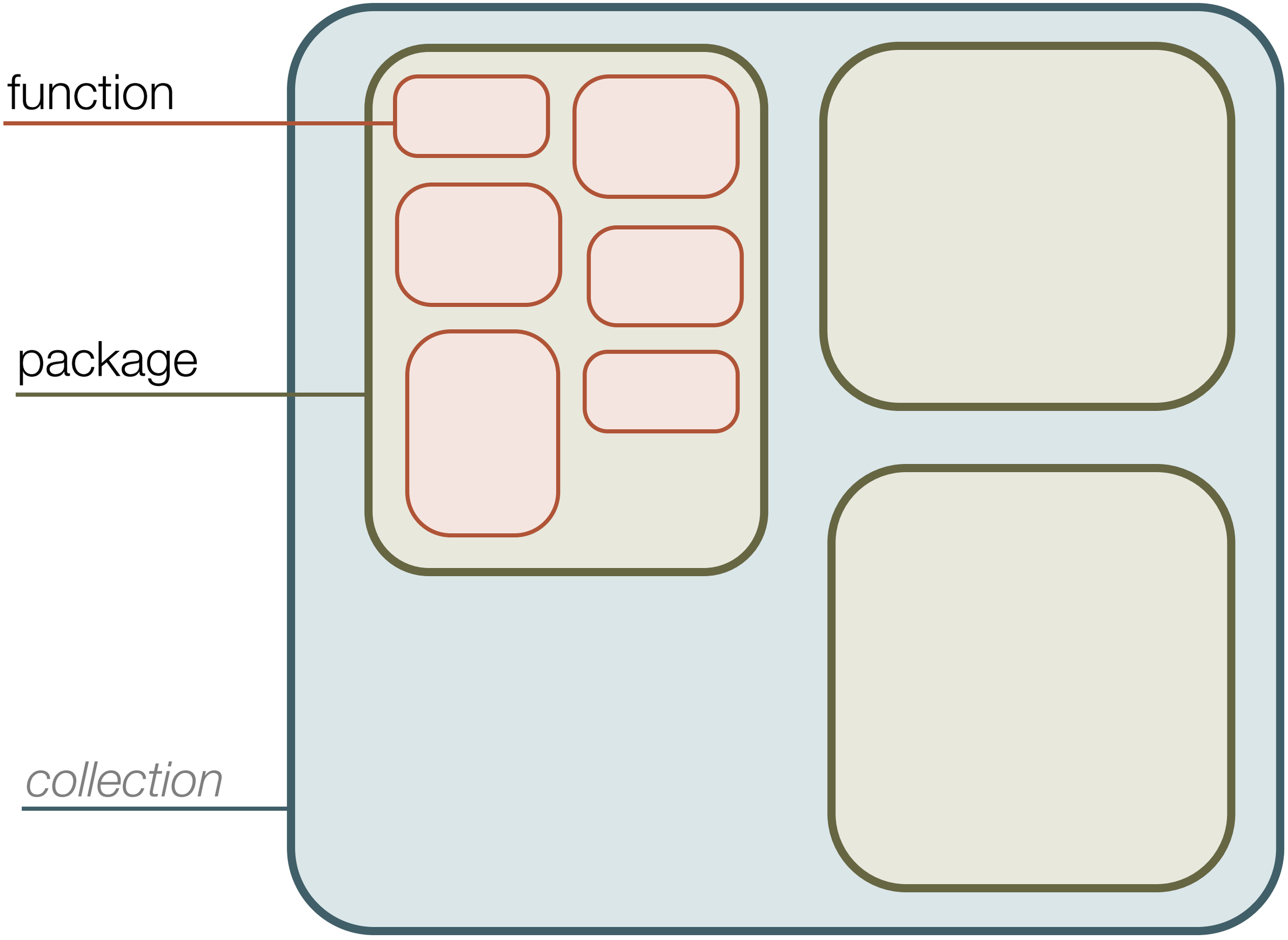
Part II: tidyverse

R for data science

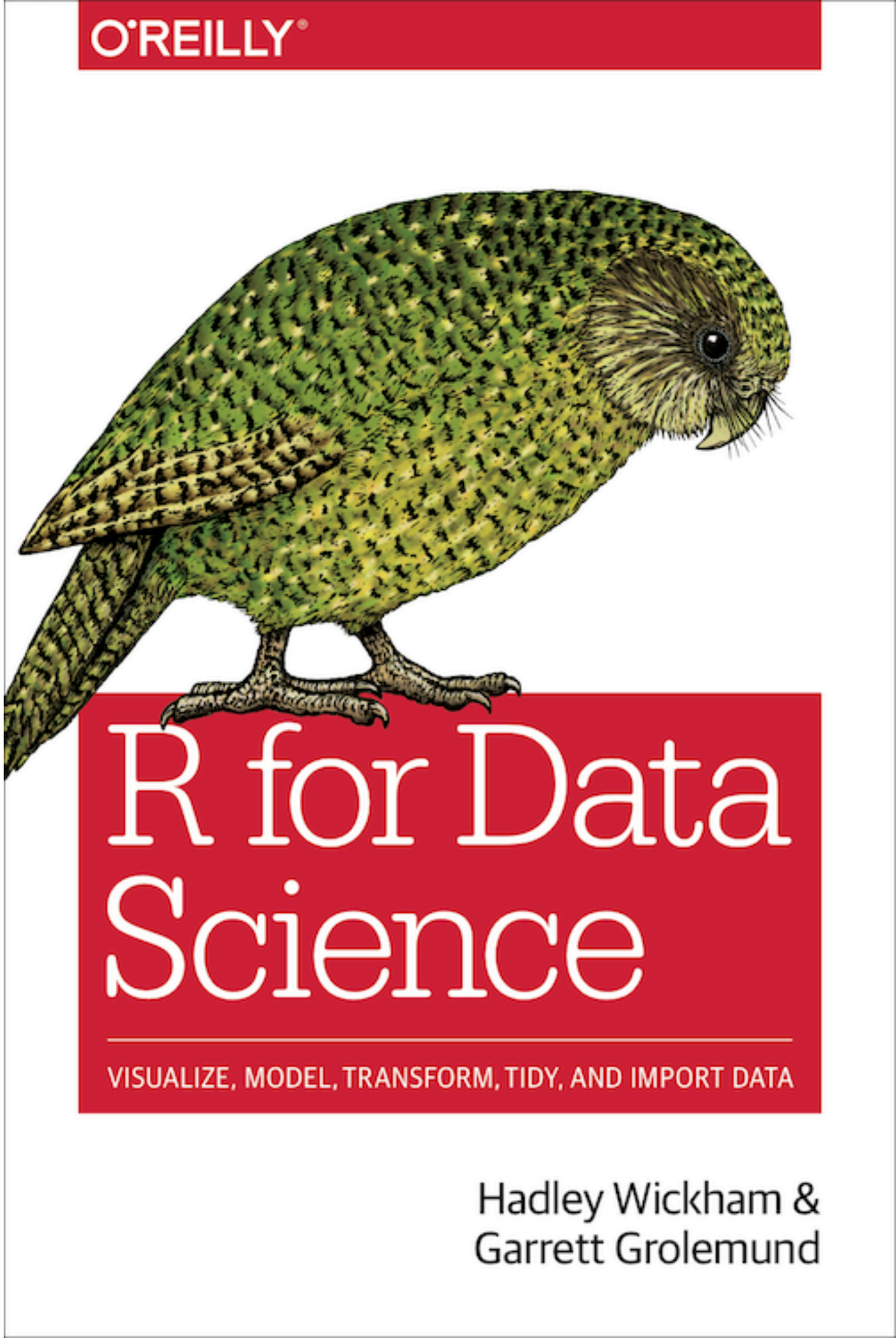
"The tidyverse is an opinionated collection of R packages **designed for data science**. All packages **share an underlying design** philosophy, grammar, and data structures."

