

# **Workshop di ricerca sociale**

Introduzione all'analisi dei dati

# Cosa faremo oggi

---

## Quando i dati sbagliano

- Errore di rappresentazione
- Errore di misura

## Analisi dati

- Esempio di analisi esplorativa
- Comunicare i risultati

# Quando i dati sbagliano

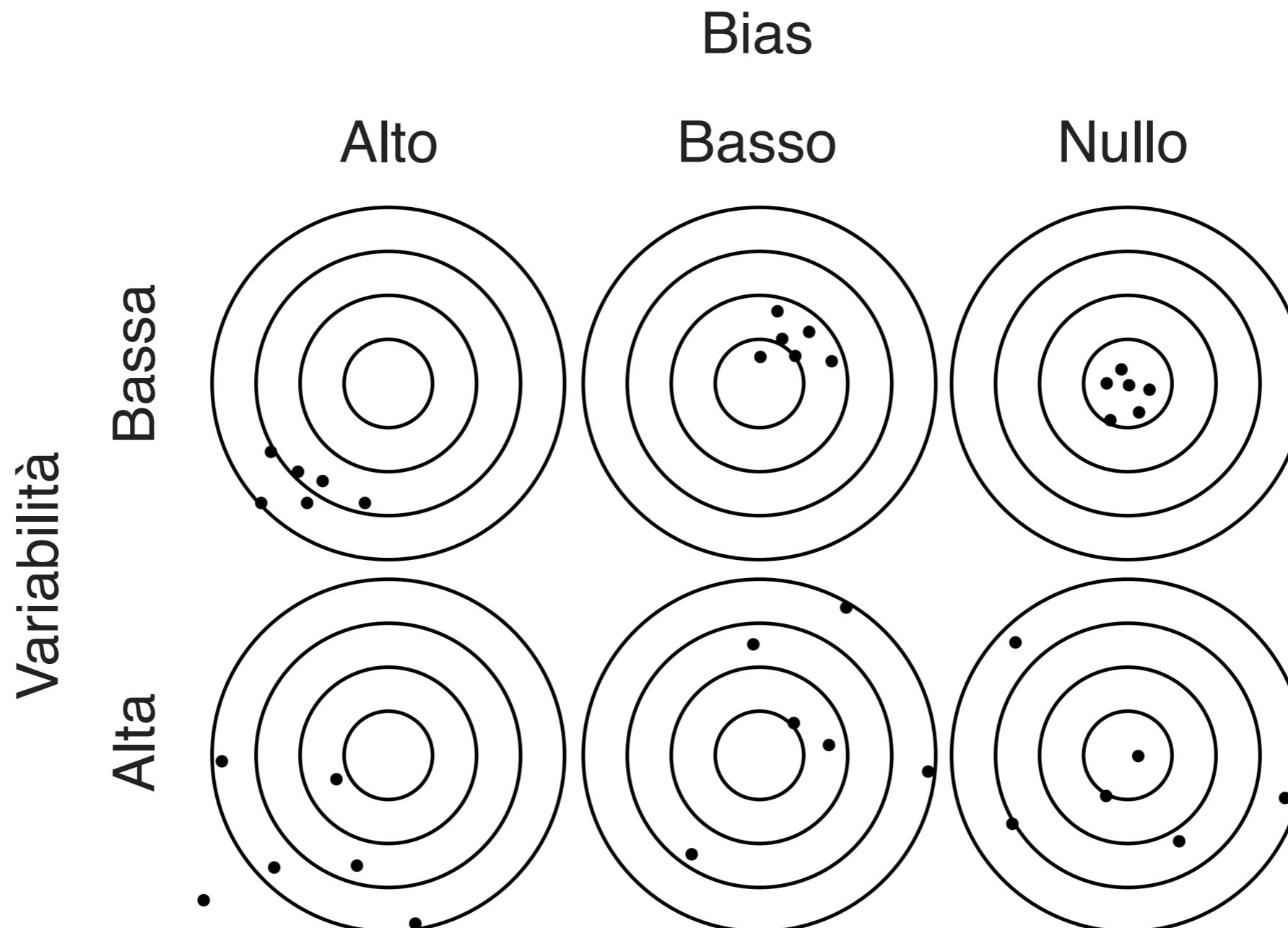
---

Workshop di ricerca sociale

# Cos'è l'errore

*Bias e variabilità*

---



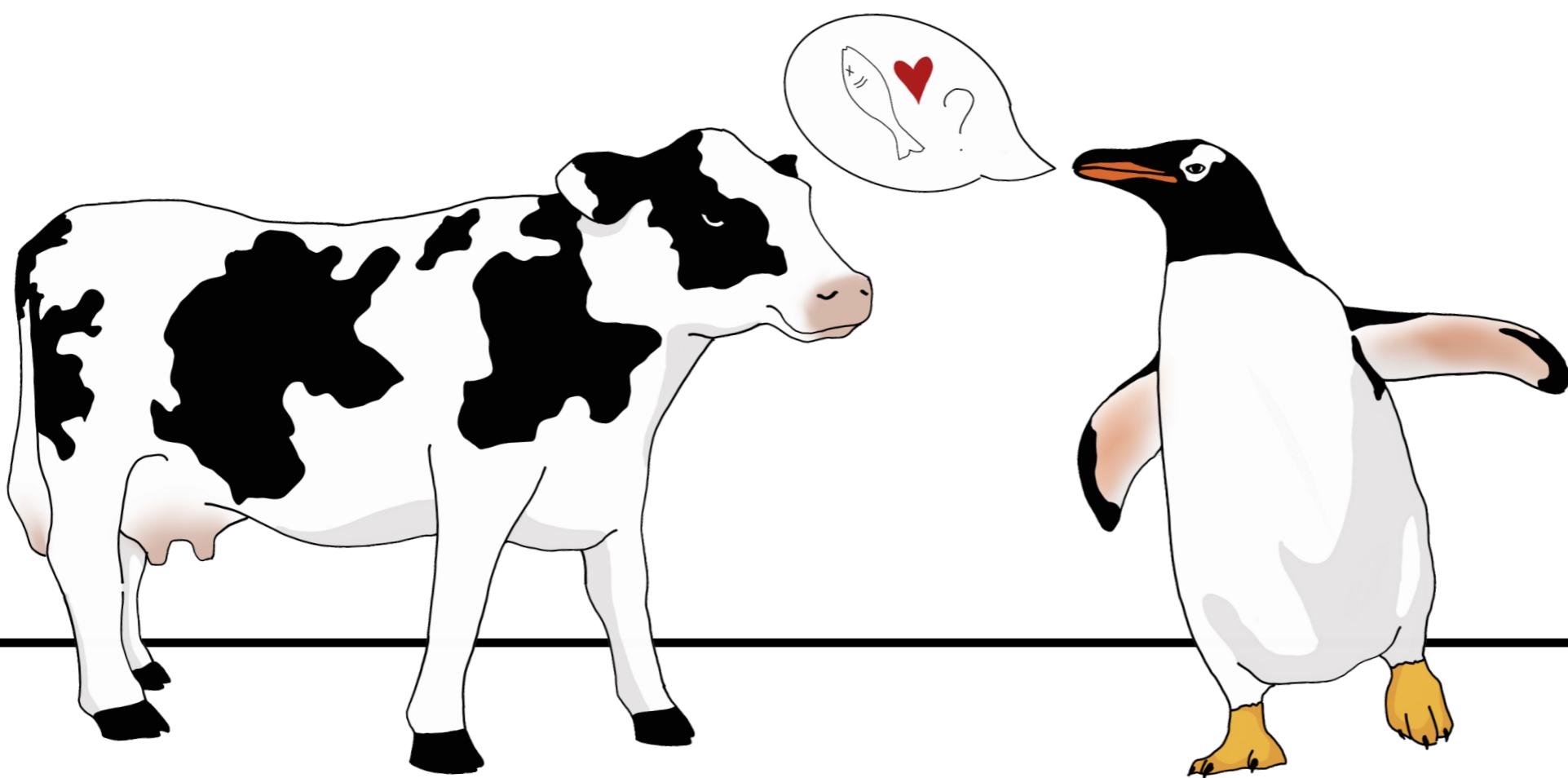
# Total survey error framework

Di quale errore parliamo

---

$$\text{Errore totale} = \text{Errore di rappresentazione} + \text{Errore di misura}$$

