

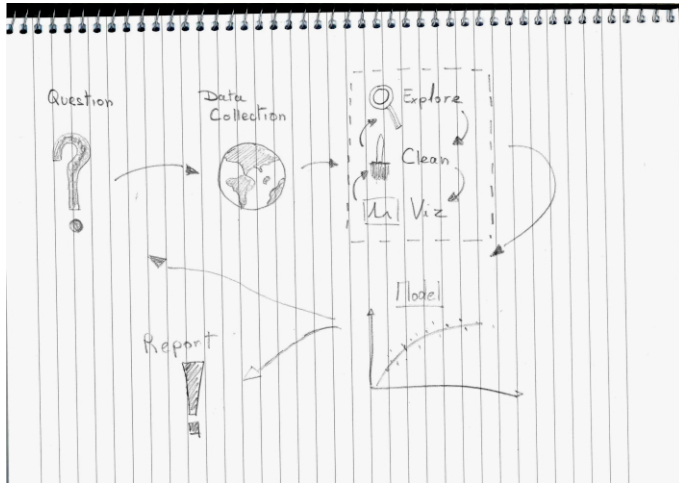
# Data manipulation with dplyr

Damien Georges

International Agency for Research on Cancer

August 2019 - Tartu

# Epidemiological study workflow



# Data manipulation tools



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

# Tidyverse (from [www.tidyverse.org](http://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()
```

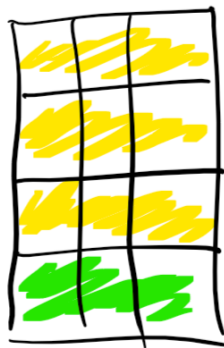
## code as you speak

Data manipulation with `dplyr` is done using a limited number of **verbs** corresponding to an action to be applied to a table.

- ▶ `slice`
- ▶ `filter`
- ▶ `arrange`
- ▶ `select`
- ▶ `mutate`
- ▶ `group_by`
- ▶ `summarize`
- ▶ `join`
- ▶ `...`

select rows

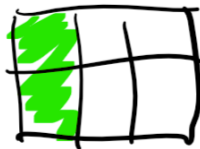
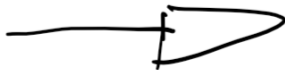
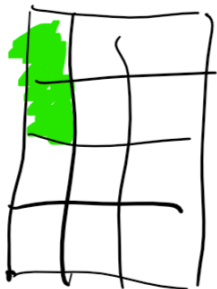
```
dat %>% slice(c(1, 1, 1, 2))
```





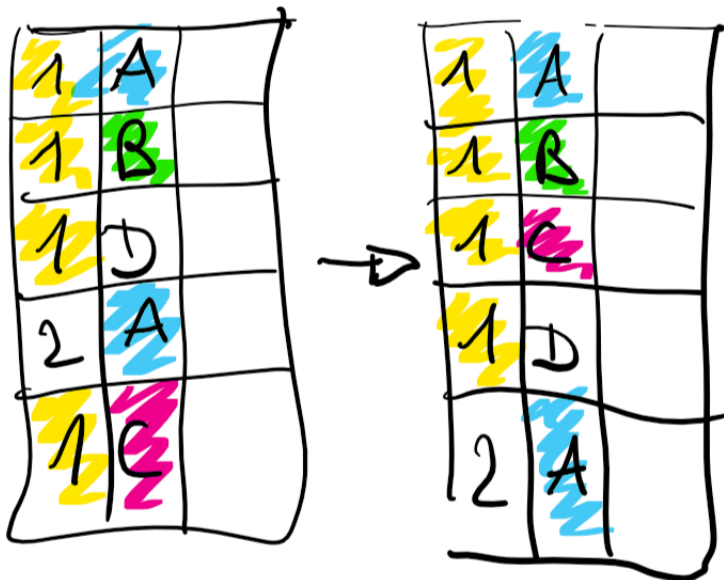
## select columns

```
dat %>% filter(C1 == 'green')
```



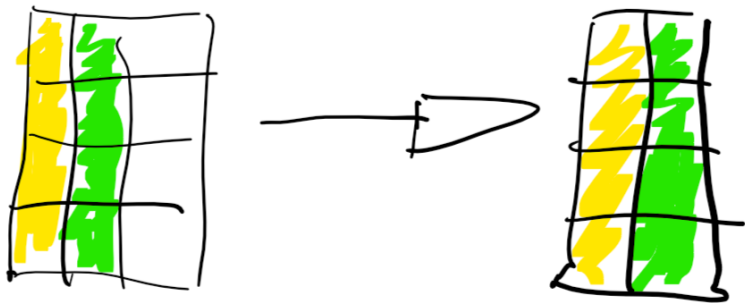
arrange rows

```
dat %>% arrange(C1, C2)
```



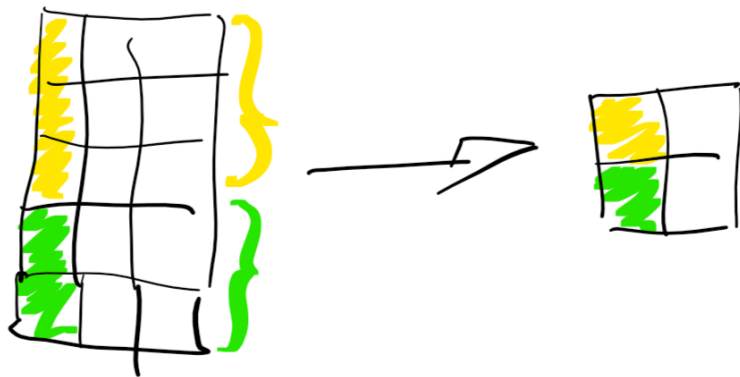
## columns selection

```
dat %>% select(C1, C2)
```



## group and summarize data

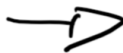
```
dat %>% group_by(C1) %>% summarize(C4 = mean(C2 + C3))
```



**note:** `summarise()` is an alias for `summarize()`

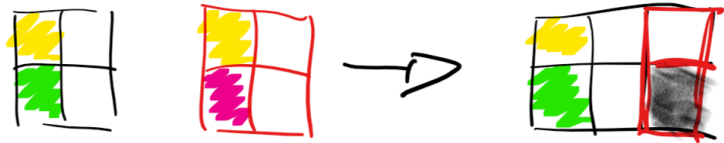
## bind and merge tables

```
dat1 %>% bind_rows(dat2)
```



## bind and merge tables

```
dat1 %>% left_join(dat2)
```



**note:** right\_join will keep all rows of dat2

## bind and merge tables

```
dat1 %>% inner_join(dat2)
```



## bind and merge tables

```
dat1 %>% full_join(dat2)
```

