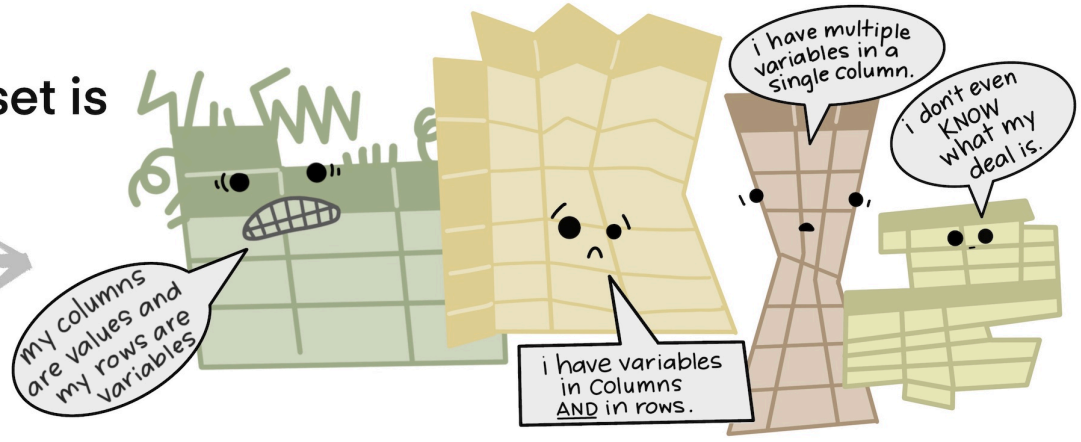
Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration by Julia Lowndes and Allison Horst

The standard structure of tidy data means that "tidy datasets are all alike..."



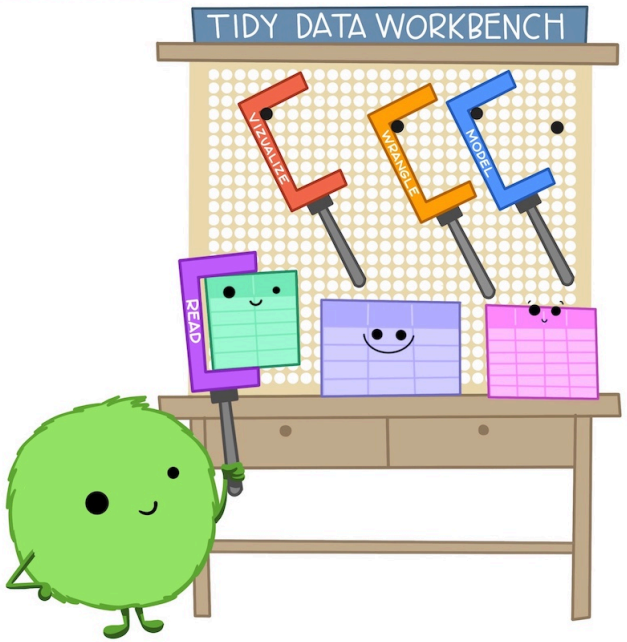
"...but every messy dataset is messy in its own way."

