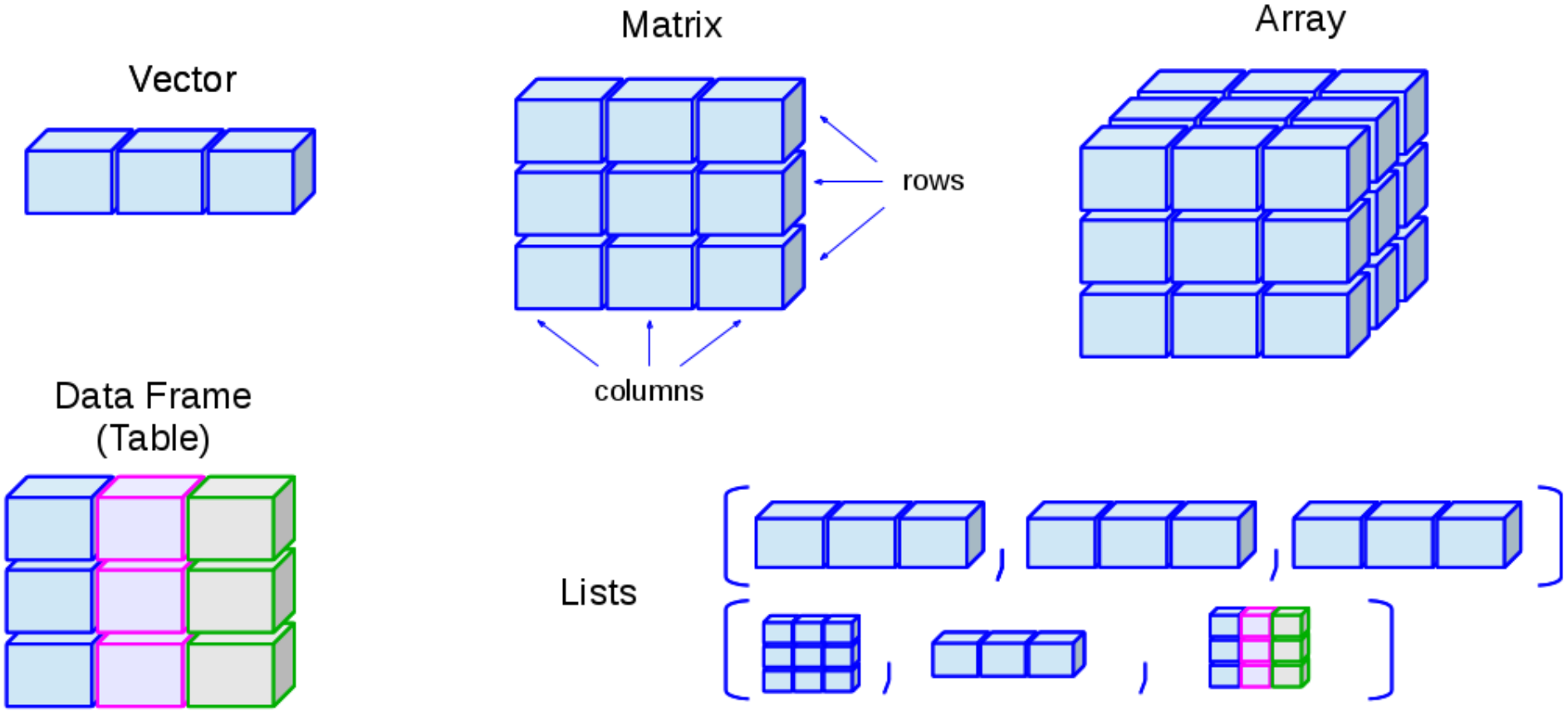
Coding e Big Data

2023-2024

Vincenzo Nardelli

vincenzo.nardelli@unicatt.it

Strutture dati in R



Tidyverse – tidyverse.org

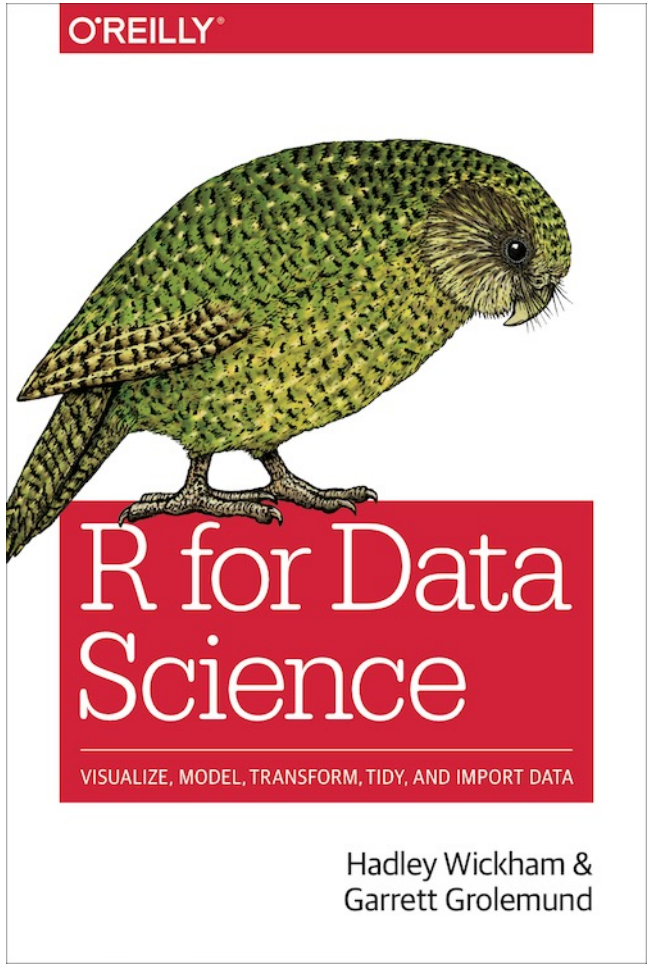
R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

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

Libro R for Data science



Versione italiana: https://it.r4ds.hadley.nz/

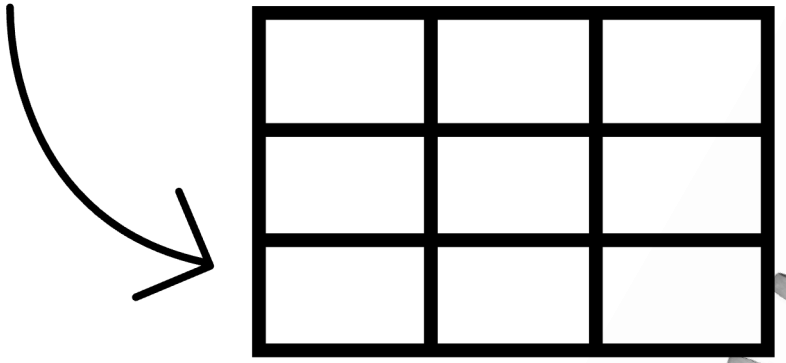
Versione inglese: https://r4ds.hadley.nz/

dplyr

dplyr è una **grammatica** della manipolazione dei dati, che fornisce un **insieme coerente di verbi** che ti aiutano a risolvere le sfide più comuni di manipolazione dei dati

d = data, data.frame

plyr = pinze (pliers)



dplyr – funzioni/verbi principali

select()

Subset variabili (colonne)

filter()

Subset osservazioni (righe)

mutate()

Crea nuove variabili

summarise()

Riassumi

dplyr – funzioni per gruppi

group_by()