# Errore di rappresentazione



Workshop di ricerca sociale

# Le elezioni USA del 1936

Roosevelt contro Landon

---

Partito democratico  
Governatore di NY



Partito repubblicano  
Governatore del Kansas



# Il sondaggio del *Literary Digest*

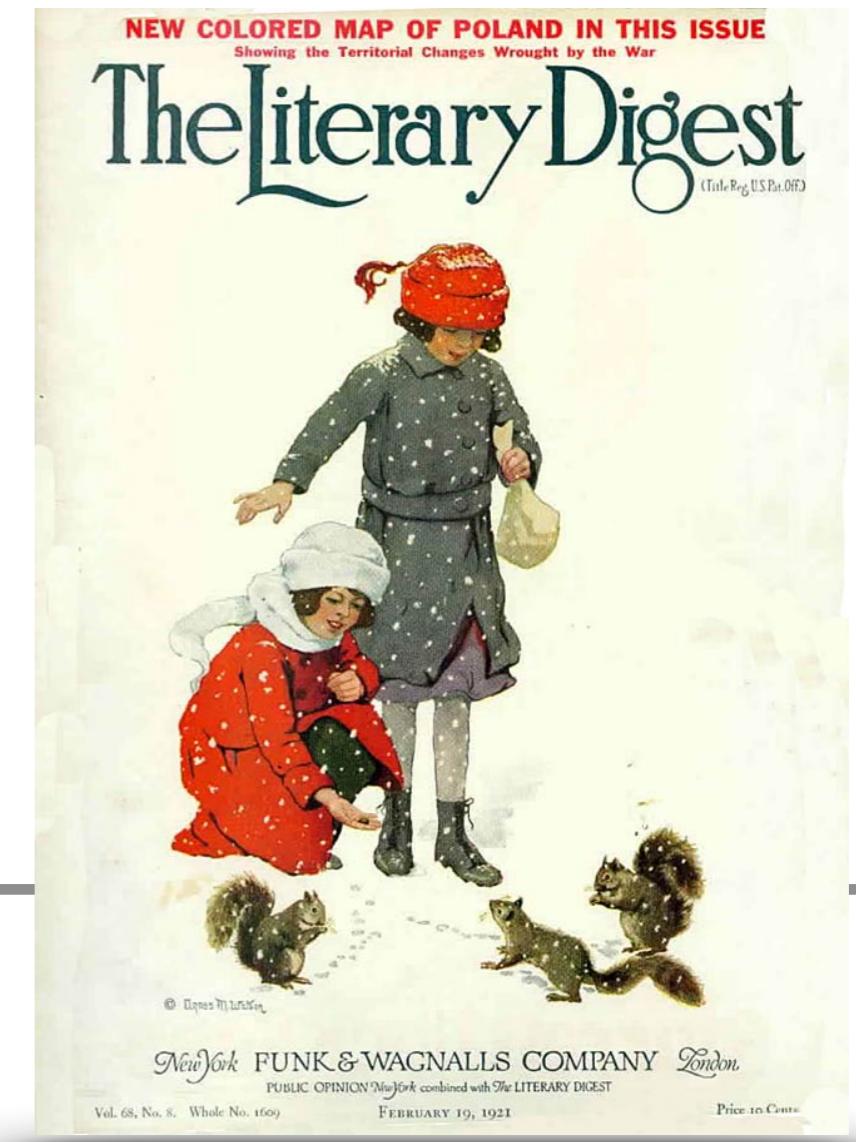
Le dimensioni non contano

---

- 10 mln di questionari
- 2,4 mln di intervistati

**1000 volte ca.** un campione contemporaneo

Predizione: Landon vincitore



# Le elezioni USA del 1936

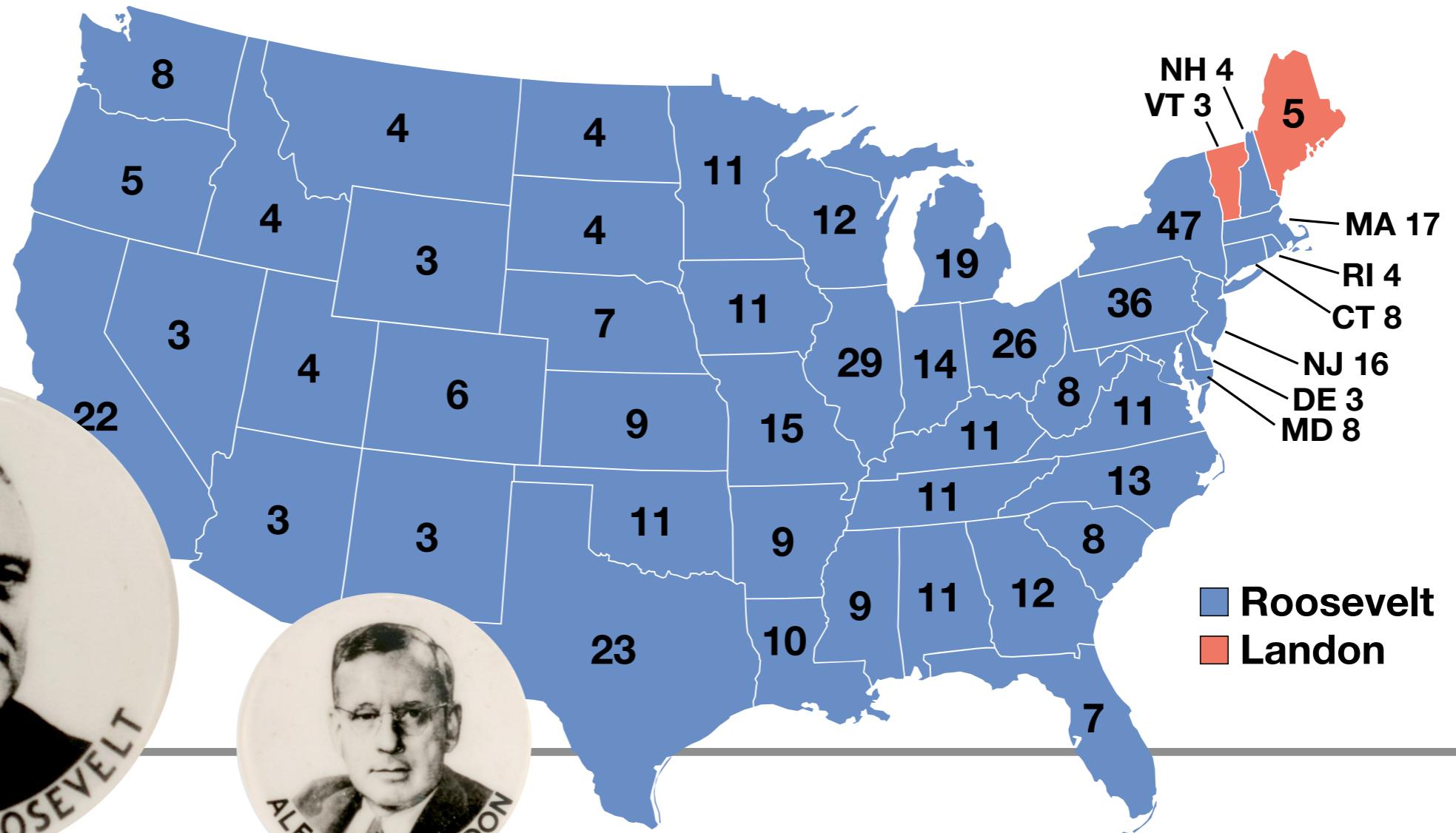
## Risultati



60,8%



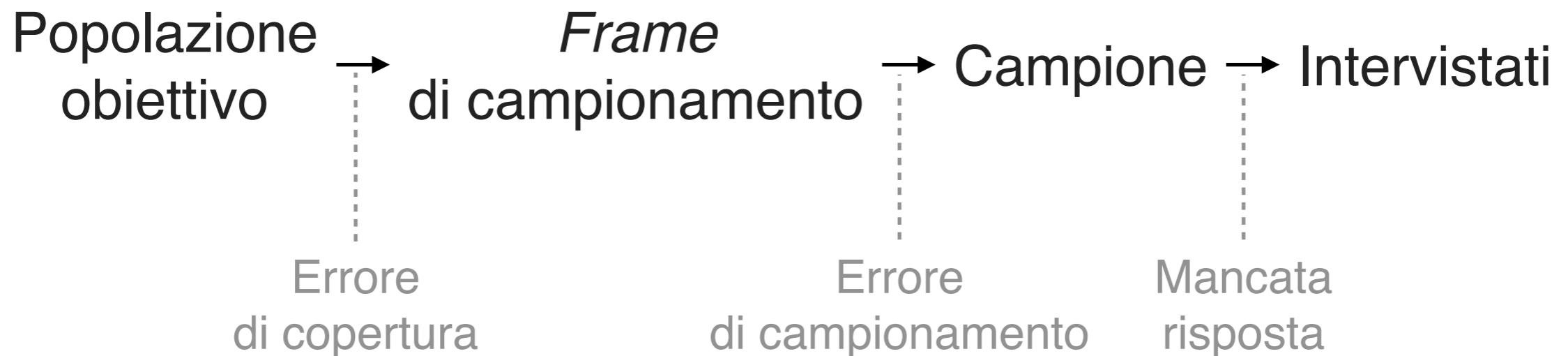
36,5%