– tidyverse.org (2018)

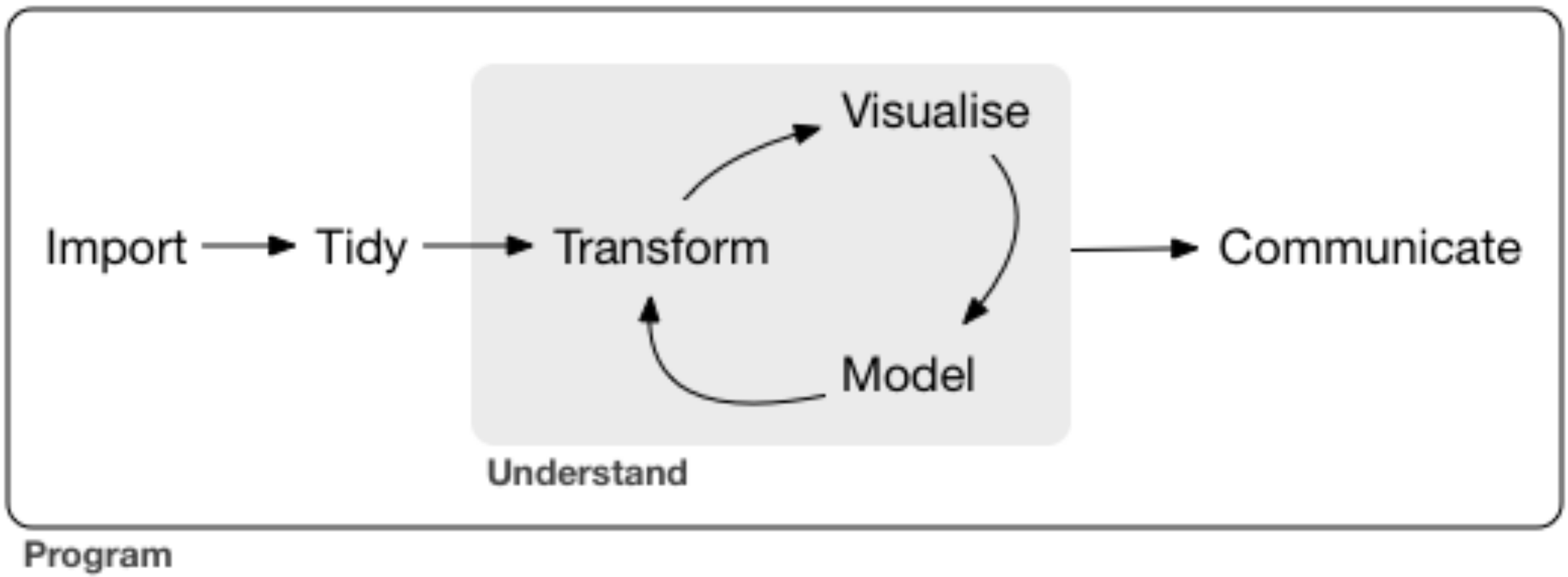

function
package
collection

Learn tidyverse

- R for Data Science (book); freely available on [r4ds.had.co.nz/](r4ds.had.co.nz/)
- cheatsheets www.rstudio.com/resources/cheatsheets/