–HADLEY WICKHAM

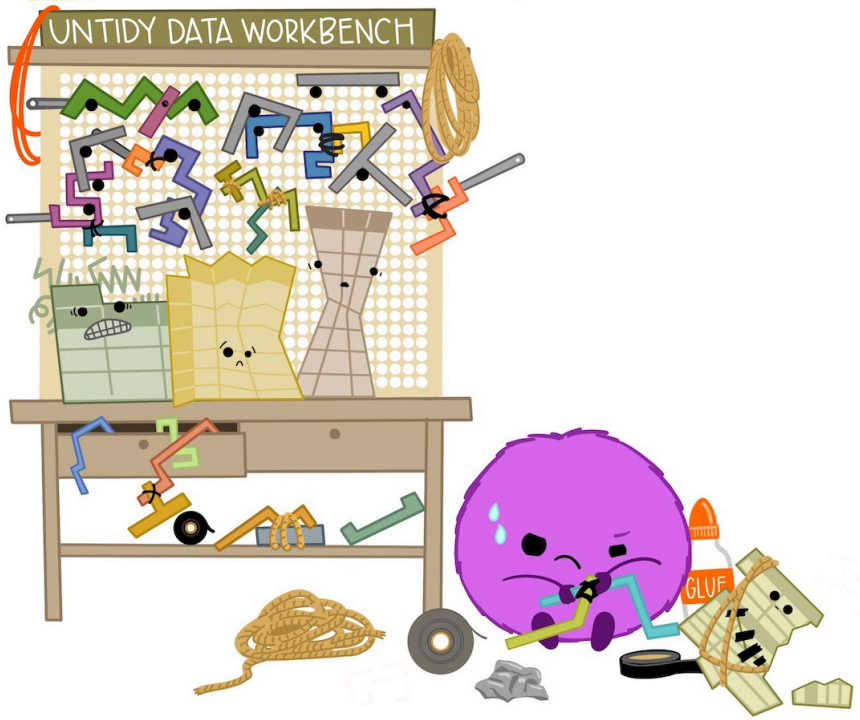


Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration by Julia Lowndes and Allison Horst

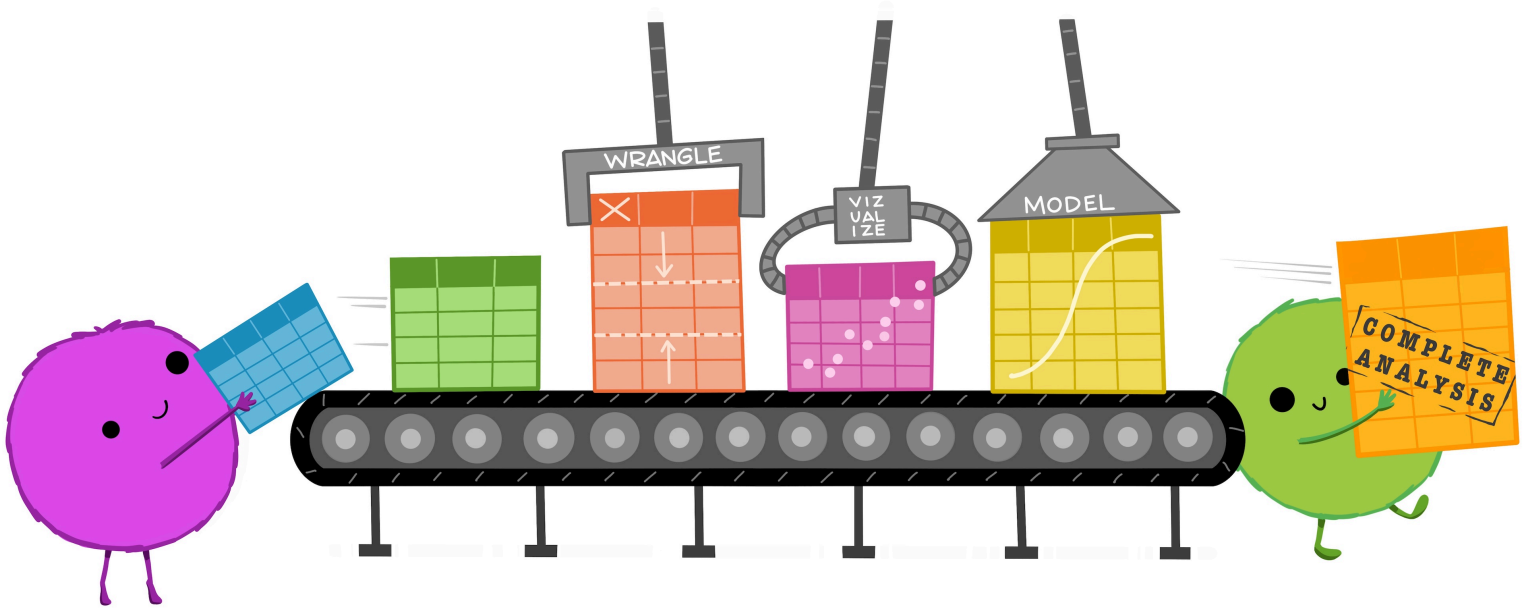
When working with tidy data, we can use the same tools in similar ways for different datasets...



