

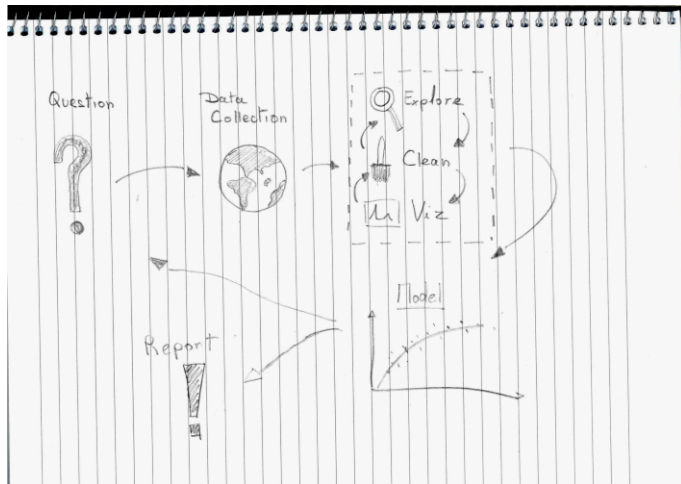
Data manipulation with dplyr

Damien Georges

International Agency for Research on Cancer

June 2023 - Tartu

Epidemiological study workflow



Data manipulation tools



- ▶ R core function
- ▶ dplyr
- ▶ data.table
- ▶ ...

=> The best tool is the one you feel the most comfortable with

Tidyverse (from www.tidyverse.org)

R packages 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.



pipe functions %>%

```
chill(fold(add(melt(add(chocolate, butter)),  
                beat(add(eggs.white, cream))))
```

pipe functions %>%

```
chill(fold(add(melt(add(chocolate, butter)),  
                beat(add(eggs.white, cream)))))
```

```
chocolate %>%  
  add(butter) %>%  
  melt() %>%  
  add(  
    eggs.white %>%  
      add(cream) %>%  
        beat()  
  ) %>%  
  fold() %>%  
  chill()
```






data manipulation with



Code as you speak: Data manipulation with dplyr is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- ▶ select rows (`slice`)
- ▶ select columns (`filter`)
- ▶ arrange rows (`arrange`)
- ▶ columns selection (`select`)
- ▶ create/modify columns (`mutate`)
- ▶ group and summarize data (`group_by` and `summarise`)
- ▶ bind different tables (`bind_rows`, `bind_cols`)
- ▶ merge different tables (`left_join`, `right_join`, `inner_join`, `full_join`)

discovering other tidyverse packages features

- ▶ data visualization with  (ggplot, geom_bars, ...)
- ▶ pivoting data with  (pivot_wider, pivo_longer)
- ▶ reading data with  (read_table, read_csv)
- ▶ manipulating lists with  (map, map_chr, reduce, ...)
- ▶ manipulating strings with  (str_remove)