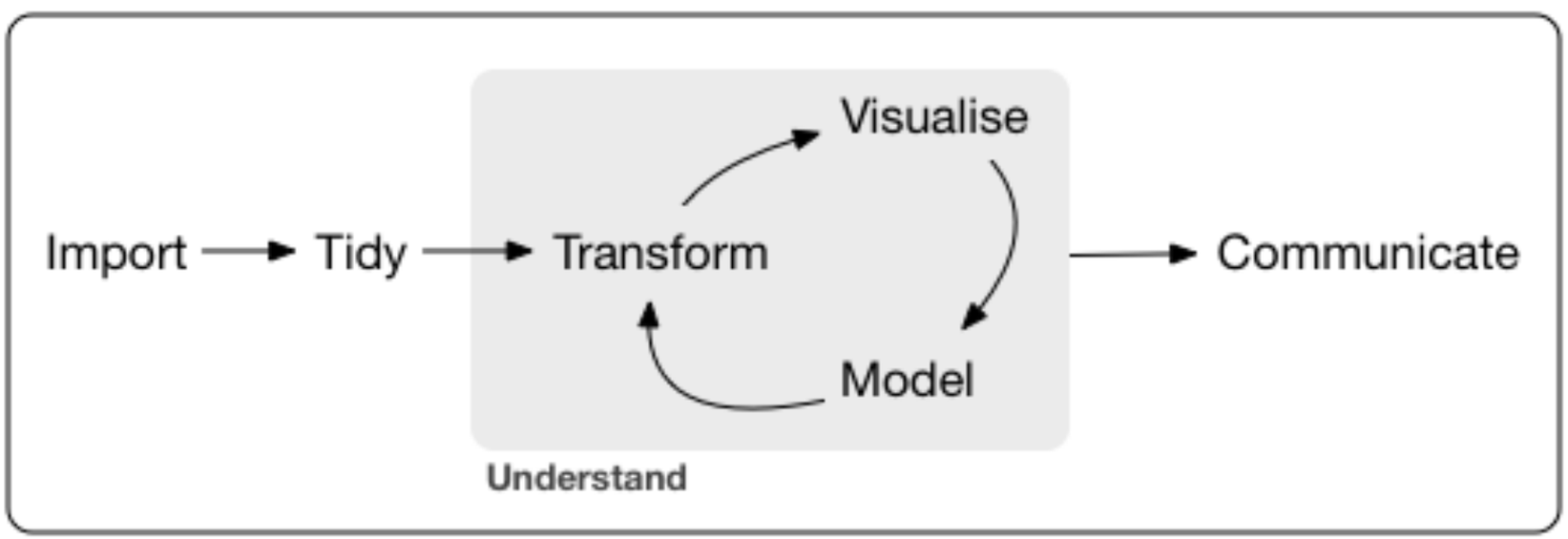
Data science workflow

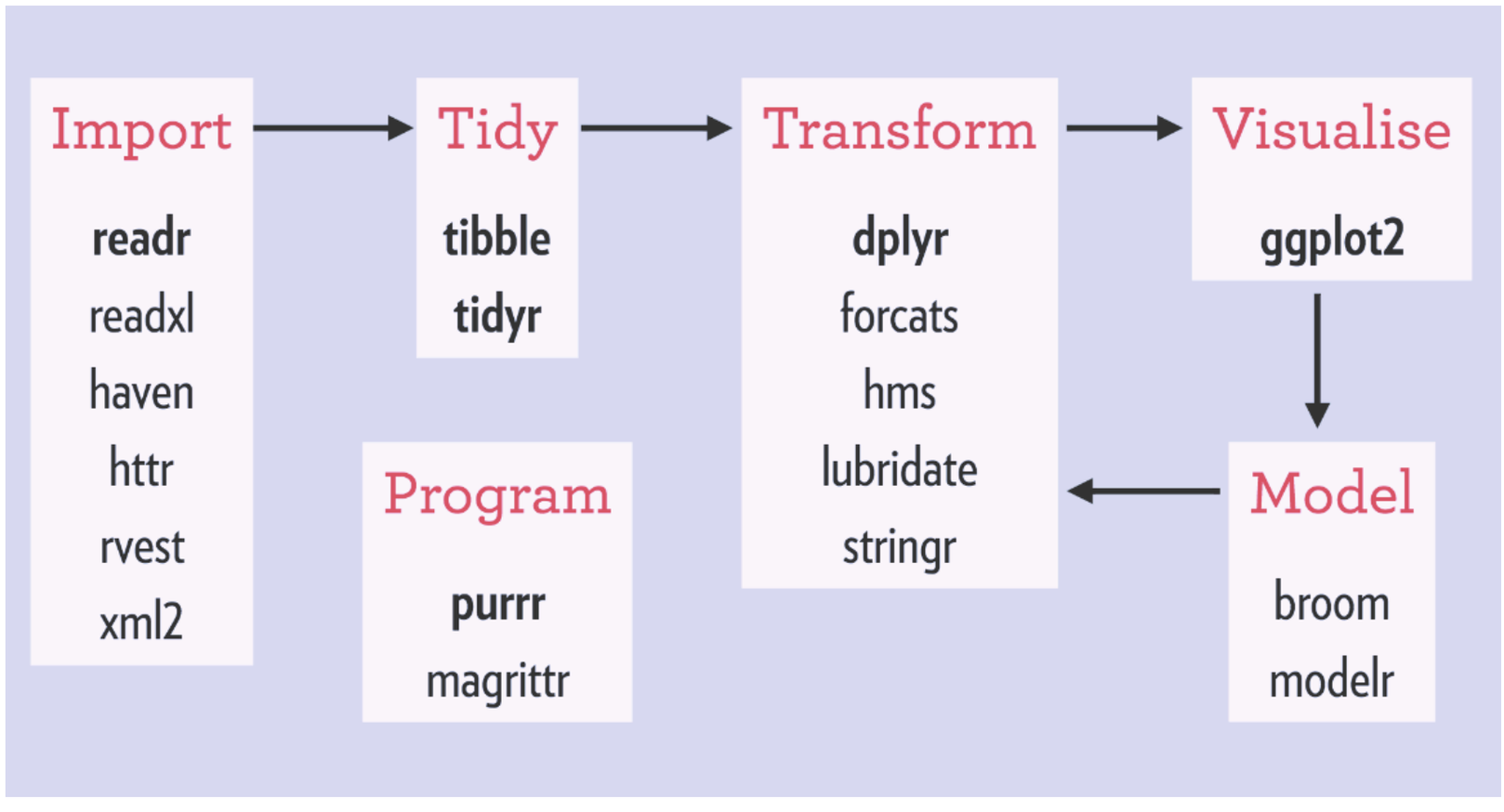


Tidy data ensures that further processing can be done efficiently, and reproducibly.

Tidy data is easy to manipulate, model, and visualize.

Tidyverse: connecting packages for every step



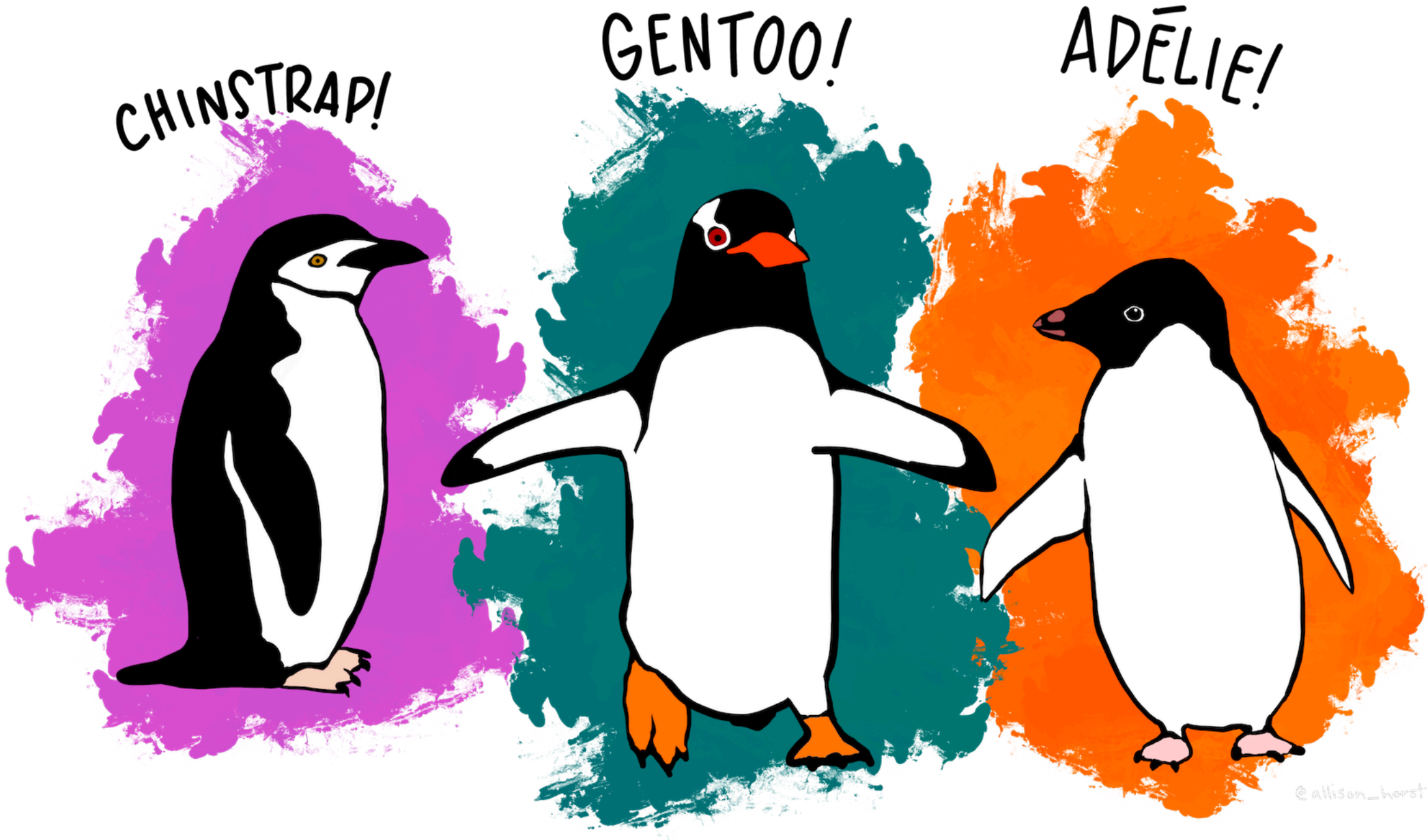


Tidy data

- Each **variable** is a column and contains **values**
- Each **observation** is a row
- Each type of **observational unit** forms a table

