

TIDY VERSE PART 1

dr Maria Kubara, WNE UW

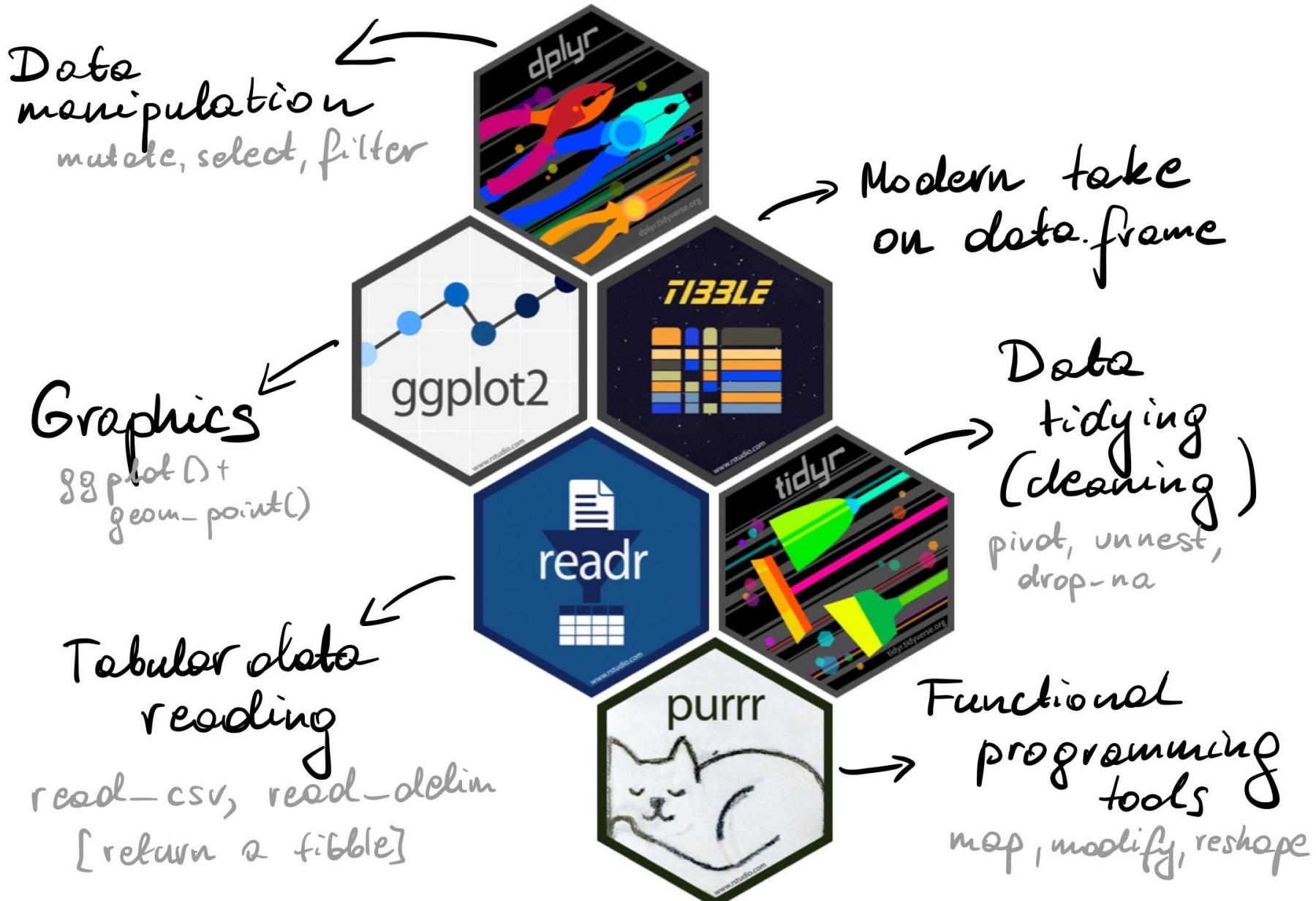
TIDYVERSE

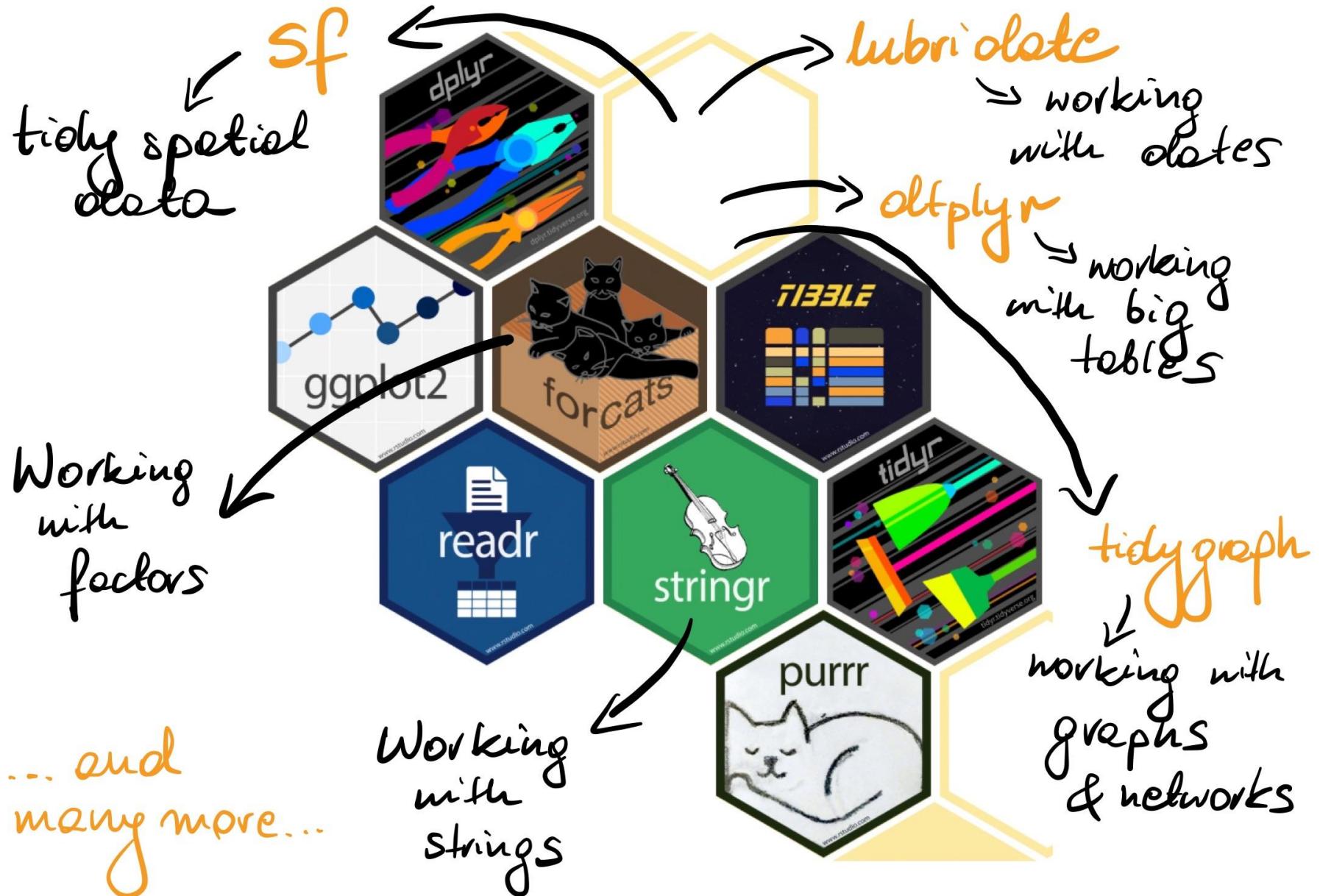
- Environment with packages for easier and tidier data processing in R
- All based on tibble structure (refreshed data.frame)
- Concerned about the tidy data
- „All packages share an underlying design philosophy, grammar, and data structures“



Install the complete tidyverse with:

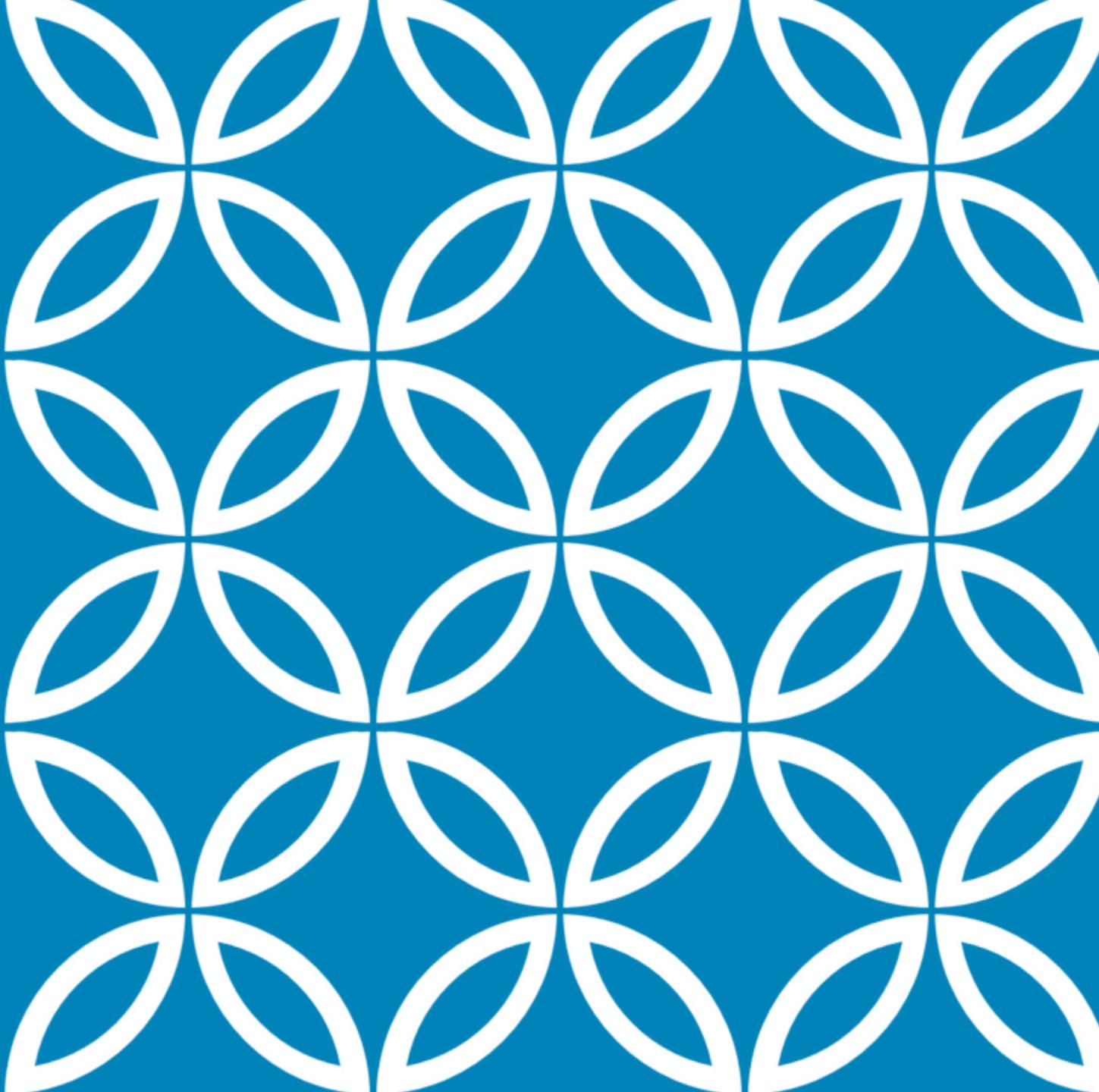
```
install.packages("tidyverse")
```





TIDY DATA

Golden standard in data analytics

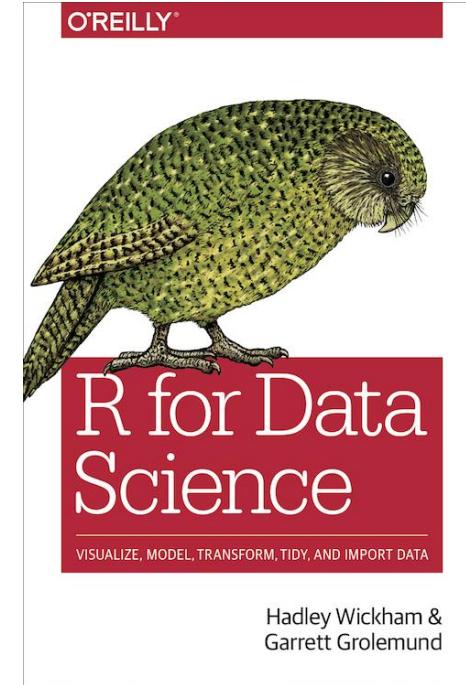


TIDYVERSE

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

In tidy data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.