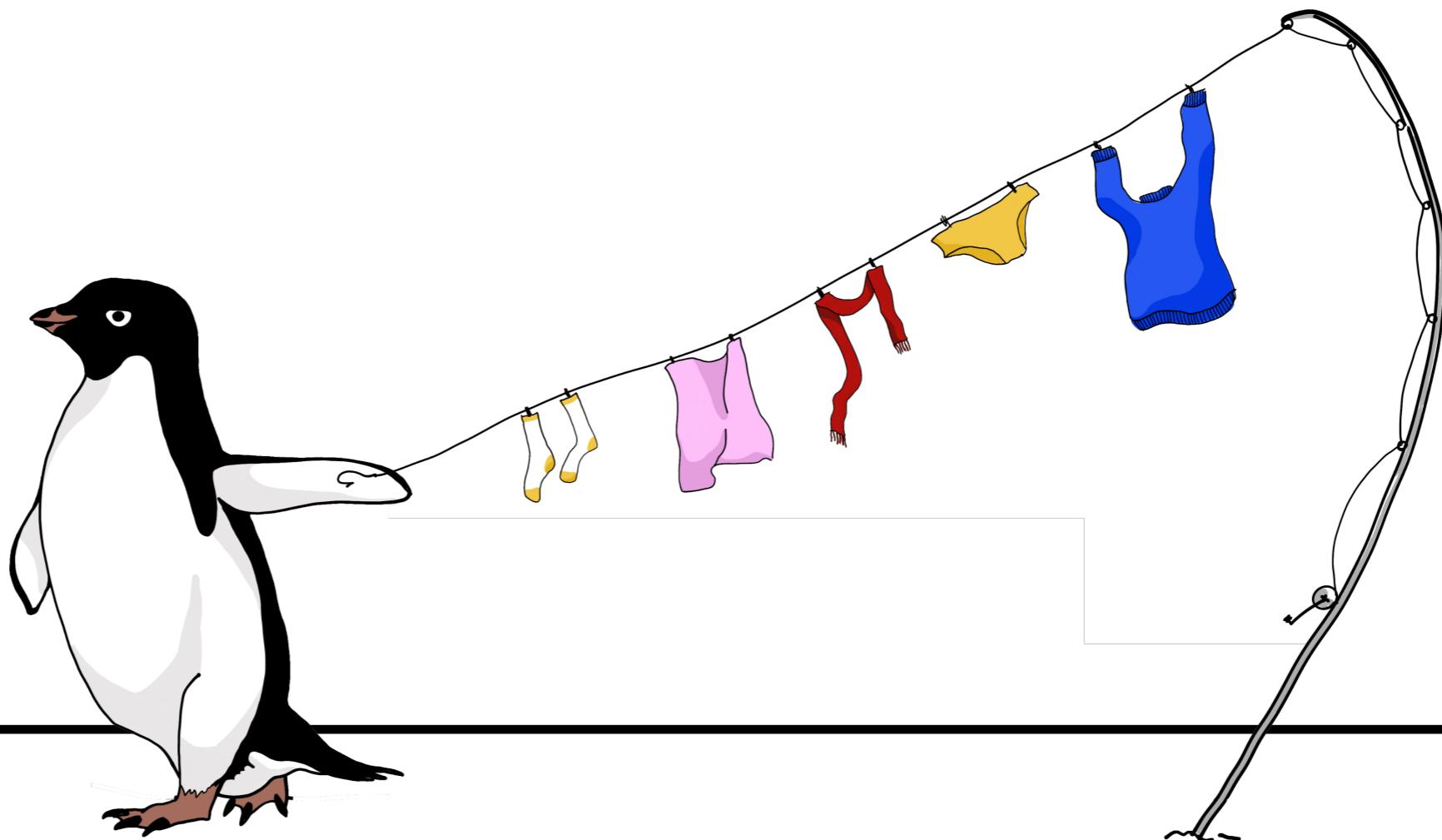
# Errori di rappresentazione

La parte per il tutto

---



# Errore di misura

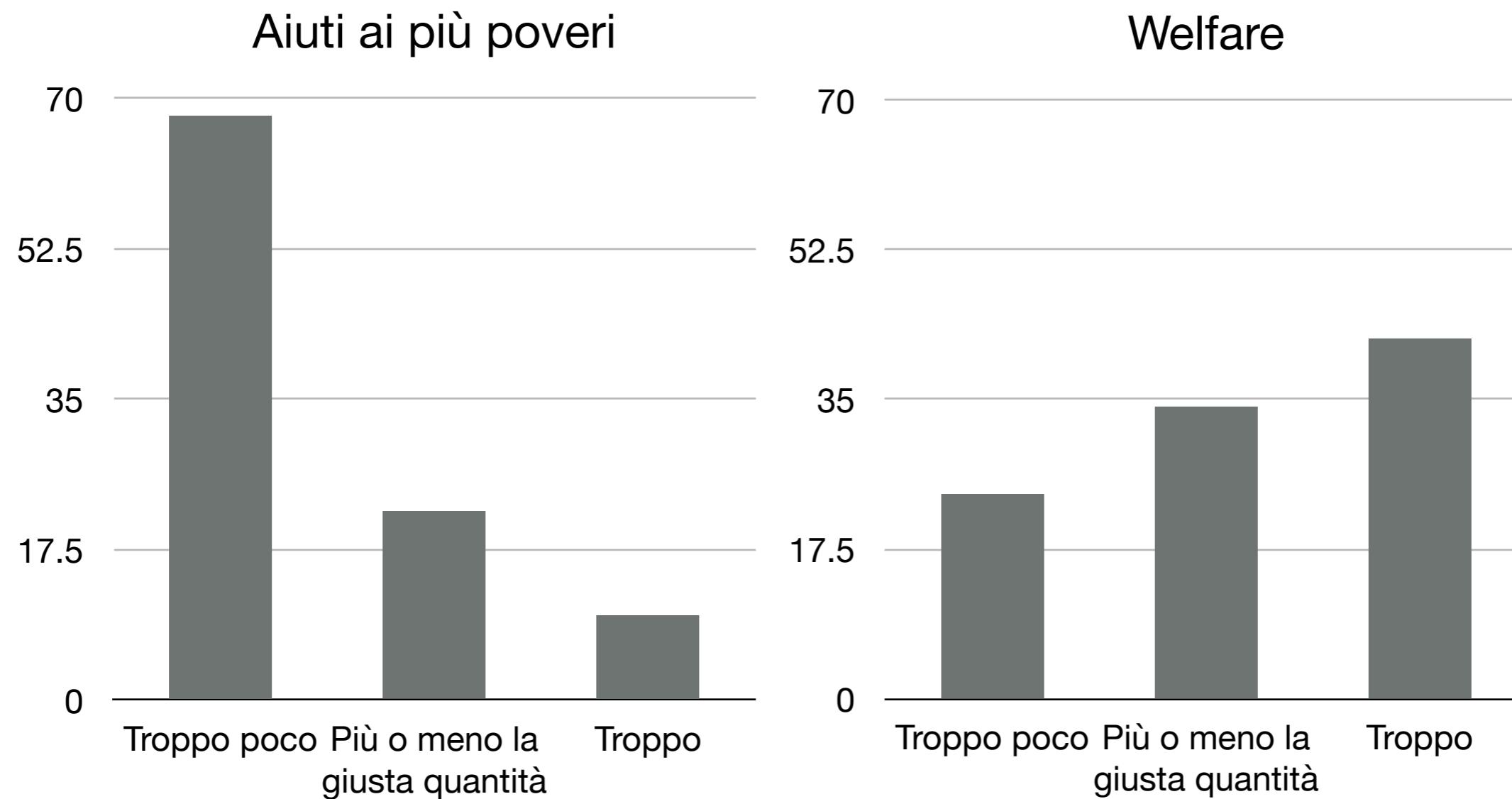


Workshop di ricerca sociale

# Spendiamo troppo o troppo poco?

Le opinioni in un sondaggio americano

---



Adattamento dalla tabella A1 di:

HUBER, G. A., & PARIS, C. (2013). ASSESSING THE PROGRAMMATIC EQUIVALENCE ASSUMPTION IN QUESTION WORDING EXPERIMENTS: UNDERSTANDING WHY AMERICANS LIKE ASSISTANCE TO THE POOR MORE THAN WELFARE. *The Public Opinion Quarterly*, 77(1), 385–397. <http://www.jstor.org/stable/24545803>