...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.



Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration by Julia Lowndes and Allison Horst



Illustrations from the Openscapes blog Tidy Data for reproducibility, efficiency, and collaboration by Julia Lowndes and Allison Horst

1. Reading data from different data sources and select/rename columns.

- Reading data from Stata, SPSS, and Excel files
- Filtering variables using select()
- Generalised select()

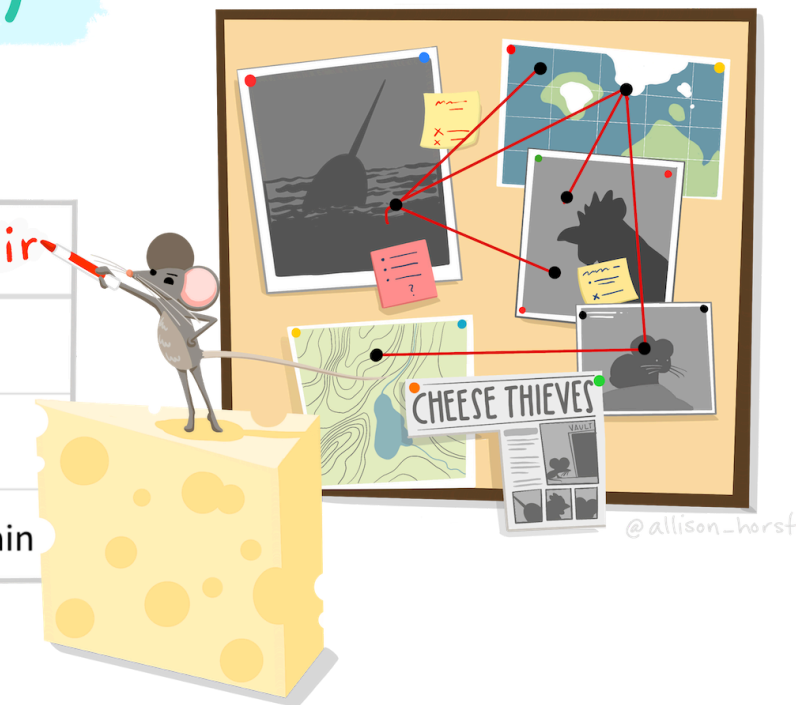
dplyr::rename()

RENAME COLUMNS*

```
df %>% rename(lair = site)
```

~~species~~ nemesis	status	~~site~~ lair
narwhal	unknown	ocean
chicken	active	coop
pika	active	mountain

*See **rename_with()** to rename using a function.



Artwork by @allison_horst

Artwork by @allison_horst

2. Creating new variables using mutate()

- Creating new variables that are transformations of existing variables
- Recoding categorical variables using case_when()
- Extracting numbers with parse_number()
- Performing repeated/generalised transformations

Artwork by @allison_horst

dplyr::case_when()