TIDY DATA MANIPULATION

Four fundamental **verbs** of data manipulation:

- Filter: subsetting or removing observations based on some condition.
- Transform: adding or modifying variables. These modifications can involve either a single variable (e.g., log-transformation), or multiple variables (e.g., computing density from weight and volume).
- Aggregate: collapsing multiple values into a single value (e.g., by summing or taking means).
- Sort: changing the order of observations.

BASIC DATA MANIPULATION WITH DPLYR

Filter

`filter()`

Aggregate

`summarize()`

`count()`

*Group-wise
operations:
`group_by()`
`rowwise()`*

Transform

`mutate()`

Sort

`arrange()`

CONTEXT FOR TIDYVERSE FUNCTIONS

All tidyverse functions take DATA as their first argument and work with its context. In this way we can just specify the name of a variable (column) and use it without any \$ symbols to show its source.

```
filter(iris, Sepal.Length > 2)
```

filter function „knows” that Sepal.Length is a variable within iris dataset

```
mutate(iris, Double.Sepal.Length = Sepal.Length*2)
```

A new variable Double.Sepal.Length is created in the iris dataset
which uses the values in Sepal.Length to create its own

In tidyverse there is a convention: new goes first, older comes second
Here we first specify the name of the new variable and then get its
value (based on the older info already present in the dataset)

PIPELINE OPERATOR

Handy shortcut in
RStudio to write a
pipeline: Ctrl+Shift+M



%>% pipeline operator allows for passing on the result from the previous operation as the first input for the next function (or where the dot is – for more control)

```
filter(iris, Sepal.Length > 2)  
iris %>% filter(Sepal.Length > 2)  
iris %>% filter(., Sepal.Length > 2)
```

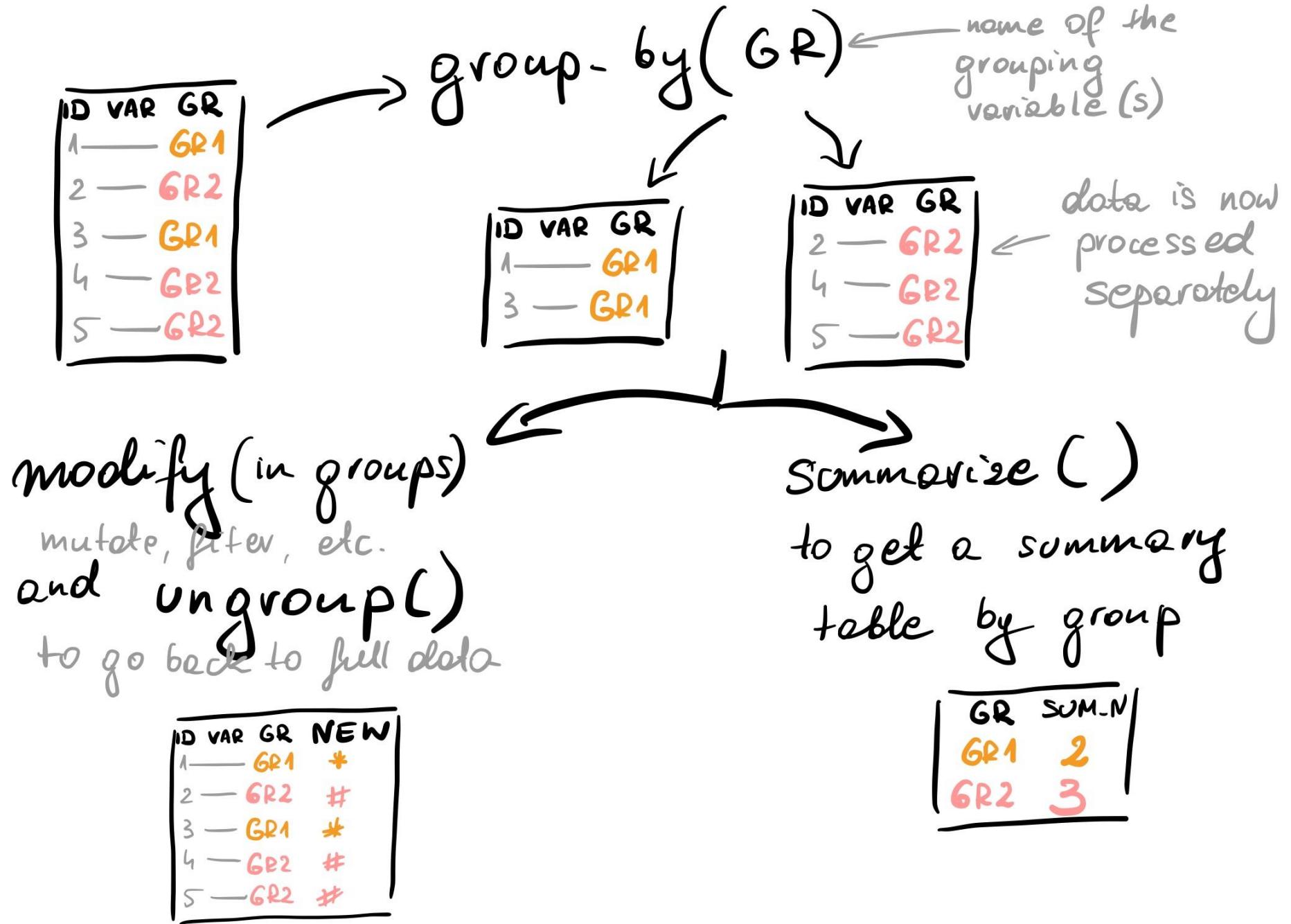
Note: in R 4.0 „base pipe” has been introduced, written as |>
It controls the function flow in a similar way like the pipeline operator. You can use it without loading any additional package.