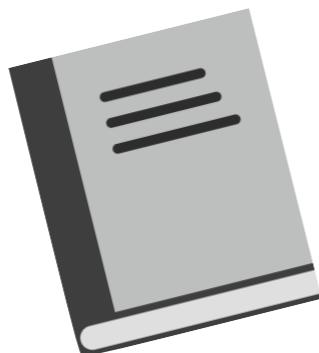
# Operazionalizzazione

Dal concetto al valore

---

Concetto      Proprietà      Operazionalizzazione      Valore

Peso



0,6 Kg

# Evitare gli errori

# Domande di ricerca

Porsi le giuste domande

---

Una domanda di ricerca dovrebbe essere:

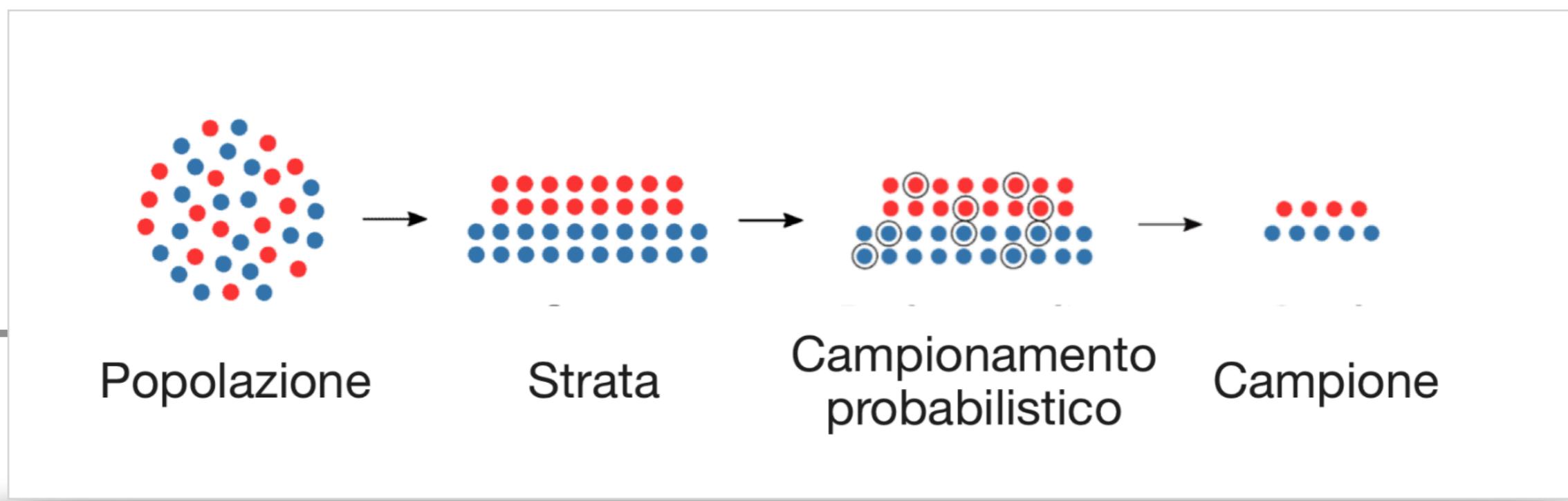
- Importante per la società
- Interessante per il ricercatore
- Fattibile ed eticamente difendibile

# Errori di rappresentazione

A chi chiedere?

---

- Campionamento probabilistico
- Post-stratificazione

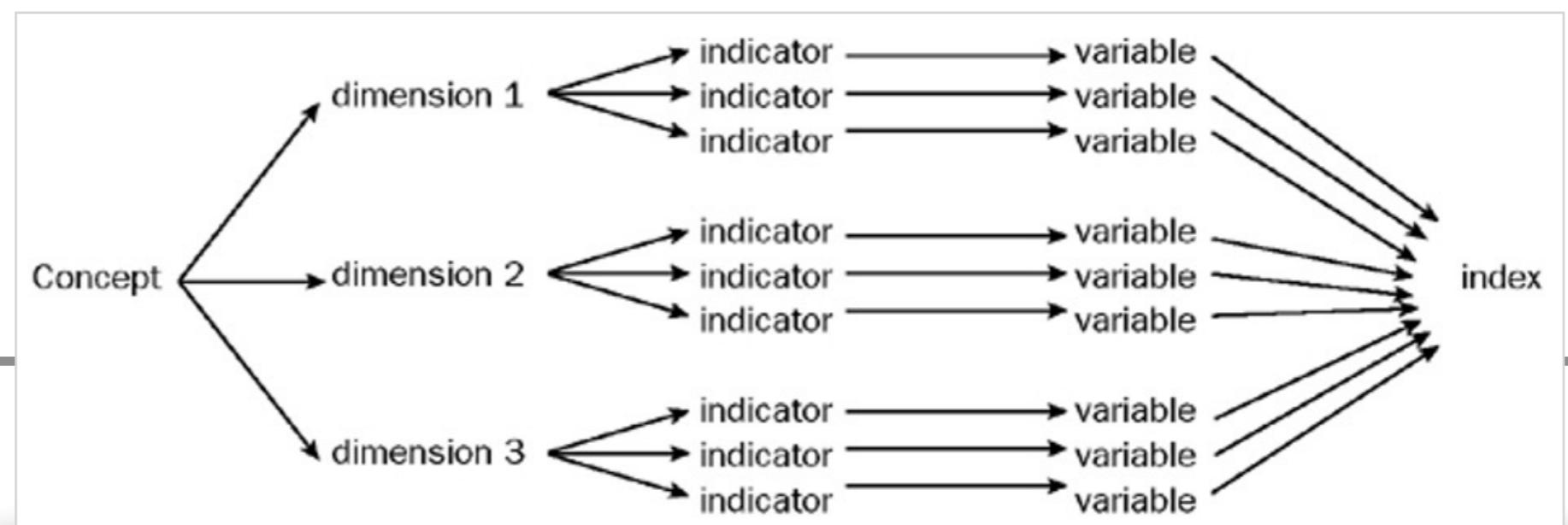


# Errori di misura

Come chiedere?

---

- Ampia **preparazione** sulla letteratura di riferimento
- Costruire indici **multidimensionali**



# Analisi dati

---

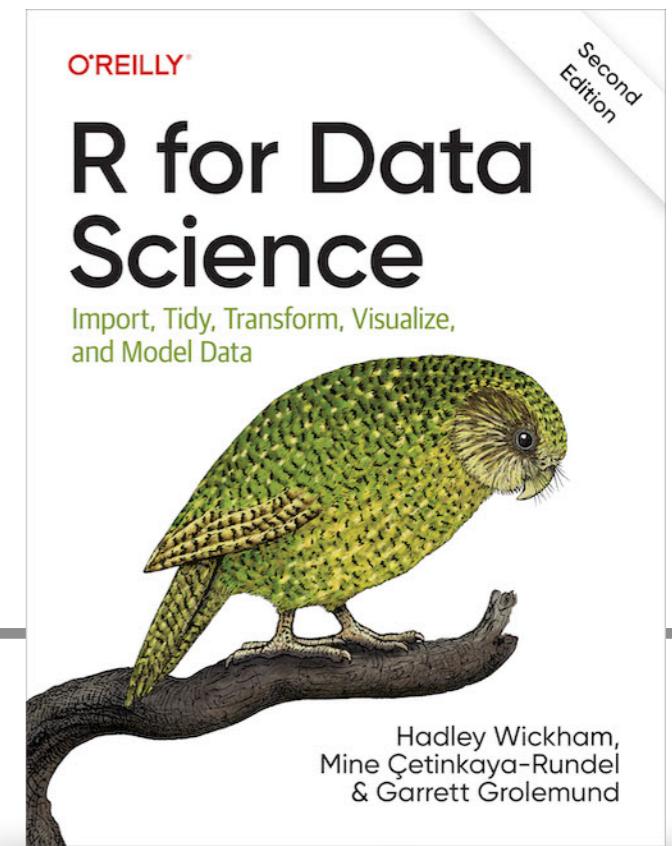
**Workshop di ricerca sociale**

# Approccio *tidy*

Un'osservazione per riga, una variabile per colonna

---

	y1	y2	y3	...	yk
x1	v11	v12	v13	...	v1k
x2	v21	v22	v23	...	v2k
x3	v31	v32	v33	...	v3k
...	...	...	...	...	...
xn	vn1	vn2	vn3	...	vnk