Raggruppa per

+

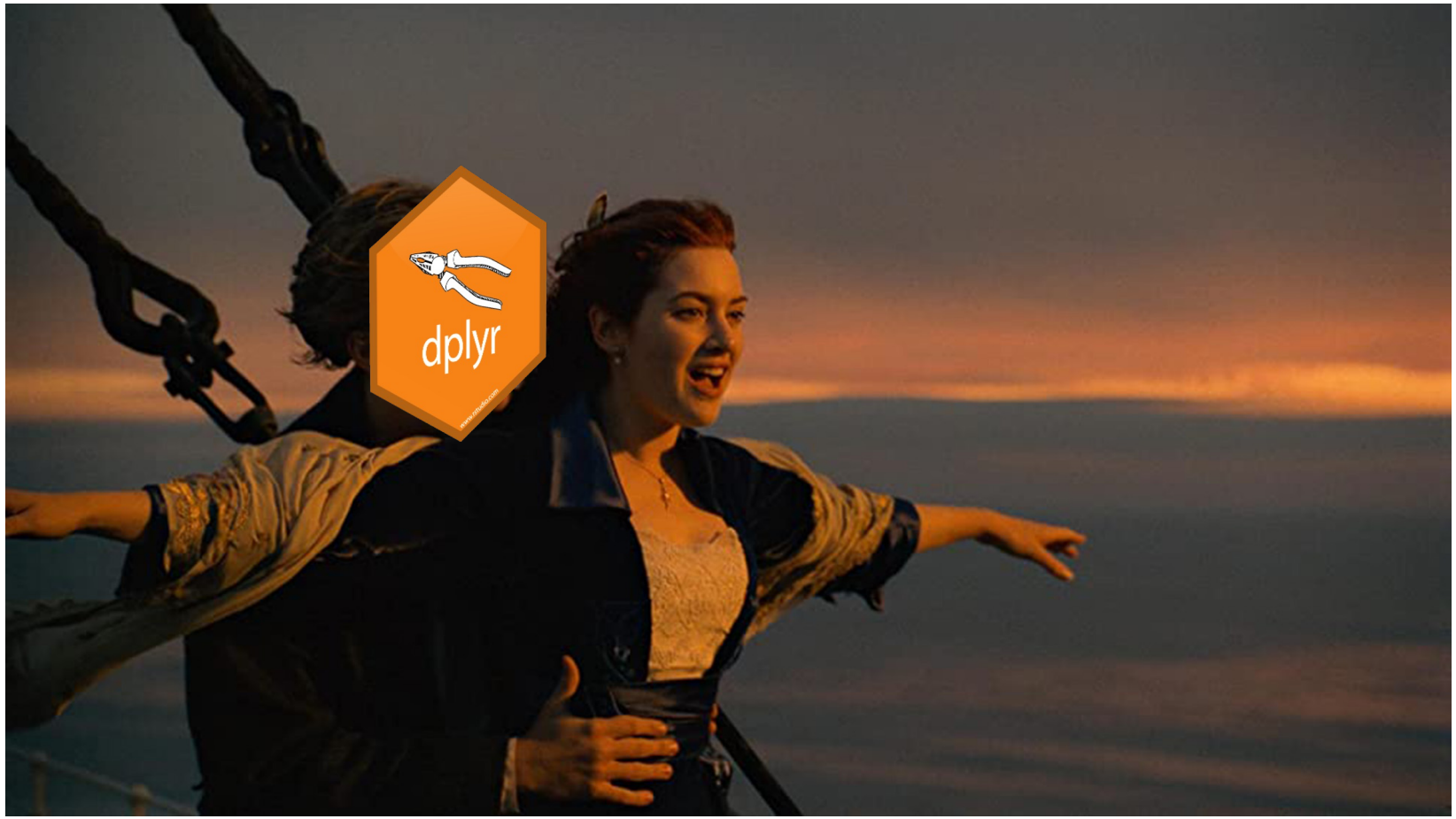
mutate()

Crea nuove variabili

summarise()

Riassumi

LAB Analizzare dati del Titanic con dplyr



LAB Titanic

- Quante persone oltre i 50 anni erano sul titanic?
- Quale era la percentuale di uomini e donne tra gli over 50?
- In valore assoluto ci sono più uomini tra gli over 50 o tra gli under 50? Quante donne sono under 50?
- In percentuale sono sopravvissuti più passeggeri in prima o terza classe?
- Quale è stato il costo del biglietto minimo, medio e massimo, per ogni classe e per ogni sesso?
- I sopravvissuti hanno pagato di più il viaggio?

LAB Titanic

- Le famiglie più numerose hanno avuto una percentuale più alta di sopravvivenza?
- C'è differenza nelle percentuali di sopravvivenza rispetto al porto d'imbarco?

Extra!

- Chi ha pagato i 10 biglietti più costosi?
- Quali sono i biglietti associati a più passeggeri?