Our data: Palmer penguins

- Thanks to Dr. Kristen Gorman and the Palmer Station, Antarctica LTER
- Brought to R by Allison Horst



Plant data

Plant_no	Treatment	Stem length day 1	Leaf width day 1	Stem length day 2	Leaf width day 2	Stem length day 3	Leaf width day 3
A1_14	control	120	21	122	23	124	25
A1_18	control	132	23	135	25	138	27
A1_21	control	131	18	133	20	135	21
A2_09	UV	109	29	114	31	115	31
A3_02	UV	125	25	127	27	129	28
A3_10	UV	130	12	133	14	136	16

Plant data

values in column names

Plant_no	Treatment	Stem length day 1	Leaf width day 1	Stem length day 2	Leaf width day 2	Stem length day 3	Leaf width day 3
A1_14	control	120	21	122	23	124	25
A1_18	control	132	23	135	25	138	27
A1_21	control	131	18	133	20	135	21
A2_09	UV	109	29	114	31	115	31
A3_02	UV	125	25	127	27	129	28
A3_10	UV	130	12	133	14	136	16

Plant data

multiple observations per row

Plant_no	Treatment	Stem length day 1	Leaf width day 1	Stem length day 2	Leaf width day 2	Stem length day 3	Leaf width day 3
A1_14	control	120	21	122	23	124	25
A1_18	control	132	23	135	25	138	27
A1_21	control	131	18	133	20	135	21
A2_09	UV	109	29	114	31	115	31
A3_02	UV	125	25	127	27	129	28
A3_10	UV	130	12	133	14	136	16

Tidy plant data

Plant_no	Treatment	Element	Day	Measurement
A1_14	control	Stem length	1	120
A1_14	control	Leaf width	1	21
A1_14	control	Stem length	2	122
A1_14	control	Leaf width	2	23
A1_14	control	Stem length	3	124
A1_14	control	Leaf width	3	25
A1_18	control	Stem length	1	132
A1_18	control	Leaf width	1	23
A1_18	control	Stem length	2	135
A1_18	control	Leaf width	2	25
A1_18	control	Stem length	3	138
A1_18	control	Leaf width	3	27
A1_21	control	Stem length	1	131

+ 23 more rows

Wide vs. Long

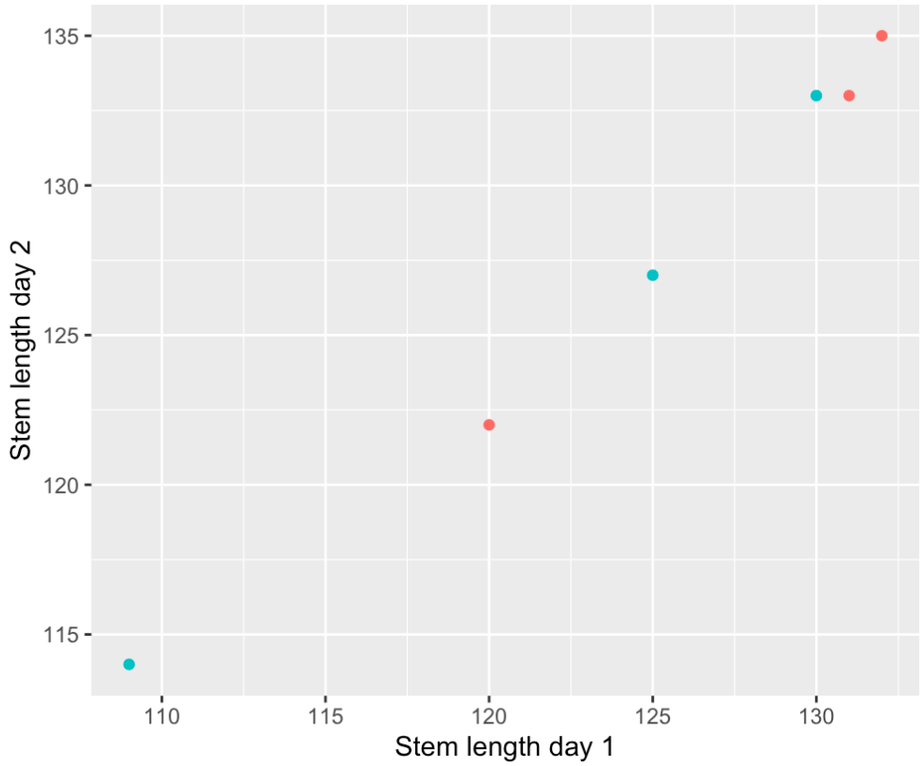
Plant_no	Treatment	Stem length day 1	Leaf width day 1	Stem length day 2	Leaf width day 2	Stem length day 3	Leaf width day 3
A1_14	control	120	21	122	23	124	25
A1_18	control	132	23	135	25	138	27
A1_21	control	131	18	133	20	135	21
A2_09	UV	109	29	114	31	115	31
A3_02	UV	125	25	127	27	129	28
A3_10	UV	130	12	133	14	136	16

- More information per row
- Combines all measurements on a single individual
- Necessary to plot matching measurements

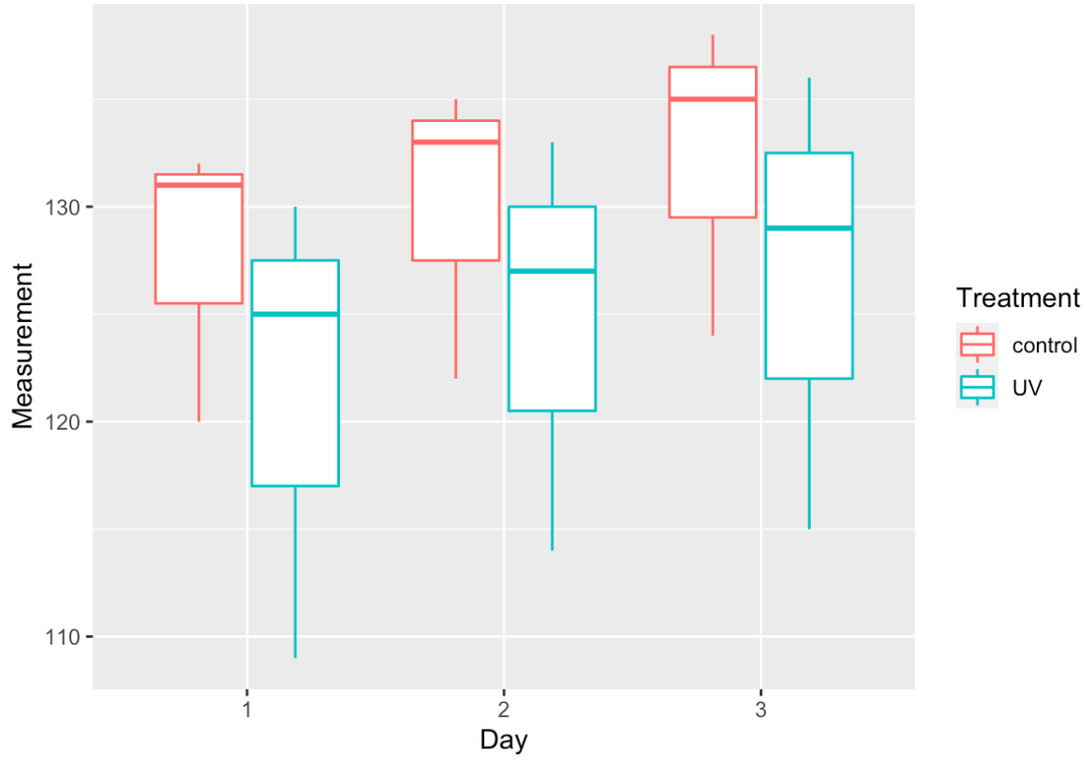
Plant_no	Treatment	Element	Day	Measurement
A1_14	control	Stem length	1	120
A1_14	control	Leaf width	1	21
A1_14	control	Stem length	2	122
A1_14	control	Leaf width	2	23
A1_14	control	Stem length	3	124
A1_14	control	Leaf width	3	25
A1_18	control	Stem length	1	132
A1_18	control	Leaf width	1	23
A1_18	control	Stem length	2	135
A1_18	control	Leaf width	2	25
A1_18	control	Stem length	3	138
A1_18	control	Leaf width	3	27
A1_21	control	Stem length	1	131