# Il nostro dataset

## Risultati esami californiani

Data literacy - RStudio Source Editor

CASchools

Filter

	district	school	county	grades	students	teachers	calworks	lunch	...
1	75119	Sunol Glen Unified	Alameda	KK-08	195	10.900	0.5102	2.0408	
2	61499	Manzanita Elementary	Butte	KK-08	240	11.150	15.4167	47.9167	
3	61549	Thermalito Union Elementary	Butte	KK-08	1550	82.900	55.0323	76.3226	
4	61457	Golden Feather Union Elementary	Butte	KK-08	243	14.000	36.4754	77.0492	
5	61523	Palermo Union Elementary	Butte	KK-08	1335	71.500	33.1086	78.4270	
6	62042	Burrel Union Elementary	Fresno	KK-08	137	6.400	12.3188	86.9565	
7	68536	Holt Union Elementary	San Joaquin	KK-08	195	10.000	12.9032	94.6237	
8	63834	Vineland Elementary	Kern	KK-08	888	42.500	18.8063	100.0000	
9	62331	Orange Center Elementary	Fresno	KK-08	379	19.000	32.1900	93.1398	
10	67306	Del Paso Heights Elementary	Sacramento	KK-06	2247	108.000	78.9942	87.3164	
11	65722	Le Grand Union Elementary	Merced	KK-08	446	21.000	18.6099	85.8744	
12	62174	West Fresno Elementary	Fresno	KK-08	987	47.000	71.7131	98.6056	
13	71795	Allensworth Elementary	Tulare	KK-08	103	5.000	22.4299	98.1308	
14	72181	Sunnyside Union Elementary	Tulare	KK-08	487	24.340	24.6094	77.1484	
15	72298	Woodville Elementary	Tulare	KK-08	649	36.000	14.6379	76.2712	
16	72041	Pixley Union Elementary	Tulare	KK-08	852	42.070	24.2142	94.2957	
17	63594	Lost Hills Union Elementary	Kern	KK-08	491	28.920	11.2016	97.7597	
18	63370	Buttonwillow Union Elementary	Kern	KK-08	421	25.500	8.5511	77.9097	
19	64709	Lennox Elementary	Los Angeles	KK-08	6880	303.030	21.2824	94.9712	
20	63560	Lamont Elementary	Kern	KK-08	2688	135.000	23.4375	93.2292	

Showing 1 to 20 of 420 entries, 14 total columns

# Il nostro dataset

## Risultati esami californiani

---

A data frame containing 420 observations on 14 variables.

**district** character. District code.

**school** character. School name.

**county** factor indicating county.

**grades** factor indicating grade span of district.

**students** Total enrollment.

**teachers** Number of teachers.

**calworks** Percent qualifying for CalWorks (income assistance).

**lunch** Percent qualifying for reduced-price lunch.

**computer** Number of computers.

**expenditure** Expenditure per student.

**income** District average income (in USD 1,000).

**english** Percent of English learners.

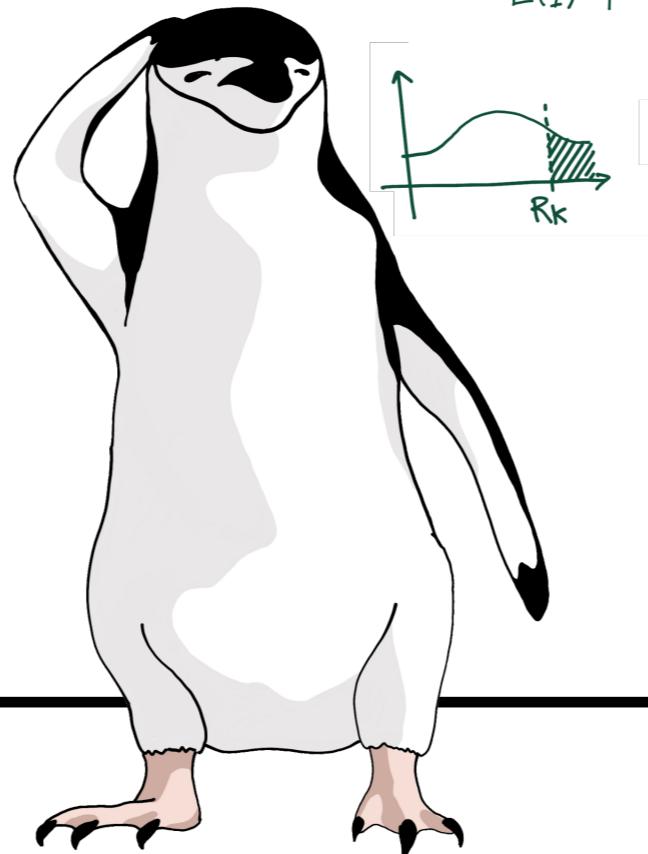
**read** Average reading score.

