ARISTOTLE UNIVERSITY OF THESSALONIKI

FACULTY OF HEALTH SCIENCES - SCHOOL OF MEDICINE

MSc Health Statistics and Data Analytics

data wrangling

KONSTANTINOS I. BOUGIOUKAS

PHYSICIST, BIOSTATISTICIAN AND RESEARCH METHODOLOGIST

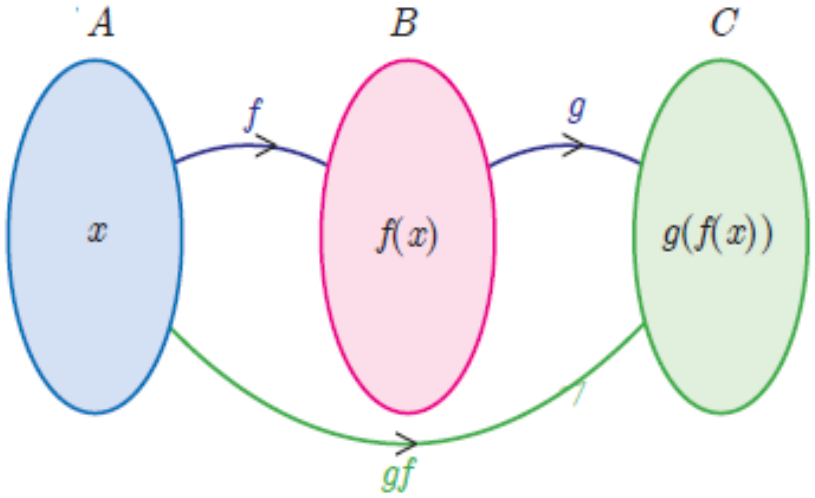
THESSALONIKI 2021-22

Core tidyverse packages

Pipe operator %>%

Perform a sequence of operations on a data frame x using the functions f(), and g() :

```
g(f(x))
```



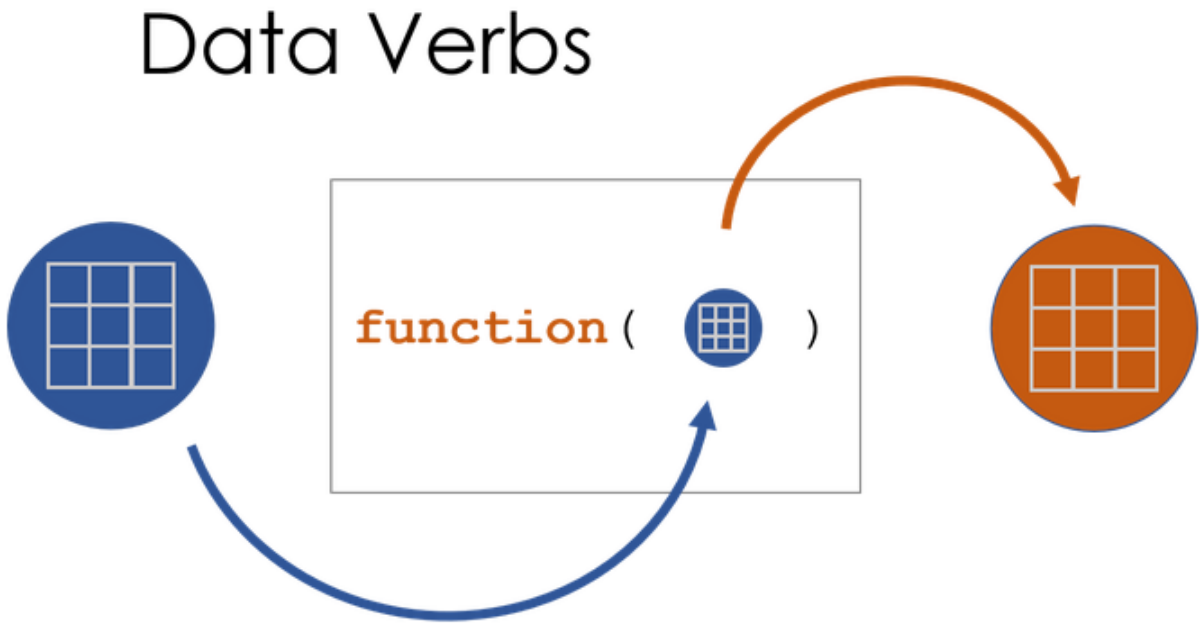
With %>%:

```
x  %>%
   f()  %>%
   g()
```