- More information per column
- No values as column headers (tidy)
- Single observation in single row (tidy)
- Necessary to plot large amounts of data in a single plot

Wide vs. Long

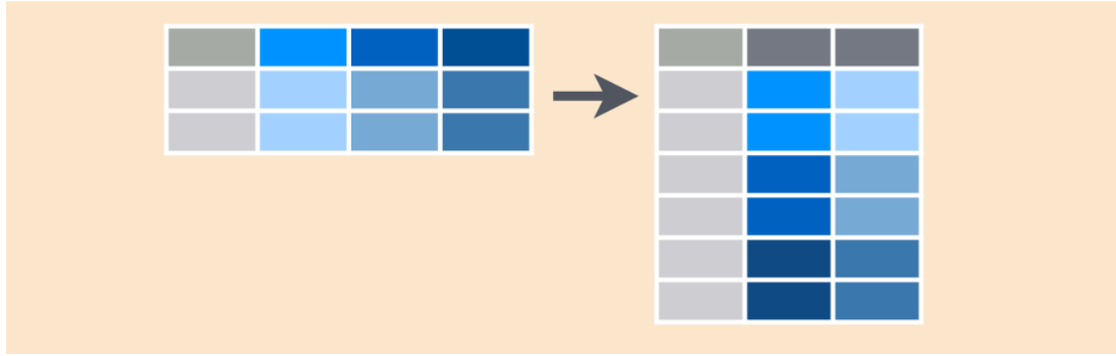


- More information per row
- Combines all measurements on a single individual
- **Necessary to plot matching measurements**



- More information per column
- No values as column headers (tidy)
- Single observation in single row (tidy)
- **Necessary to plot large amounts of data in a single plot**

Function: pivot_longer()



```
> plants
# A tibble: 6 x 8
  Plant_number Treatment `Stem length da… `Leaf width day… `Stem length da… `Leaf width day…
  <chr>        <chr>                <dbl>            <dbl>            <dbl>            <dbl>
1 A1_14        control                120               21              122               23
2 A1_18        control                132               23              135               25
3 A1_21        control                131               18              133               20
4 A2_09        UV                     109               29              114               31
5 A3_02        UV                     125               25              127               27
6 A3_10        UV                     130               12              133               14
# … with 2 more variables: `Stem length day 3` <dbl>, `Leaf width day 3` <dbl>
```

headers

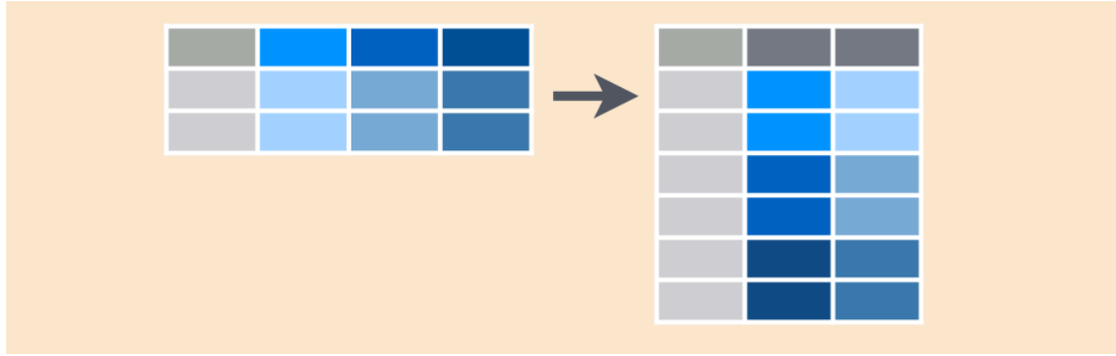
content

```
> tidy_plants <- plants %>%
+   pivot_longer(cols = c(`Leaf width day 1`, `Leaf width day 2`, `Leaf width day 3`,
 `Stem length day 1`, `Stem length day 2`, `Stem length day 3`),
+                names_to = "Element_day",
+                values_to = "Measurement")
```

column name for headers

column name for content

Function: pivot_longer()



```
> plants
# A tibble: 6 x 8
  Plant_number Treatment `Stem length da… `Leaf width day… `Stem length da… `Leaf width day…
  <chr>        <chr>                <dbl>            <dbl>            <dbl>            <dbl>
1 A1_14        control                120               21              122               23
2 A1_18        control                132               23              135               25
3 A1_21        control                131               18              133               20
4 A2_09        UV                     109               29              114               31
5 A3_02        UV                     125               25              127               27
6 A3_10        UV                     130               12              133               14
# … with 2 more variables: `Stem length day 3` <dbl>, `Leaf width day 3` <dbl>
```

headers

content

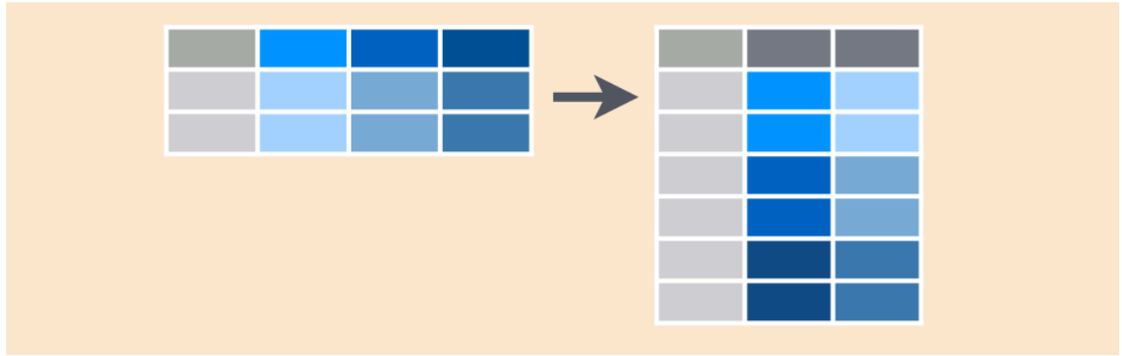
```
> tidy_plants <- plants %>%
+   pivot_longer(cols = where(is.numeric),
+                names_to = "Element_day",
+                values_to = "Measurement")
```