**math** Average math score.

# Analisi esplorativa

$$P\{A_n|B\} = \frac{P(B|A_n)P(A_n)}{\sum_i P(B|A_i)P(A_i)}$$

$$E(Y) = \mu = g^{-1}(\mathbf{x}\beta)$$

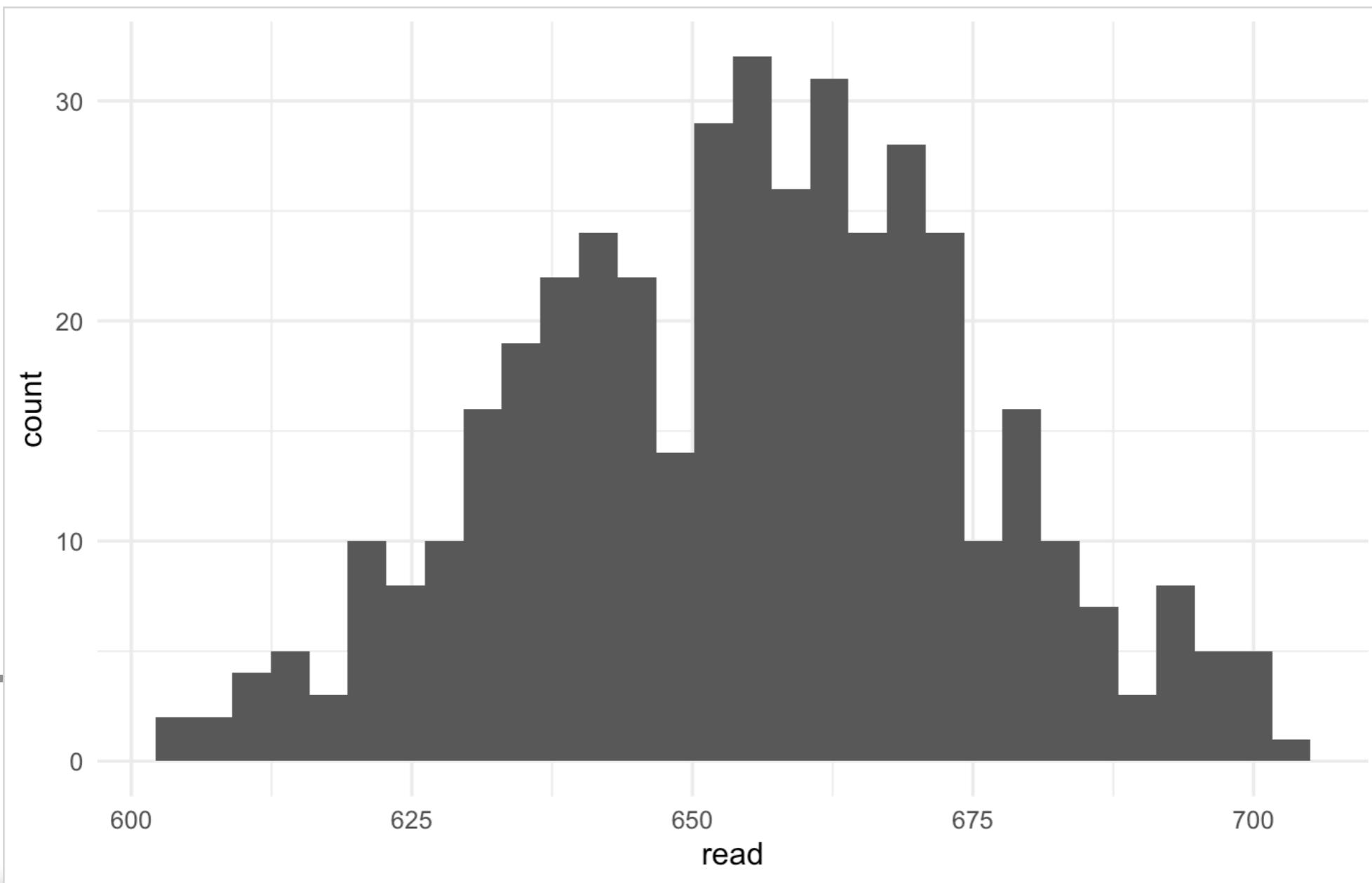


Workshop di ricerca sociale

# Analisi esplorativa

Istogramma della variabile obiettivo

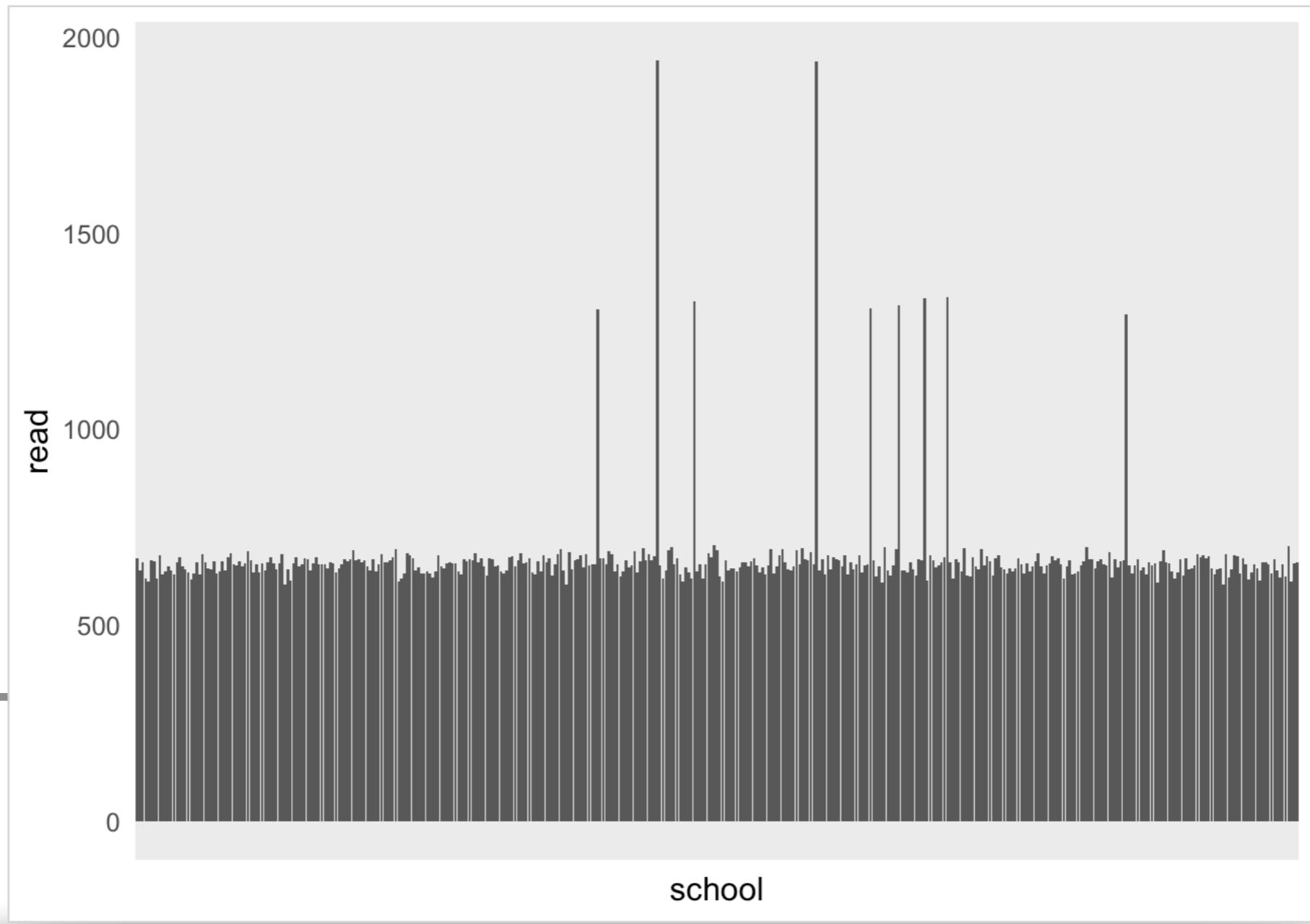
---



# Analisi esplorativa

Plot delle differenze fra scuole

---



RQ: Cosa spiega le differenze tra i voti conseguiti da studenti di scuole diverse?

# Esempi di ipotesi possibili

Avete consigli?

---

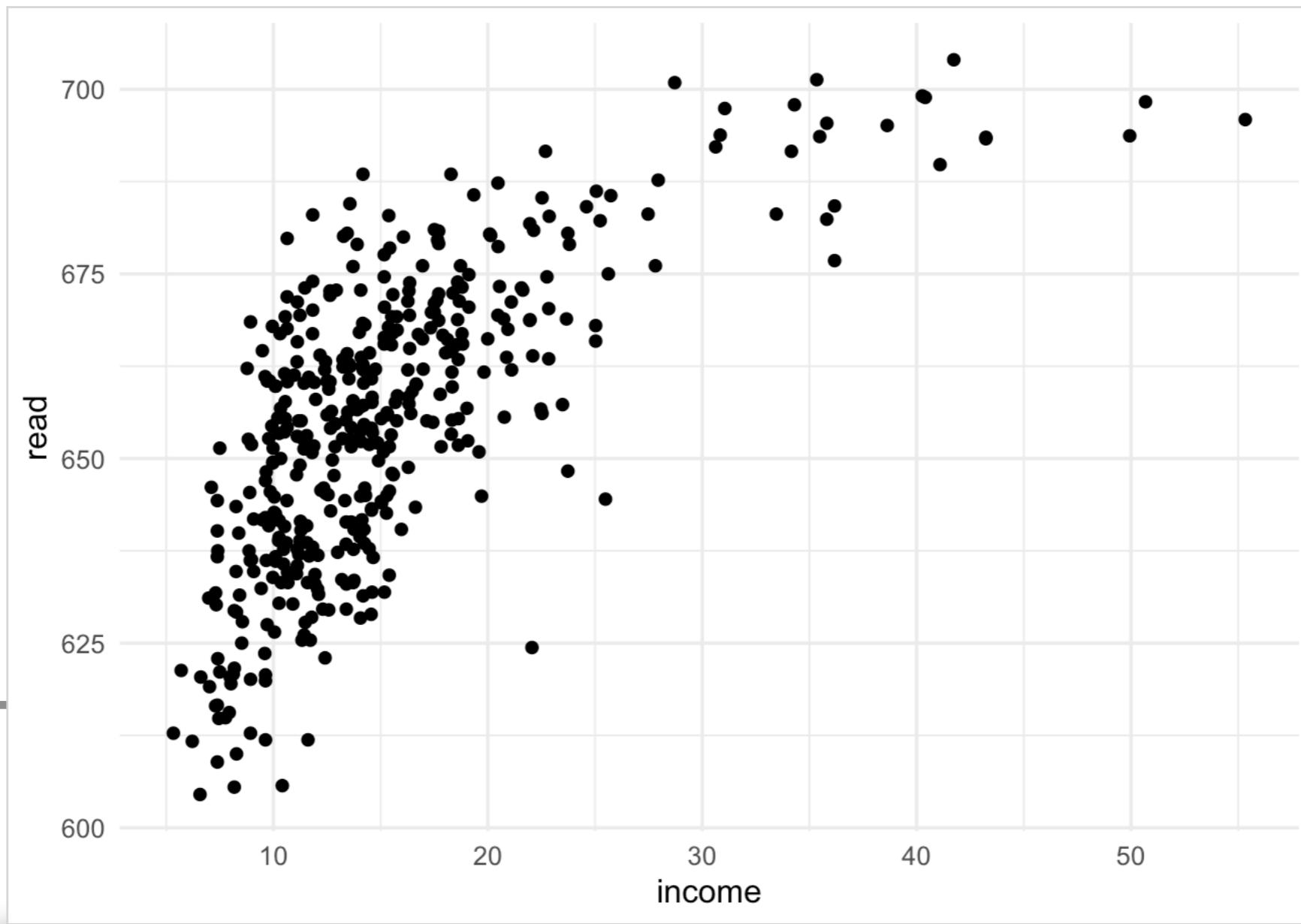
H1: Le scuole con reddito medio più alto sono avvantaggiate