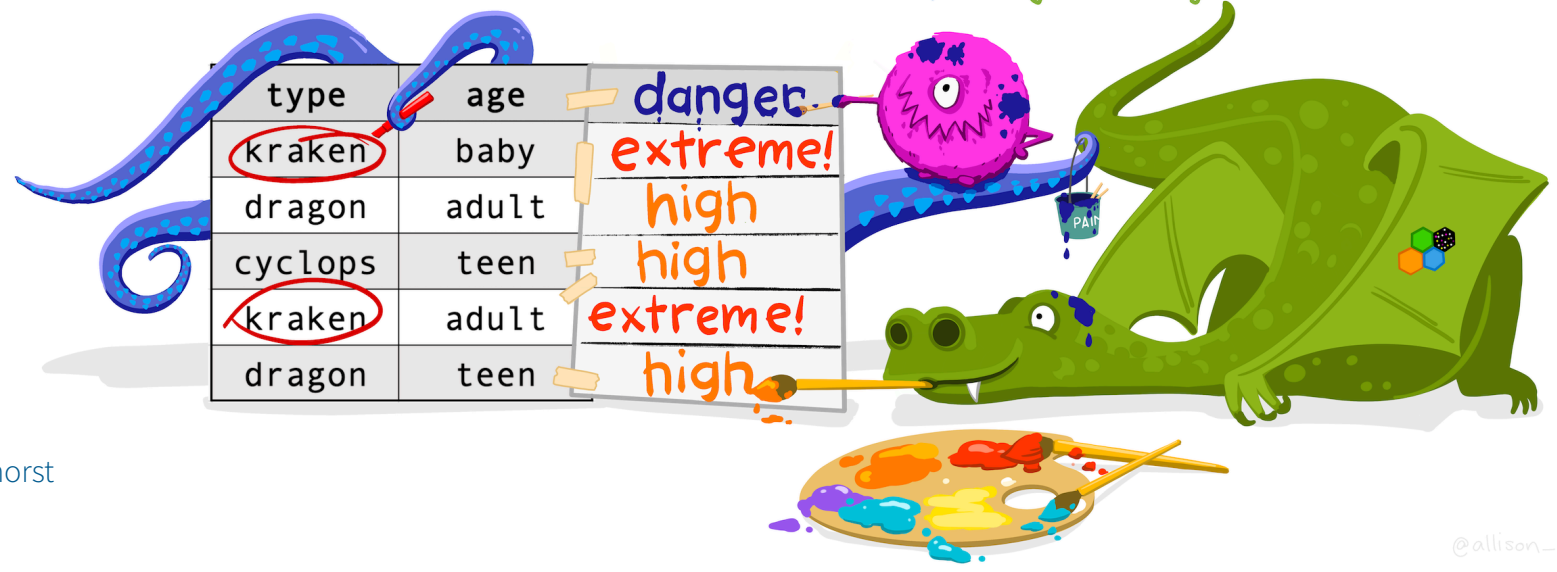
IF ELSE... (but you love it?)

```
df %>% mutate(danger = case_when(type == "kraken" ~ "extreme!",
                                 TRUE ~ "high"))
```

ADD COLUMN 'danger'

IF type is kraken THEN danger is extreme!

OTHERWISE, danger is high.



Artwork by @allison_horst

Artwork by @allison_horst

3. Aggregating data to higher levels with group_by()

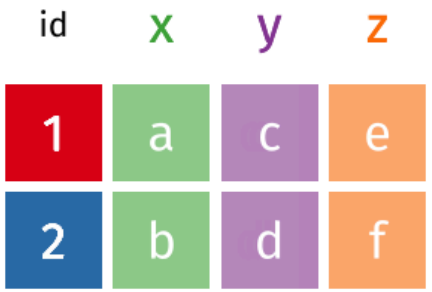
- Creating new, aggregated datasets using group_by() and summarise()
- Adding group-level variables for multilevel models using group_by() and mutate()

Animation by Andrew Heiss

4. Pivoting data between wide and long formats

- Converting wide datasets suitable for Latent Growth Structural Equation Modelling to long datasets suitable for multilevel modelling.
- ... and the reverse.

wide



Animation by Garrick Aden-Buie

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

Animation by Garrick Aden-Buie

5. Working with strings and a little bit of regex

- How to remove certain characters or strings from character type variables (especially footnotes).
- Extracting subsets of characters from longer strings.
- Splitting variables into multiple columns based on a character within a string.