We would read this sequence as:

1. Take x *then*
2. Use this output as the input to the function f() *then*
3. Use this output as the input to the next function g()

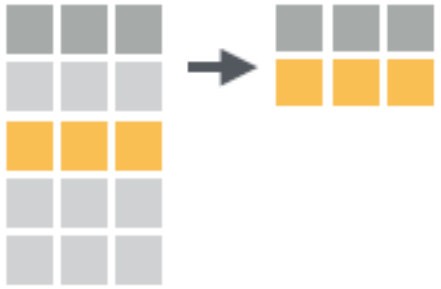
Common dplyr functions



A data verb requires a dataset as input, and returns a transformed dataset.

Common dplyr functions: filter() and select()

EXTRACT CASES



`filter()`

Extract rows that meet logical criteria.

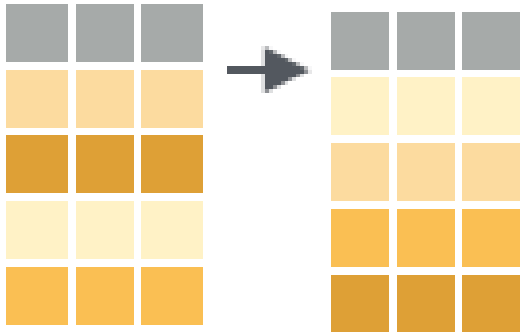
EXTRACT VARIABLES

`select()`

Extract columns as a table.

Common dplyr functions: arrange() and mutate()

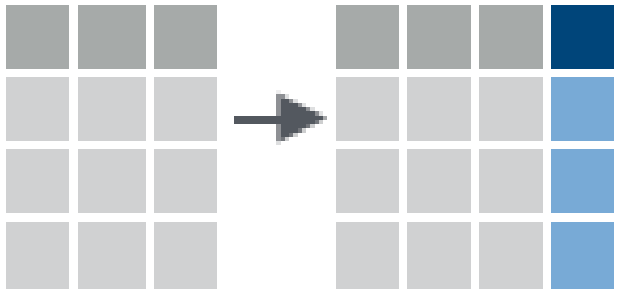
ARRANGE CASES



arrange()

Order rows by values of a column or columns

MAKE NEW VARIABLES



mutate()

Creates a new transformed variable from the formula we specify and adds it to the end of the original dataset.

Common dplyr functions: summarise()

city	particle size	amount (μg/m³)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

median	variance
22.5	1731.6

summarise()

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

city	particle size	amount (μg/m³)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	particle size	amount (μg/m³)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

group_by()

creates a "grouped" copy of a table grouped by columns in ... dplyr functions will manipulate each "group" separately and combine the results.

Common dplyr functions: group_by() + summarise()

city	particle size	amount (μg/m³)
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

city	mean	sum	n
New York	18.5	37	2
London	19.0	38	2
Beijing	88.5	177	2

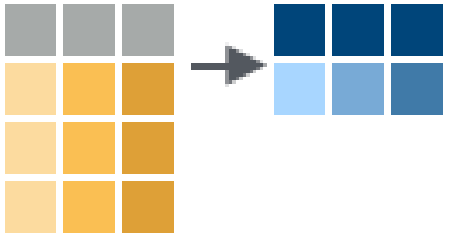
group_by() + summarise()

group_by() creates a "grouped" copy of a table grouped by columns.

summarise() manipulates each "group" separately and combine the results.

Common dplyr functions: across()

MANIPULATE MULTIPLE VARIABLES AT ONCE



across()

summarises or mutate multiple columns in the same way.

```
summarise(data, across(everything(), mean))
```