GROUP-WISE OPERATIONS

group_by() „splits” the data by given group division and allows for processing the dataset separately for given groups

Mostly used for summary table creation or by adding a variable which value differs by groups

Similarly rowwise() allows for processing data for each row individually (group_by is far more popular)



OTHER HELPFUL DPLYR FUNCTIONS

| Dplyr function | Description |
|--|--|
| rename() | Changes the name of a given variable (new name first) |
| distinct() | Remove duplicated rows in the dataset |
| slice() | Select rows by positions (specify vector of positions which you'd like to get) |
| add_row() | Add new row to data (arguments go as names of variables and their value) |
| bind_cols() | Bind two datasets by columns (like cbind) |
| bind_rows() | Bind two datasets by rows (like rbind) |
| left_join()/right_join() inner_join()/full_join() | Merging functions which take their names exactly from SQL syntax |
| %in% | Operator for the filter() function, allows for checking if a value is within the values given in a longer vector e.g. gender %in% c("male","female") |
| glimpse | Get an overview of the data (like summary/str function) |

Data transformation with dplyr :: CHEAT SHEET



dplyr functions work with pipes and expect **tidy data**. In tidy data:



Each **variable** is in its own **column**



Each **observation**, or **case**, is in its own **row**



$x \%>% f(y)$ becomes $f(x, y)$

Summarise Cases

Apply **summary functions** to columns to create a new table of summary statistics. Summary functions take vectors as input and return one value (see back).

summary function →

- **summarise(.data, ...)**
Compute table of summaries.
`summarise(mtcars, avg = mean(mpg))`
- **count(.data, ..., wt = NULL, sort = FALSE, name = NULL)**
Count number of rows in each group defined by the variables in ... Also `tally()`.
`count(mtcars, cyl)`

Group Cases

Use `group_by(.data, ..., .add = FALSE, .drop = TRUE)` to create a "grouped" copy of a table grouped by columns in ... dplyr functions will manipulate each "group" separately and combine the results.



Use `rowwise(.data, ...)` to group data into individual rows. dplyr functions will compute results for each row. Also apply functions to list-columns. See tidyverse cheat sheet for list-column workflow.

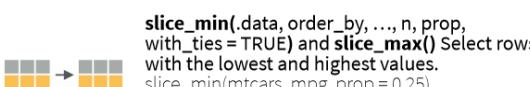
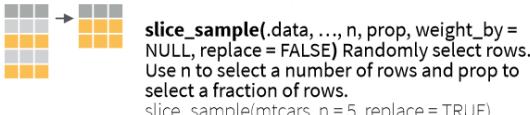


`ungroup(x, ...)` Returns ungrouped copy of table.
`ungroup(g_mtcars)`

Manipulate Cases

EXTRACT CASES

Row functions return a subset of rows as a new table.