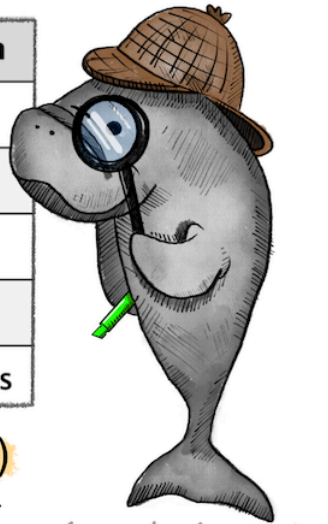
Artwork by @allison_horst

stringr::str_detect()

DETECT PRESENCE/ABSENCE* OF A PATTERN IN A STRING!

(df)

name	species	description
doug	dugong	awesome
olive	otter	super cool
greta	gorilla	superb
wilma	wolf	big, bad
barry	bonobo	superstitious



(a string detecting dugong)

Example:

```
str_detect(df$description, pattern = "super")
```

*add negate=TRUE to detect the absence of a pattern

Outcome:

```
FALSE  TRUE  TRUE  FALSE  TRUE
```

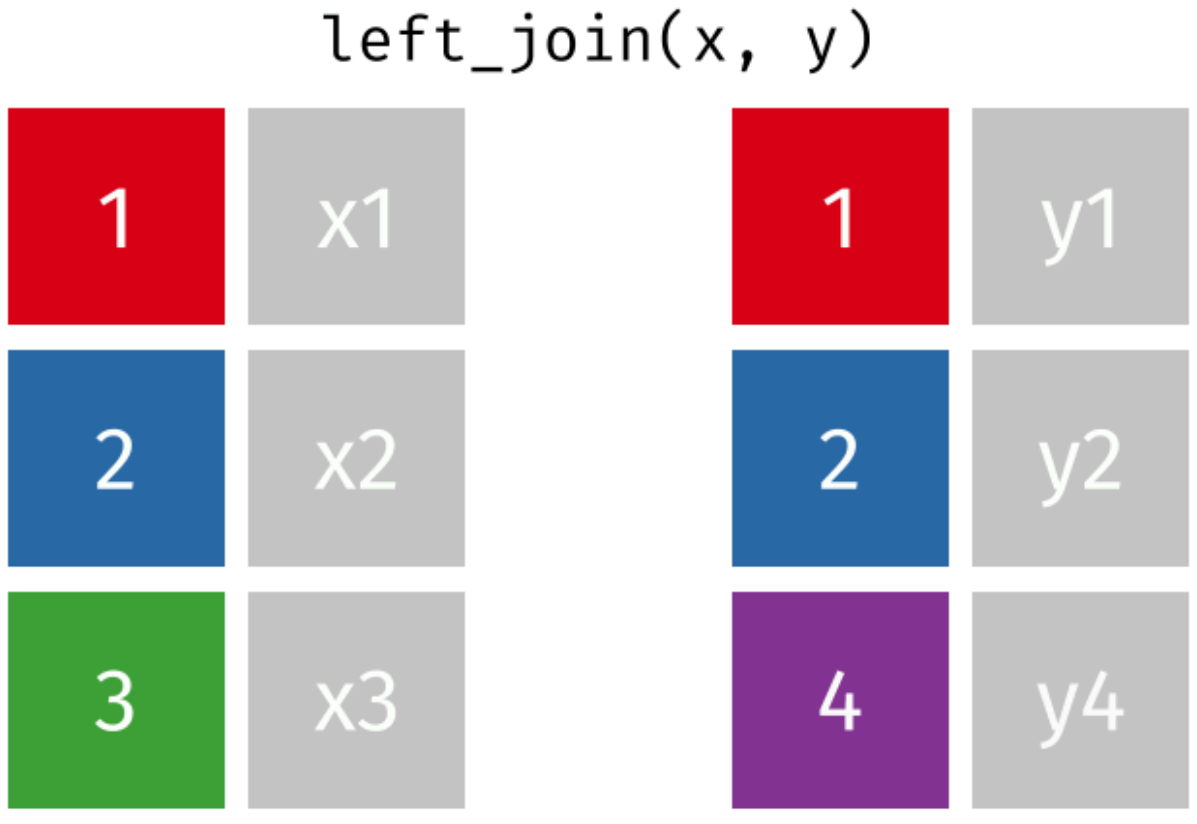
@allison_horst

Artwork by @allison_horst

6. Joining relational datasets

- Joining datasets based on a shared key
- Joining datasets together based on a combination of variables that form a key
- Joining higher level data to lower level data
- Checking for missing observations with anti_join()