column name for **headers**

column name for **content**

https://tidyselect.r-lib.org/reference/language.html

Function: pivot_longer()



```
> tidy_plants
# A tibble: 36 x 4
   Plant_number Treatment Element_day        Measurement
   <chr>        <chr>     <chr>                    <dbl>
 1 A1_14        control   Stem length day 1          120
 2 A1_14        control   Leaf width day 1            21
 3 A1_14        control   Stem length day 2          122
 4 A1_14        control   Leaf width day 2            23
 5 A1_14        control   Stem length day 3          124
 6 A1_14        control   Leaf width day 3            25
 7 A1_18        control   Stem length day 1          132
 8 A1_18        control   Leaf width day 1            23
 9 A1_18        control   Stem length day 2          135
10 A1_18        control   Leaf width day 2            25
# … with 26 more rows
```

headers

content

Separate into columns

```
> tidy_plants %>%
+   separate(col = Element_day, sep = " day ", into = c("Element", "Day"))
# A tibble: 36 x 5
   Plant_number Treatment Element      Day   Measurement
   <chr>        <chr>     <chr>        <chr>       <dbl>
 1 A1_14        control   Stem length  1             120
 2 A1_14        control   Leaf width   1              21
 3 A1_14        control   Stem length  2             122
 4 A1_14        control   Leaf width   2              23
 5 A1_14        control   Stem length  3             124
 6 A1_14        control   Leaf width   3              25
 7 A1_18        control   Stem length  1             132
 8 A1_18        control   Leaf width   1              23
 9 A1_18        control   Stem length  2             135
10 A1_18        control   Leaf width   2              25
# … with 26 more rows
```

Function: pivot_wider()

columns to **headers**

column to be split as **content**

	Plant_number	Treatment	Element	Day	Measurement
	<chr>	<chr>	<chr>	<chr>	<dbl>
1	A1_14	control	Stem length	1	120
2	A1_14	control	Leaf width	1	21
3	A1_14	control	Stem length	2	122
4	A1_14	control	Leaf width	2	23
5	A1_14	control	Stem length	3	124
6	A1_14	control	Leaf width	3	25

Unique (combination of) variable(s): necessary for grouping!

```
> tidy_plants %>%
+   pivot_wider(names_from = c(Element, Day),
+               names_sep = " day ",
+               values_from = Measurement)
```

columns with **headers**

column with **content**

Function: pivot_wider()

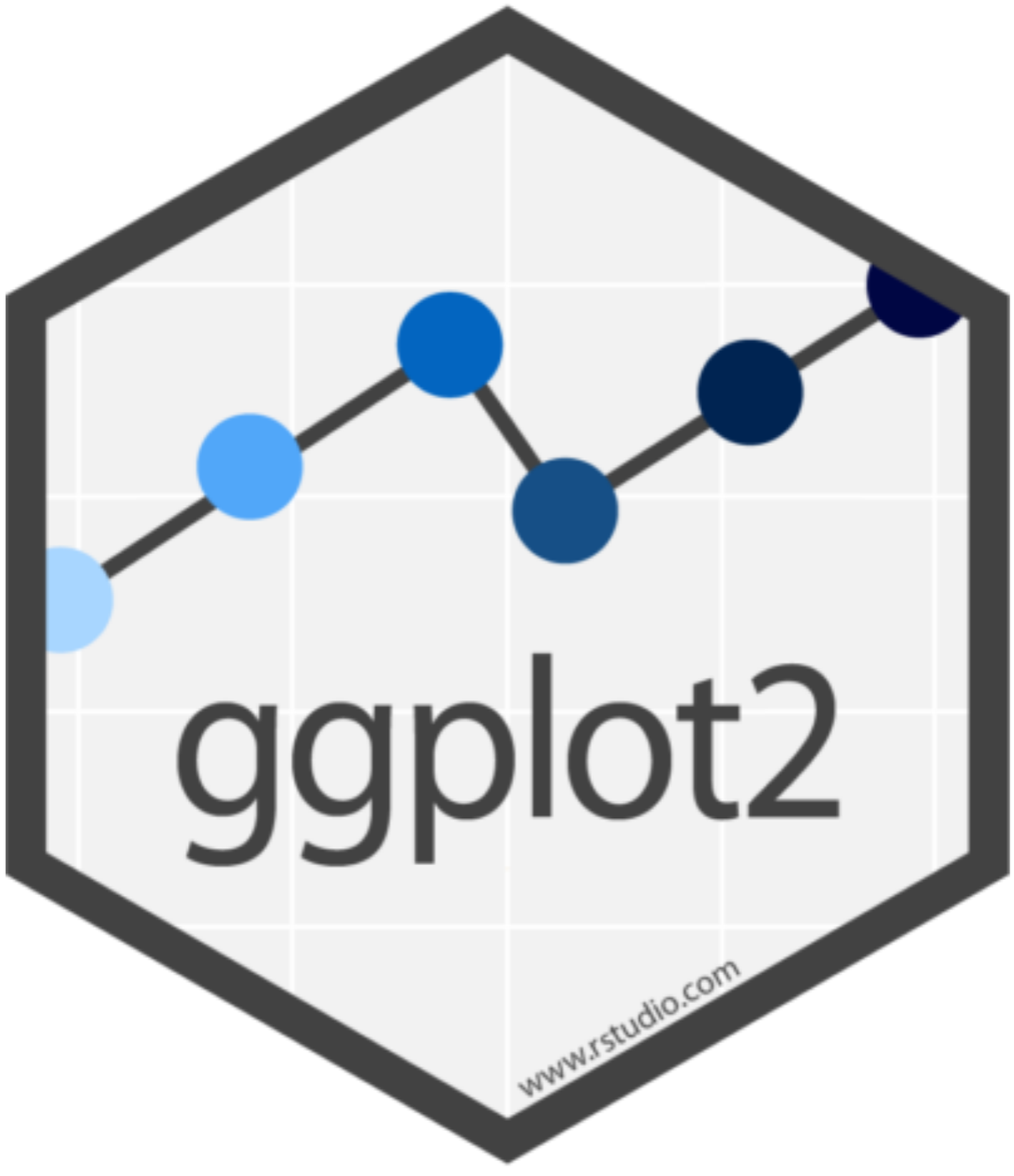
```
> tidy_plants %>%
+   pivot_wider(names_from = c(Element, Day),
+               names_sep = " day ",
+               values_from = Measurement)
# A tibble: 6 x 8
  Plant_number Treatment `Stem length da… `Leaf width day… `Stem length da…
  <chr>        <chr>                <dbl>            <dbl>            <dbl>
1 A1_14        control                120               21              122
2 A1_18        control                132               23              135
3 A1_21        control                131               18              133
4 A2_09        UV                     109               29              114
5 A3_02        UV                     125               25              127
6 A3_10        UV                     130               12              133
# … with 3 more variables: `Leaf width day 2` <dbl>, `Stem length day 3` <dbl>,
#   `Leaf width day 3` <dbl>
```