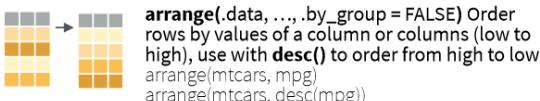


Logical and boolean operators to use with filter()

| | | | | | | |
|----|---|----|----------|------|---|-------|
| == | < | <= | is.na() | %in% | | xor() |
| != | > | >= | !is.na() | ! | & | |

See `?base::Logic` and `?Comparison` for help.

ARRANGE CASES



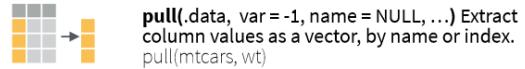
ADD CASES



Manipulate Variables

EXTRACT VARIABLES

Column functions return a set of columns as a new vector or table.



Use these helpers with `select()` and `across()`
e.g. `select(mtcars, mpg:cyl)`

`contains(match)` `num_range(prefix, range)` ; e.g. `mpg:cyl`
`ends_with(match)` `all_of(x)/any_of(x, ..., vars)` ; e.g. `-gear`
`starts_with(match)` `matches(match)` `everything()`

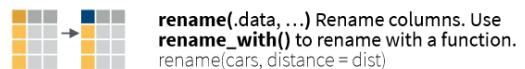
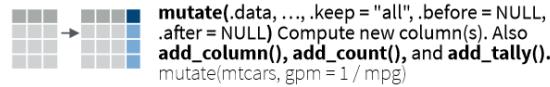
MANIPULATE MULTIPLE VARIABLES AT ONCE



MAKE NEW VARIABLES

Apply **vectorized functions** to columns. Vectorized functions take vectors as input and return vectors of the same length as output (see back).

vectorized function →



Vectorized Functions

TO USE WITH MUTATE ()

mutate() and **transmute()** apply vectorized functions to columns to create new columns. Vectorized functions take vectors as input and return vectors of the same length as output.

vectorized function

OFFSET

dplyr::lag() - offset elements by 1
dplyr::lead() - offset elements by -1

CUMULATIVE AGGREGATE

dplyr::cumall() - cumulative all()
dplyr::cumany() - cumulative any()
dplyr::cummax() - cumulative max()
dplyr::cummean() - cumulative mean()
dplyr::cummin() - cumulative min()
dplyr::cumprod() - cumulative prod()
dplyr::cumsum() - cumulative sum()

RANKING

dplyr::cume_dist() - proportion of all values <=
dplyr::dense_rank() - rank w/ ties = min, no gaps
dplyr::min_rank() - rank with ties = min
dplyr::ntile() - bins into n bins
dplyr::percent_rank() - min_rank scaled to [0,1]
dplyr::row_number() - rank with ties = "first"

MATH

+, -, *, /, ^, %/%, %% - arithmetic ops
log(), log2(), log10() - logs
<, <=, >, >=, !=, == - logical comparisons
dplyr::between() - x >= left & x <= right
dplyr::near() - safe == for floating point numbers

MISCELLANEOUS

dplyr::case_when() - multi-case if_else()
starwars %>%
 mutate(type = case_when(
 height > 200 | mass > 200 ~ "large",
 species == "Droid" ~ "robot",
 TRUE ~ "other"))
dplyr::coalesce() - first non-NA values by element across a set of vectors
dplyr::if_else() - element-wise if() + else()
dplyr::na_if() - replace specific values with NA
 pmax() - element-wise max()
 pmin() - element-wise min()

Summary Functions

TO USE WITH SUMMARISE ()

summarise() applies summary functions to columns to create a new table. Summary functions take vectors as input and return single values as output.

summary function

COUNT

dplyr::n() - number of values/rows
dplyr::n_distinct() - # of uniques
 sum(!is.na()) - # of non-NAs

POSITION

mean() - mean, also mean(!is.na())
median() - median

LOGICAL

mean() - proportion of TRUE's
sum() - # of TRUE's

ORDER

dplyr::first() - first value
dplyr::last() - last value
dplyr::nth() - value in nth location of vector

RANK

quantile() - nth quantile
min() - minimum value
max() - maximum value

SPREAD

IQR() - Inter-Quartile Range
mad() - median absolute deviation
sd() - standard deviation
var() - variance

Row Names

Tidy data does not use rownames, which store a variable outside of the columns. To work with the rownames, first move them into a column.

A B C A B
1 a t → 1 a t 1 a t
2 b u → 2 b u 2 b u
3 c v → 3 c v 3 c v

A B C A B
1 a t → 1 a 1 a
2 b u → 2 b 2 b
3 c v → 3 c 3 c

Also **tibble::has_rownames()** and **tibble::remove_rownames()**.