Key: A value, usually a string, that uniquely identifies each observation across multiple related (relational) datasets



Animation by Garrick Aden-Buie

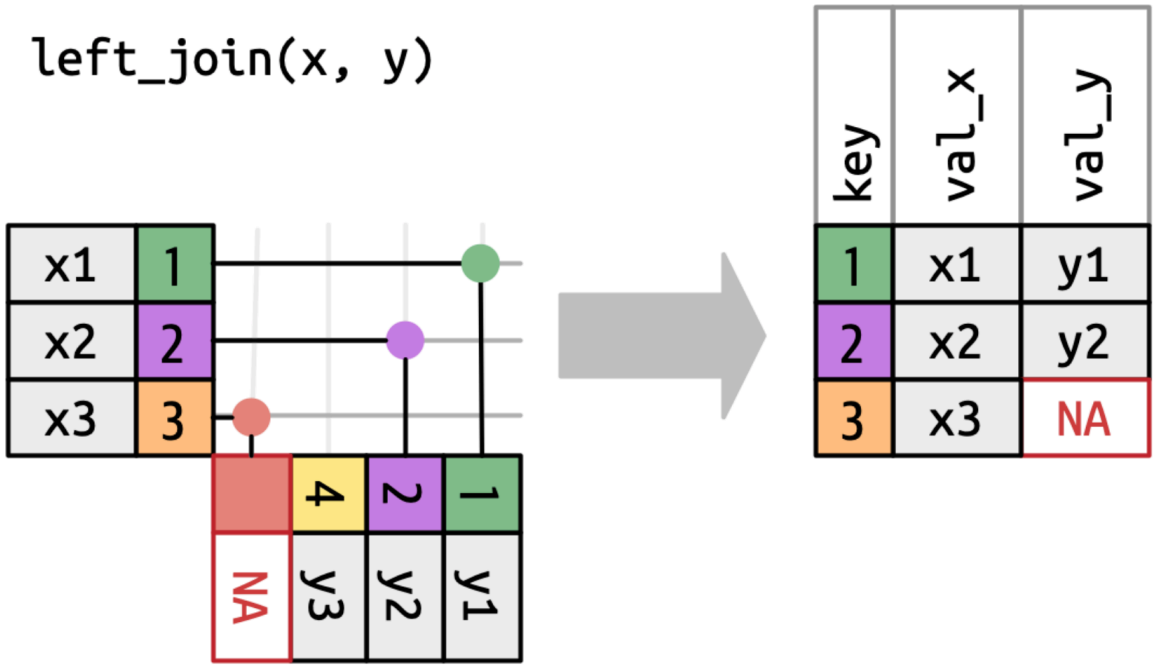


Diagram from R for Data Science 2e

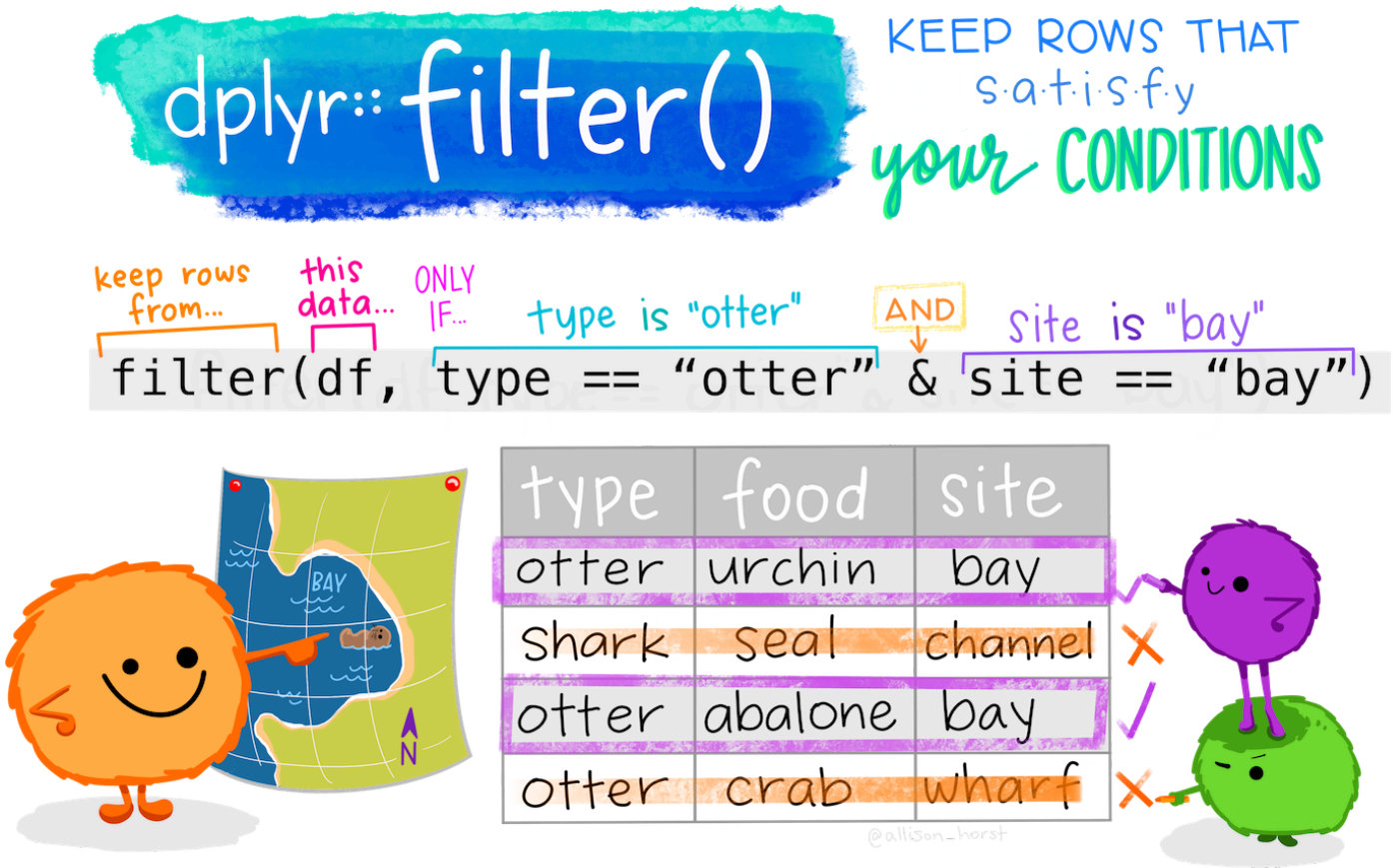
7. Working with dates

- How to fix how R interprets dates when they aren't in YYYY-MM-DD format.

Artwork by @allison_horst

8. Filtering rows of observations

- How to filter data based on values in character/factor type variables
- How to filter data based on numeric type variables
- How to filter data based on dates