headers

content

Data visualisation

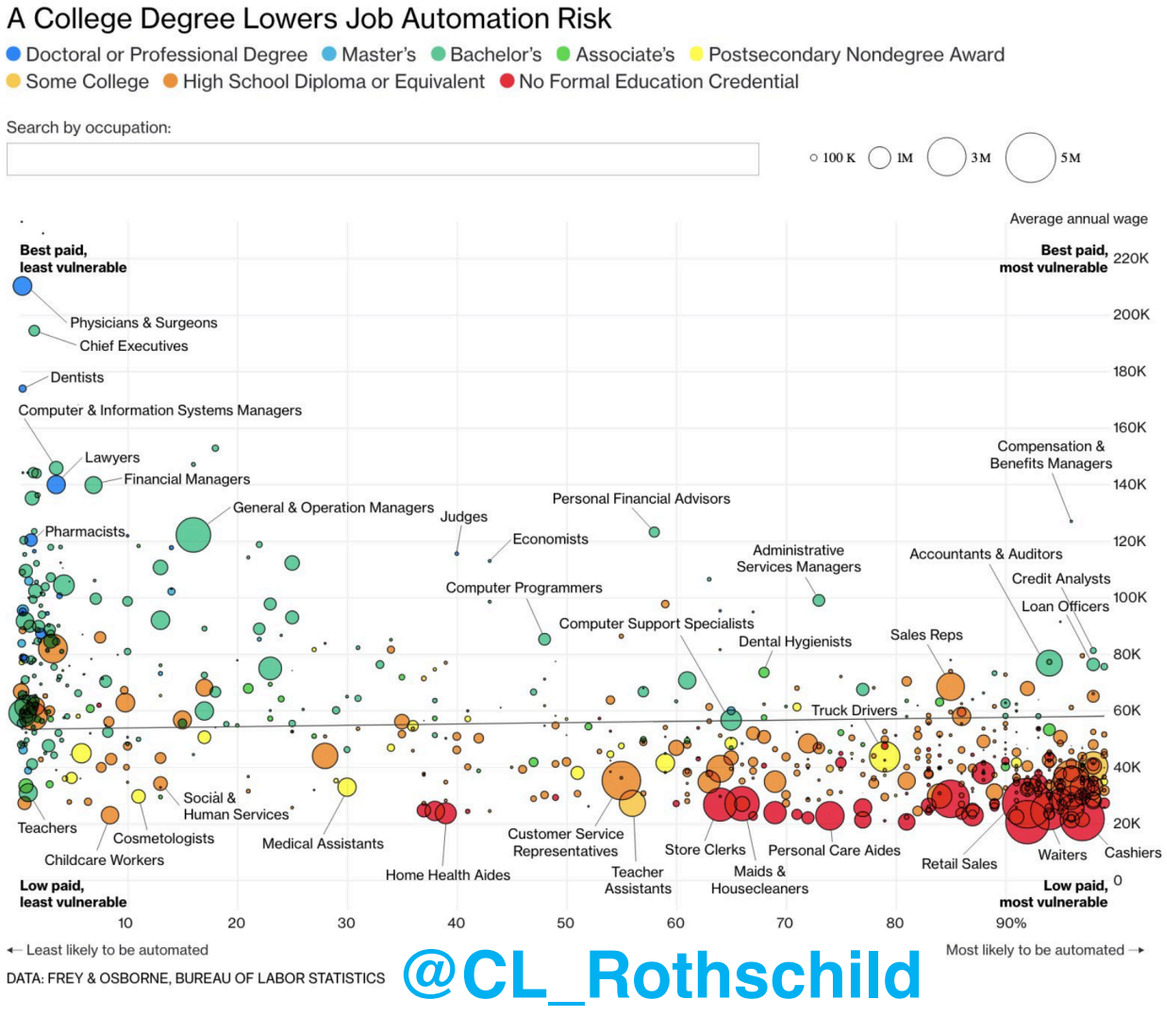
Create eye candy with ggplot2



PHILADELPHIA PARKING TICKETS

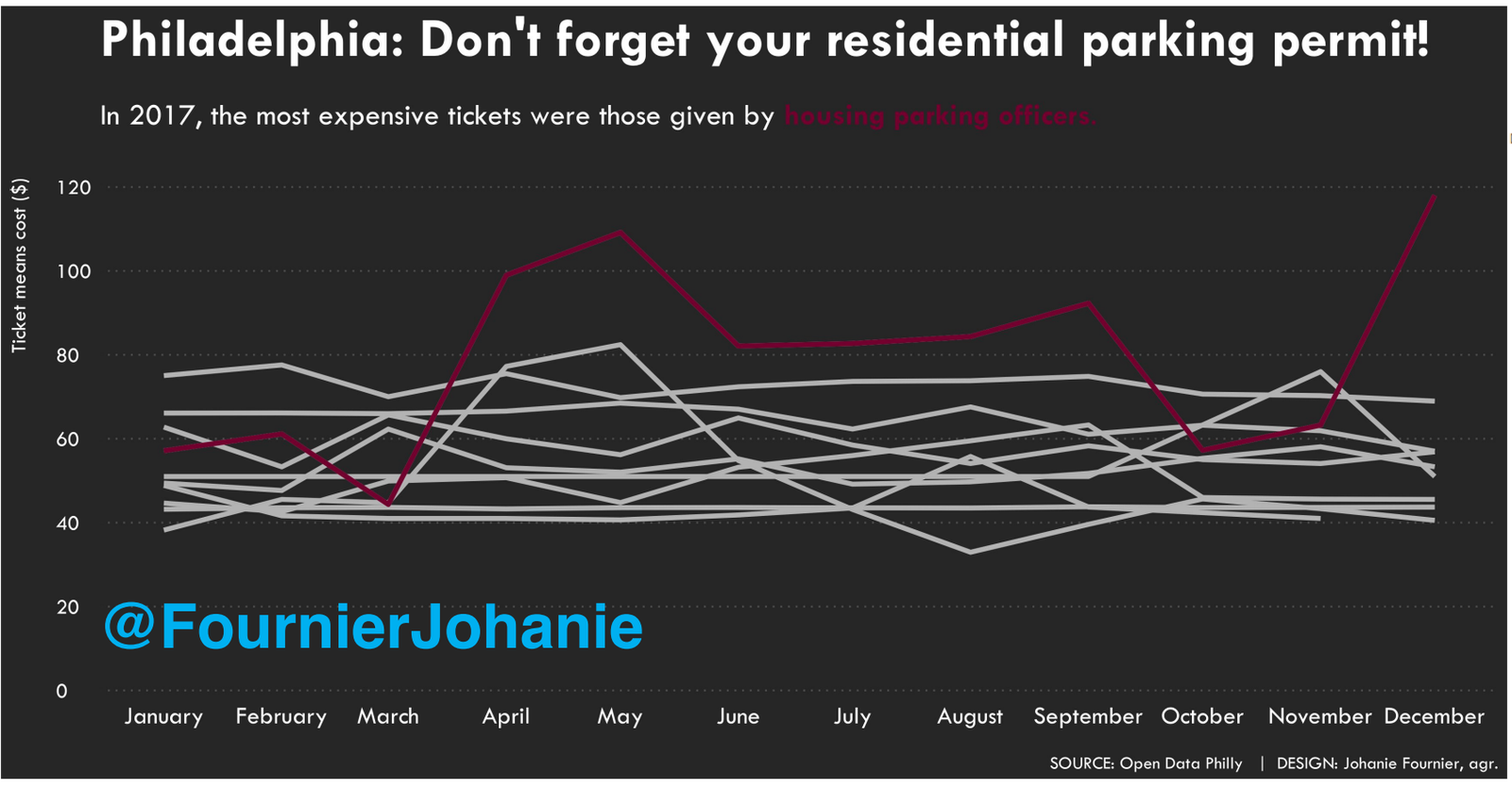
Density map of parking tickets issued in 2017

