

Data manipulation with dplyr

Damien G.

August 2019 - Tartu

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.



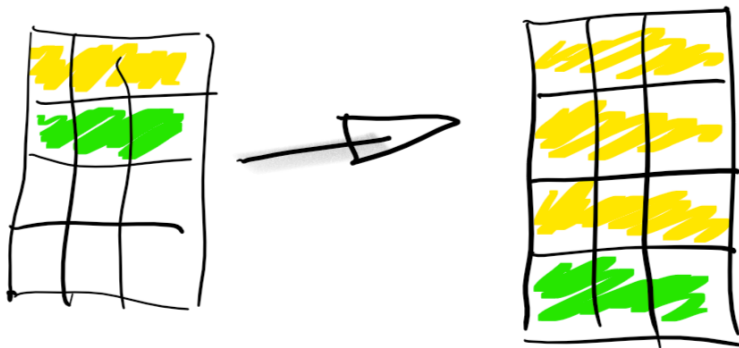
code as you speak

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

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

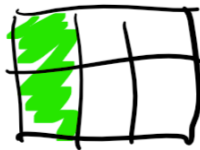
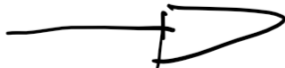
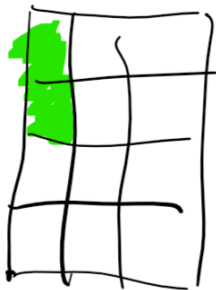
select rows

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



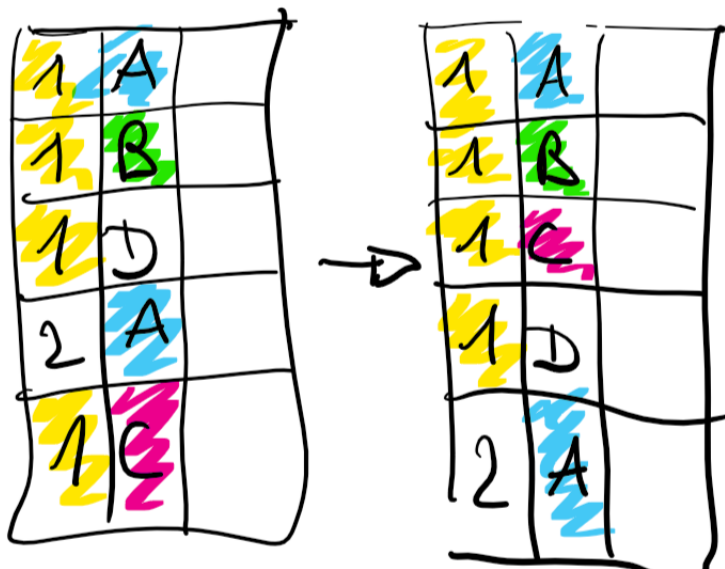
select columns

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



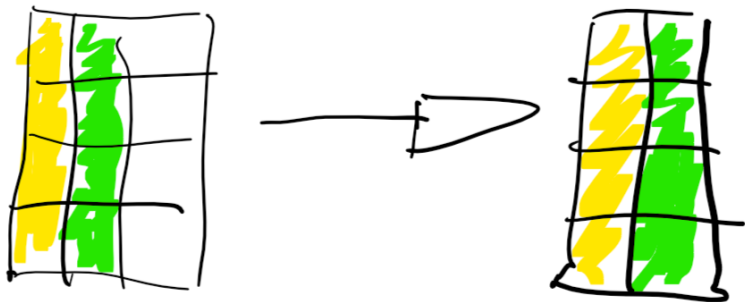
arrange rows

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



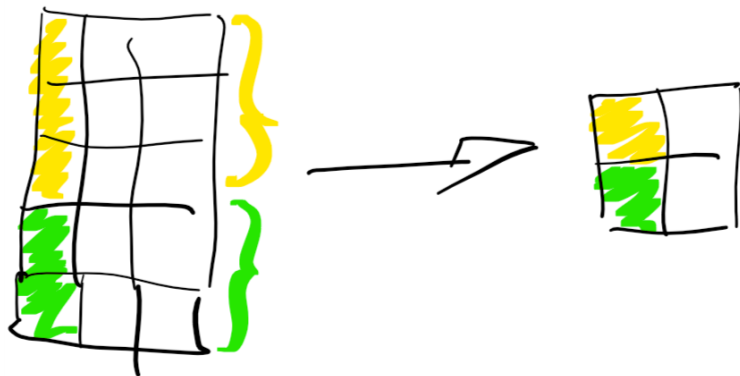
columns selection

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



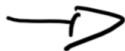
group and summarize data

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



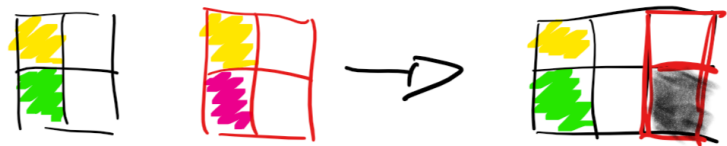
bind and merge tables

```
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)
```