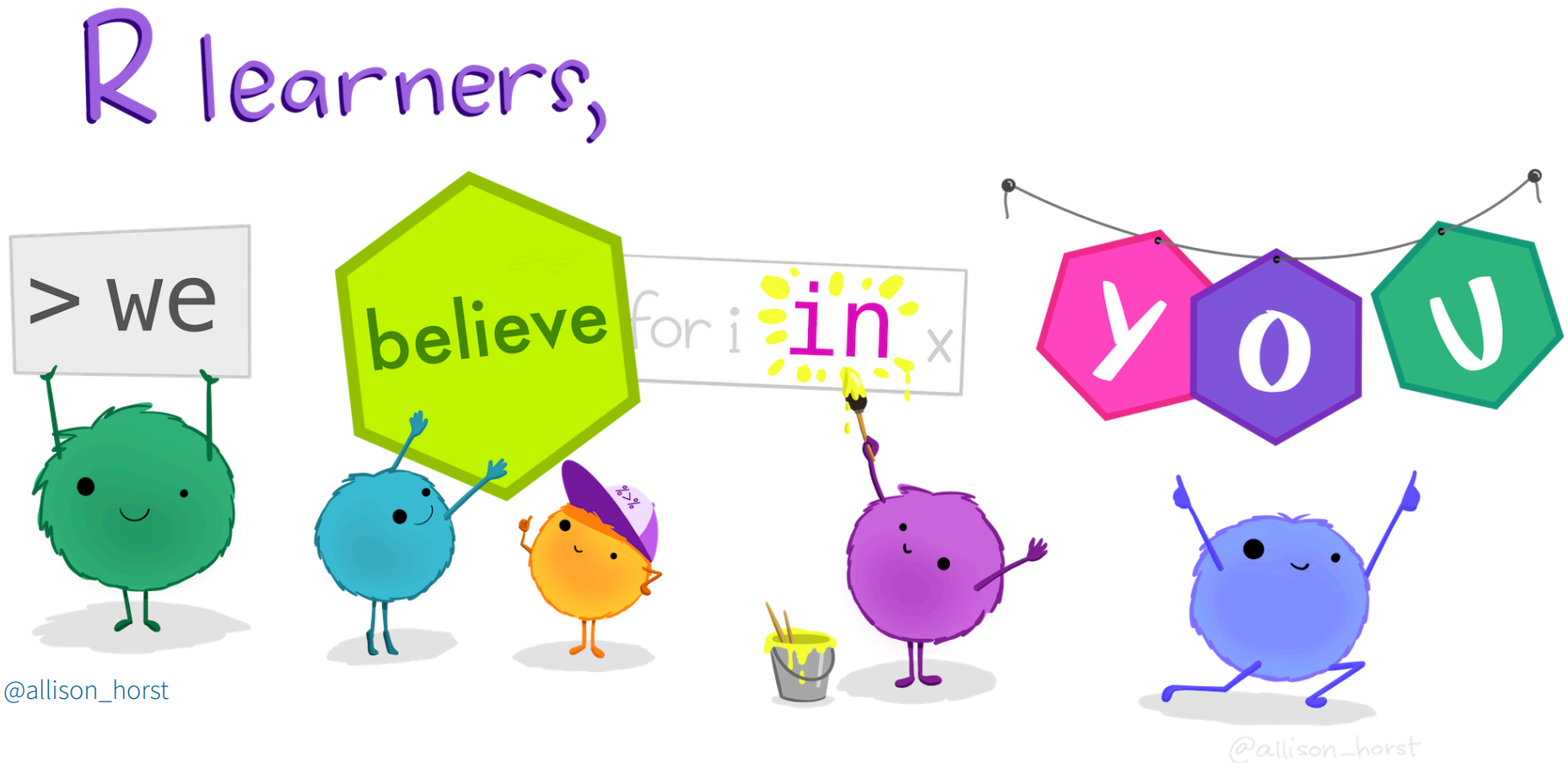
Artwork by @allison_horst

You now have all the tools — all that's left is the practice.

- In this course I've tried to give you, in as short a time as possible, information and practical examples of how to use all of the tools I've picked up over more than 10 years of using R.

You now have all the tools — all that's left is the practice.

- If you keep practicing tidying untidy datasets, using these tools will eventually become effortless. Untidy data becomes a puzzle to solve. But when you're just starting out, the puzzles will be frustrating.



Artwork by @allison_horst