@The_Anna_L

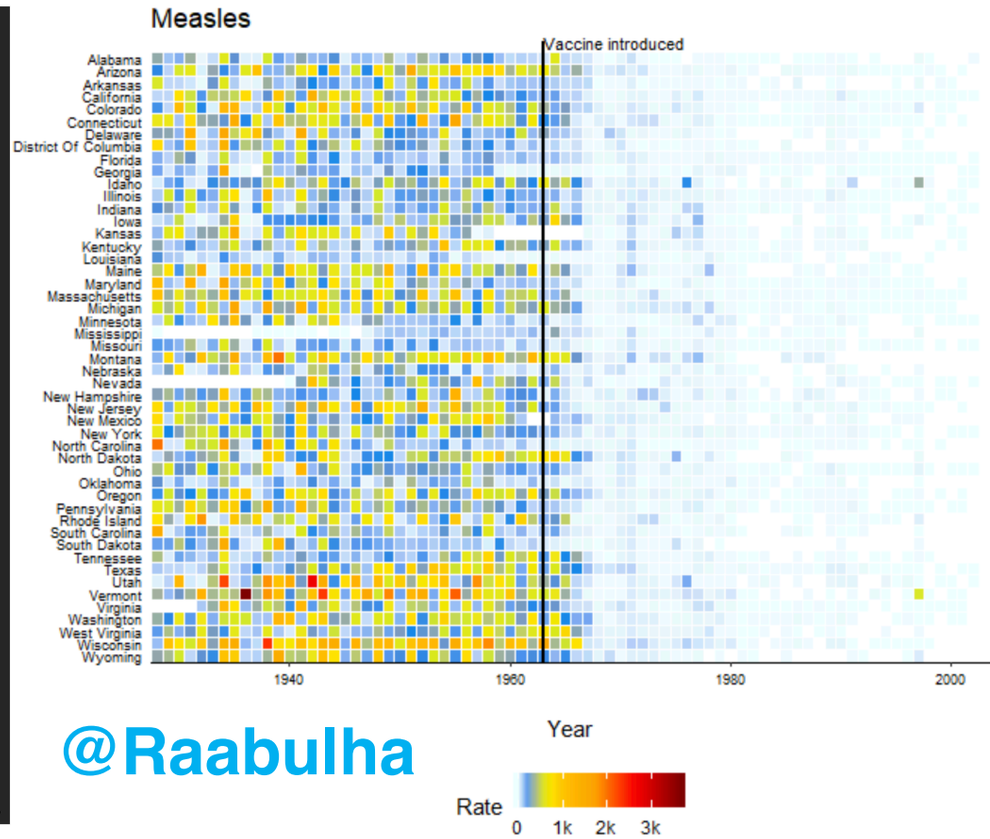


@CL_Rothschild

#tidytuesday



@FournierJohanie

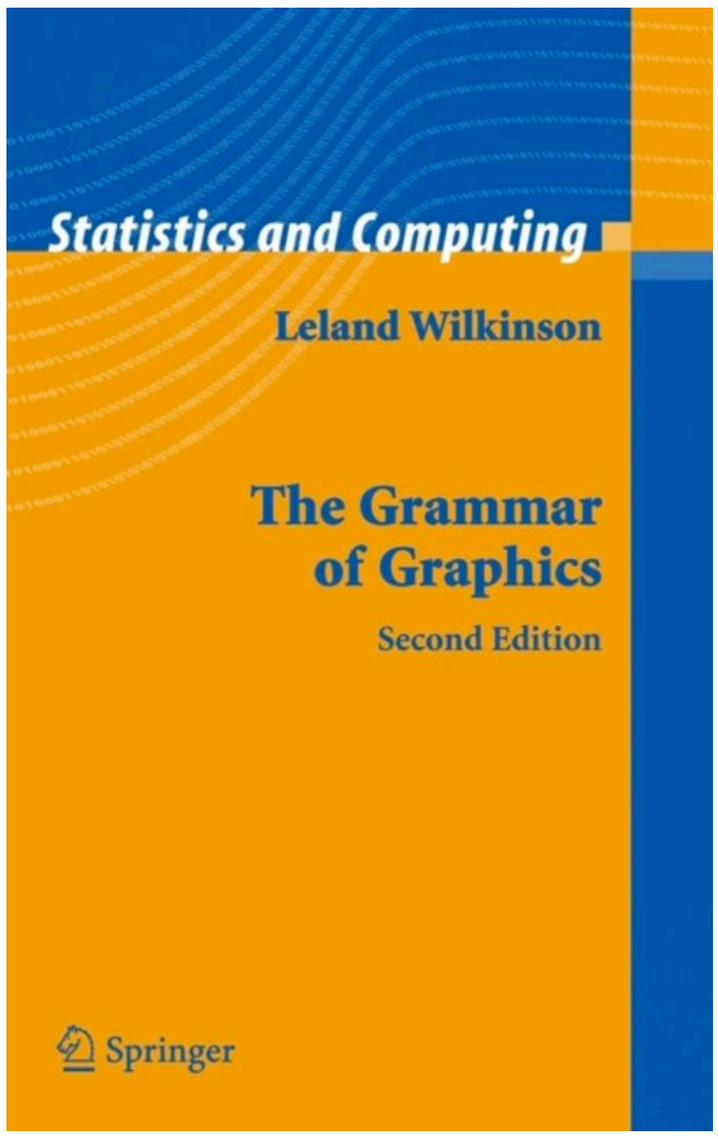


@Raabulha

Data visualisation with ggplot2

- **ggplot2** is an extremely popular data visualisation package for R
- Simple syntax, easy to learn, nice plots
- Developed and maintained by Hadley Wickham
- Based on the book: The Grammar of Graphics (Wilkinson, 2005)

The grammar of graphics



Data	The variables in a tibble or data.frame
Aesthetics	x- and y-axis, colour, size, alpha, shape
Geometries	point, line, bar, histogram

Source: https://www.youtube.com/watch?v=uiTc55clwuA

Other RDM workshops

- uu.nl/rdm > Training & workshops
- Learn to write your Data Management Plan (online course)
- Quickstart to Research Data Management
- Best Practices for Writing Reproducible Code
- Custom courses (such as this one!) at your department *(contact us!)*

Services and solutions to make research data management work

```
useR <- function(){
    print("Good luck and see you!")
}
useR()
```