Combine Tables

COMBINE VARIABLES

| | |
|-------|-------|
| x | y |
| A B C | E F G |
| a t 1 | a t 3 |
| b u 2 | b u 2 |
| c v 3 | d w 1 |

bind_cols(..., .name_repair) Returns tables placed side by side as a single table. Column lengths must be equal. Columns will NOT be matched by id (to do that look at Relational Data below), so be sure to check that both tables are ordered the way you want before binding.

RELATIONAL DATA

Use a "Mutating Join" to join one table to columns from another, matching values with the rows that they correspond to. Each join retains a different combination of values from the tables.

| | |
|----------|---|
| A B C D | y |
| a t 1 3 | |
| b u 2 2 | |
| c v 3 NA | |

left_join(x, y, by = NULL, copy = FALSE,

suffix = c("x", "y"), ..., keep = FALSE,

na_matches = "na") Join matching

values from y to x.

| | |
|-----------|---|
| A B C D | y |
| a t 1 3 | |
| b u 2 2 | |
| d w N A 1 | |

right_join(x, y, by = NULL, copy = FALSE,

suffix = c("x", "y"), ..., keep = FALSE,

na_matches = "na") Join matching

values from x to y.

| | |
|---------|---|
| A B C D | y |
| a t 1 3 | |
| b u 2 2 | |

inner_join(x, y, by = NULL, copy = FALSE,

suffix = c("x", "y"), ..., keep = FALSE,

na_matches = "na") Join data. Retain

only rows with matches.

| | |
|-----------|---|
| A B C D | y |
| a t 1 3 | |
| b u 2 2 | |
| c v 3 NA | |
| d w N A 1 | |

full_join(x, y, by = NULL, copy = FALSE,

suffix = c("x", "y"), ..., keep = FALSE,

na_matches = "na") Join data. Retain all

values, all rows.

COLUMN MATCHING FOR JOINS

| | |
|---------------|---|
| A B x C B y D | y |
| a t 1 t 3 | |
| b u 2 u 2 | |
| c v 3 N A N A | |

Use **by = c("col1", "col2", ...)** to specify one or more common columns to match on.

left_join(x, y, by = "A")

| | |
|-------------------|---|
| A x B x C A y B y | y |
| a t 1 d w | |
| b u 2 b u | |
| c v 3 a t | |

Use a named vector, **by = c("col1" = "col2", ...)**, to match on columns that have different names in each table.

left_join(x, y, by = c("C" = "D"))

| | |
|-------------------|---|
| A 1 B 1 C A 2 B 2 | y |
| a t 1 d w | |
| b u 2 b u | |
| c v 3 a t | |

Use **suffix** to specify the suffix to give to unmatched columns that have the same name in both tables.

left_join(x, y, by = c("C" = "D"), suffix = c("1", "2"))

COMBINE CASES

| | |
|-------|-------|
| x | y |
| A B C | a t 1 |
| b u 2 | b u 2 |
| c v 3 | d w 1 |

bind_rows(..., id = NULL)
Returns tables one on top of the other as a single table. Set **.id** to a column name to add a column of the original table names (as pictured).



Use a "Filtering Join" to filter one table against the rows of another.

| | |
|-------|-------|
| x | y |
| A B C | a t 1 |
| b u 2 | b u 2 |
| c v 3 | d w 1 |

semi_join(x, y, by = NULL, copy = FALSE, ..., na_matches = "na") Return rows of x that have a match in y. Use to see what will be included in a join.

anti_join(x, y, by = NULL, copy = FALSE, ..., na_matches = "na") Return rows of x that do not have a match in y. Use to see what will not be included in a join.

Use a "Nest Join" to inner join one table to another into a nested data frame.

| | |
|-------|--------------|
| A B C | y |
| a t 1 | tibble [1x2] |
| b u 2 | tibble [1x2] |
| c v 3 | tibble [1x2] |

nest_join(x, y, by = NULL, copy = FALSE, keep = FALSE, name = "y", ...) Join data, nesting matches from y in a single new data frame column.

SET OPERATIONS

intersect(x, y, ...)
Rows that appear in both x and y.

setdiff(x, y, ...)
Rows that appear in x but not y.

union(x, y, ...)
Rows that appear in x or y.
(Duplicates removed). **union_all()** retains duplicates.

Use **setequal()** to test whether two data sets contain the exact same rows (in any order).