H2: Le scuole con meno studenti non americani sono avvantaggiate

# Reddito e rendimento

Ipotesi 1

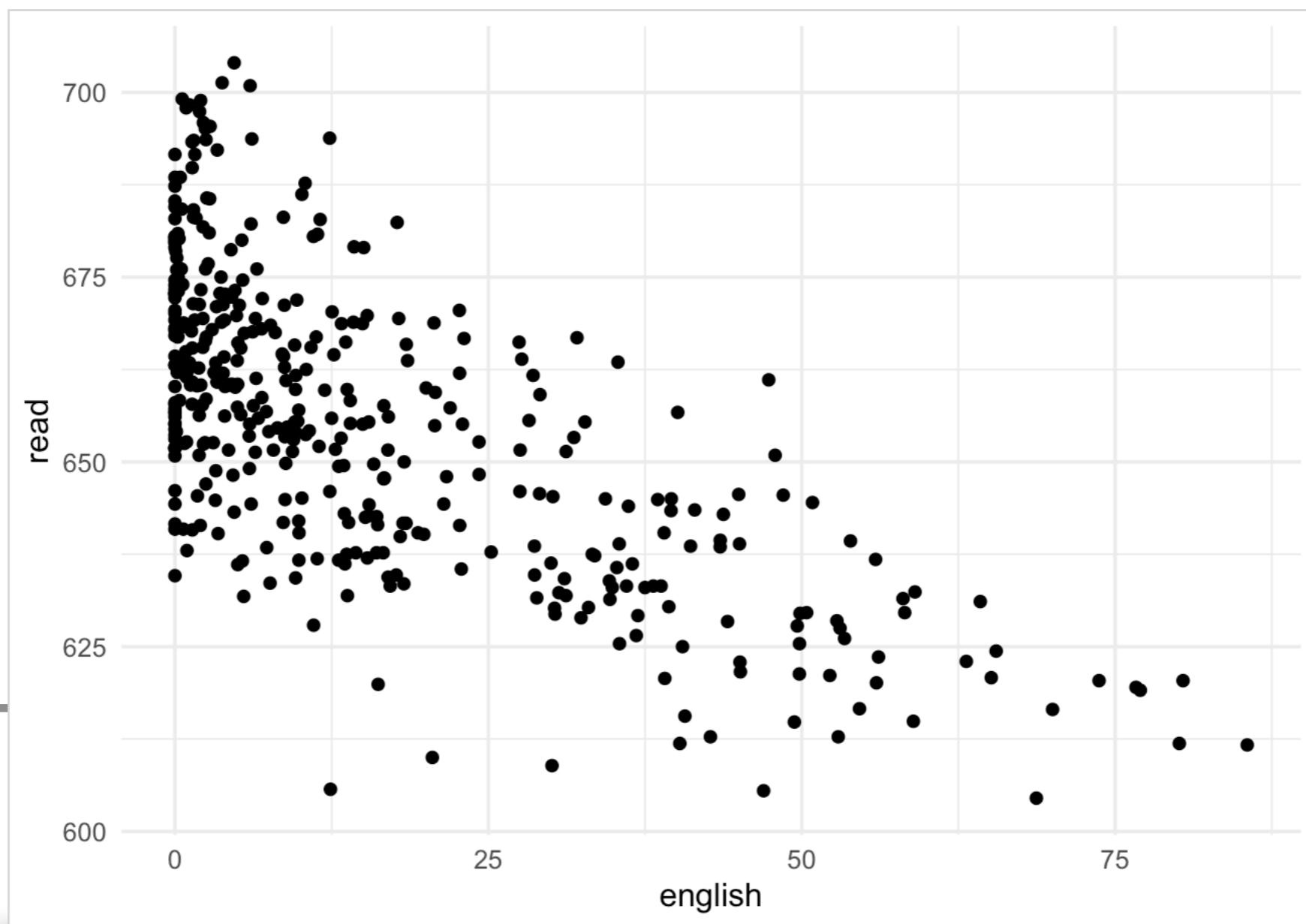
---



# Quanto è alta la barriera linguistica

Ipotesi 2

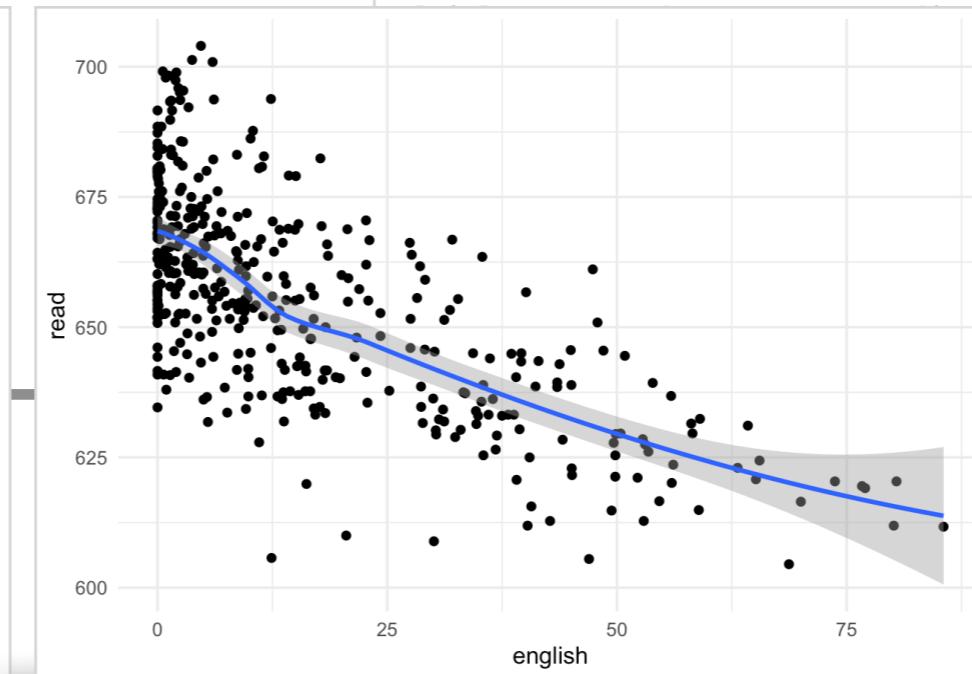
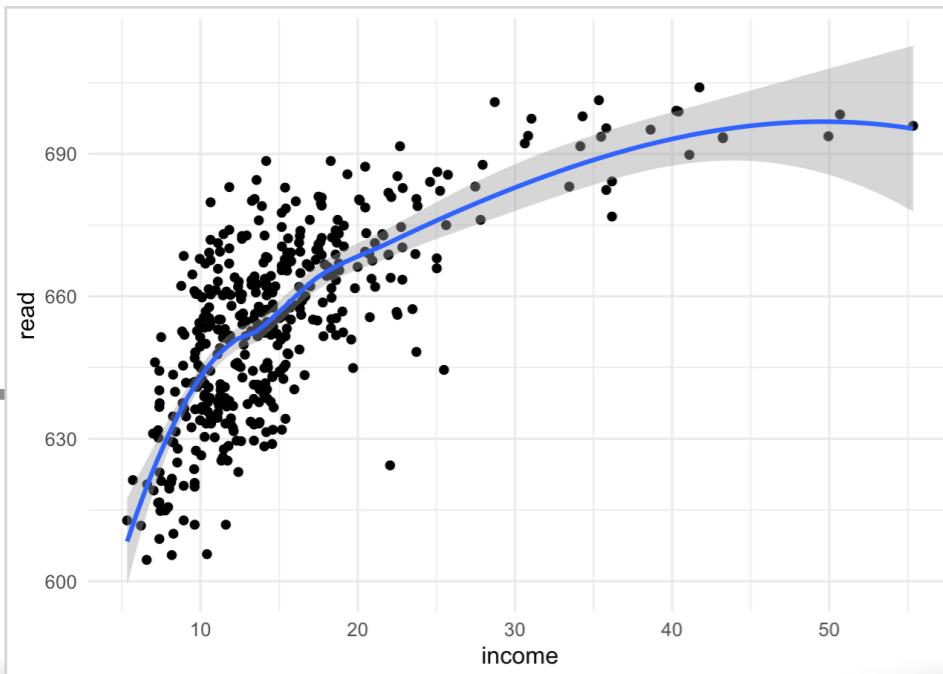
---



# Modellizzazione statistica

## Quantificare la diseguaglianza

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2^2 + \epsilon$$



```
Call:  
lm(formula = read ~ english + income2, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-46.676	-7.205	0.774	7.518	31.511

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	657.424772	0.937128	701.53	<2e-16 ***
english	-0.639549	0.031292	-20.44	<2e-16 ***
income2	0.026615	0.001629	16.34	<2e-16 ***
---				
Signif. codes:	0	***	0.001	**
	0.01	*	0.05	.
	0.1	'	1	

Residual standard error: 11.39 on 417 degrees of freedom

Adjusted R-squared: 0.6792  
F-value: < 2.2e-16

# Comunicare i risultati

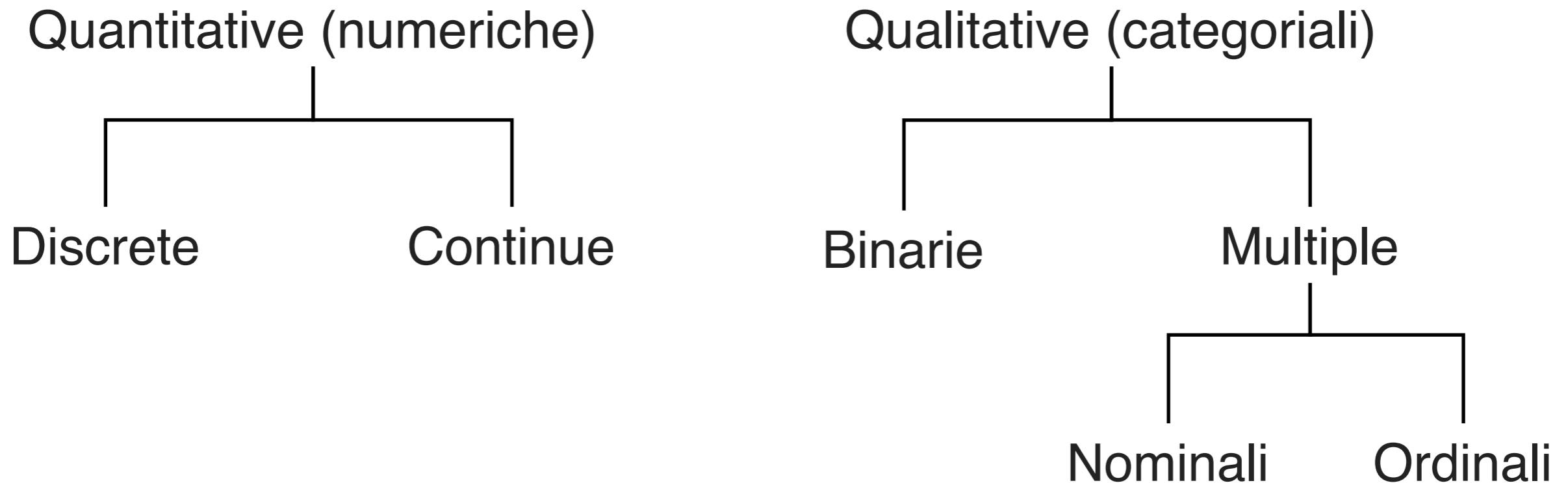
---

**Workshop di ricerca sociale**

# Scegliere il grafico giusto

## Classificazione delle variabili

---





# from Data to Viz

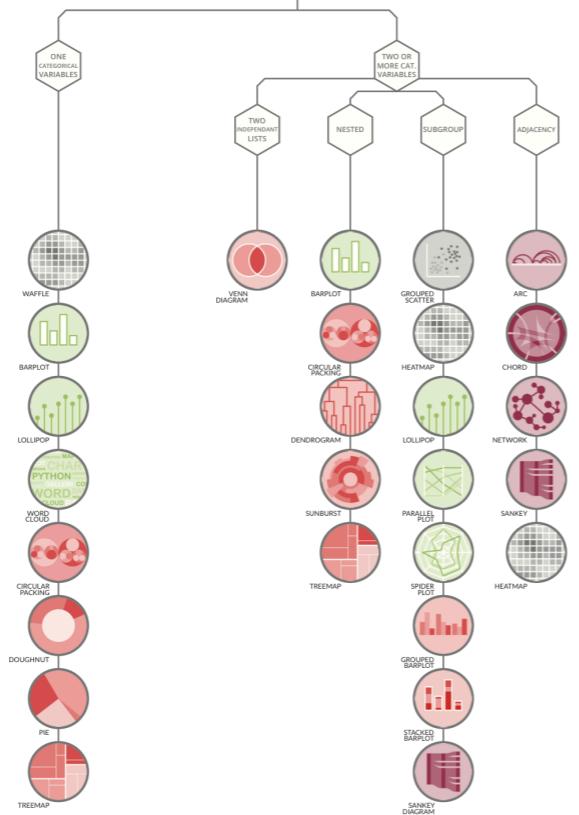
'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

- 1 Identify what type of data you have.
- 2 Go to the corresponding decision tree and follow it down to a set of possible charts.
- 3 Choose the chart from the set that will suit your data and your needs best.

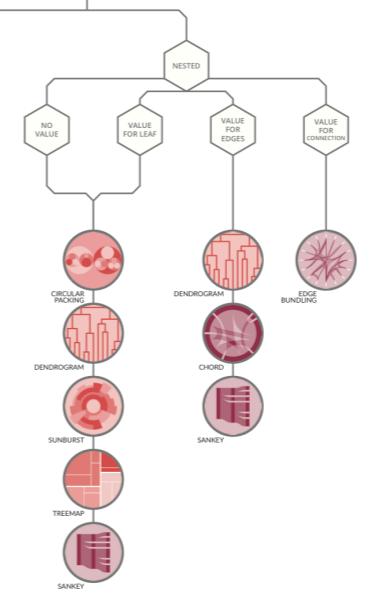
Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

[data-to-viz.com](http://data-to-viz.com)

## CATEGORIC



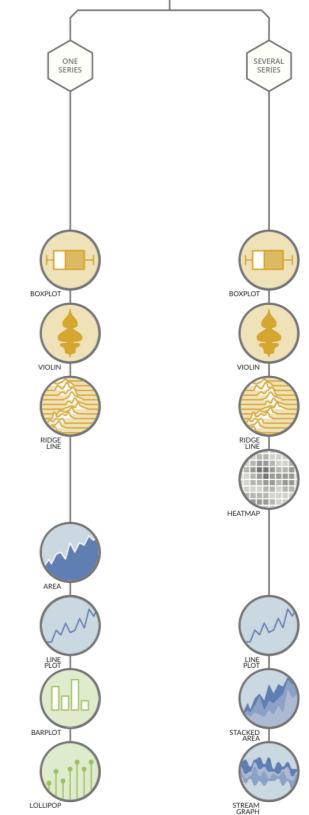
## RELATIONAL



## MAP



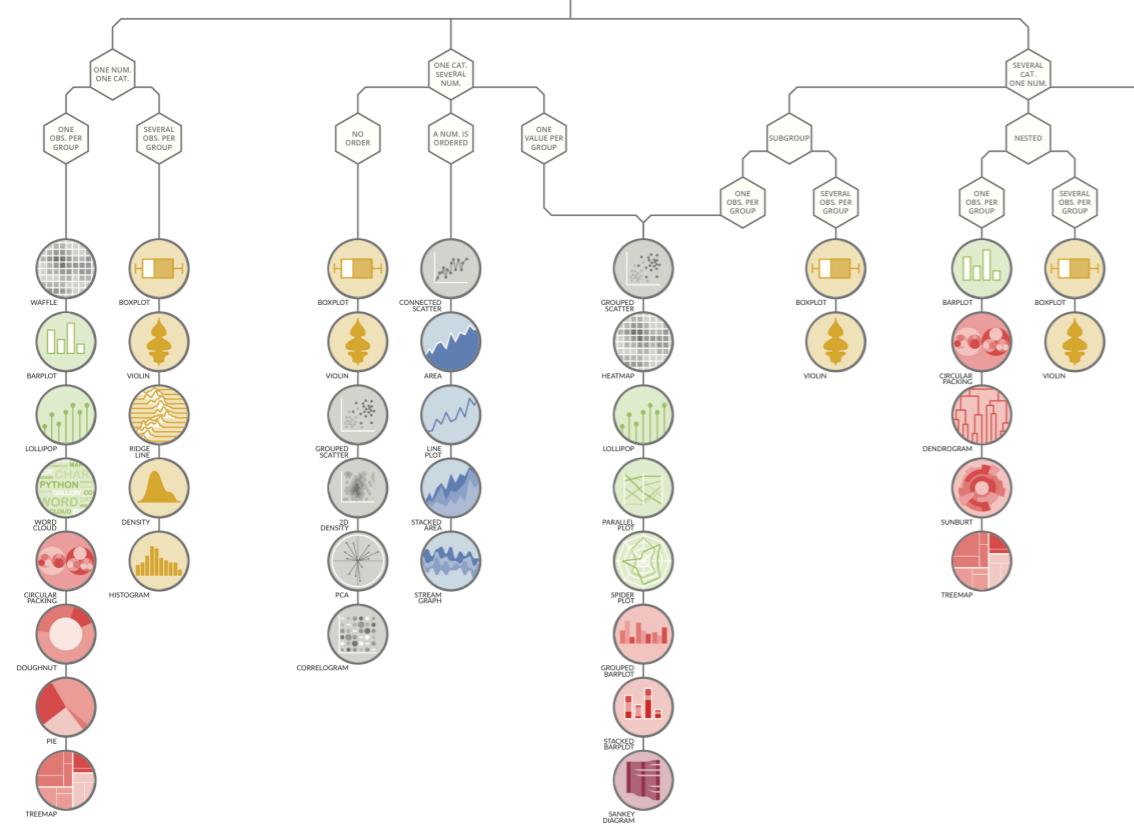
## TIME SERIES



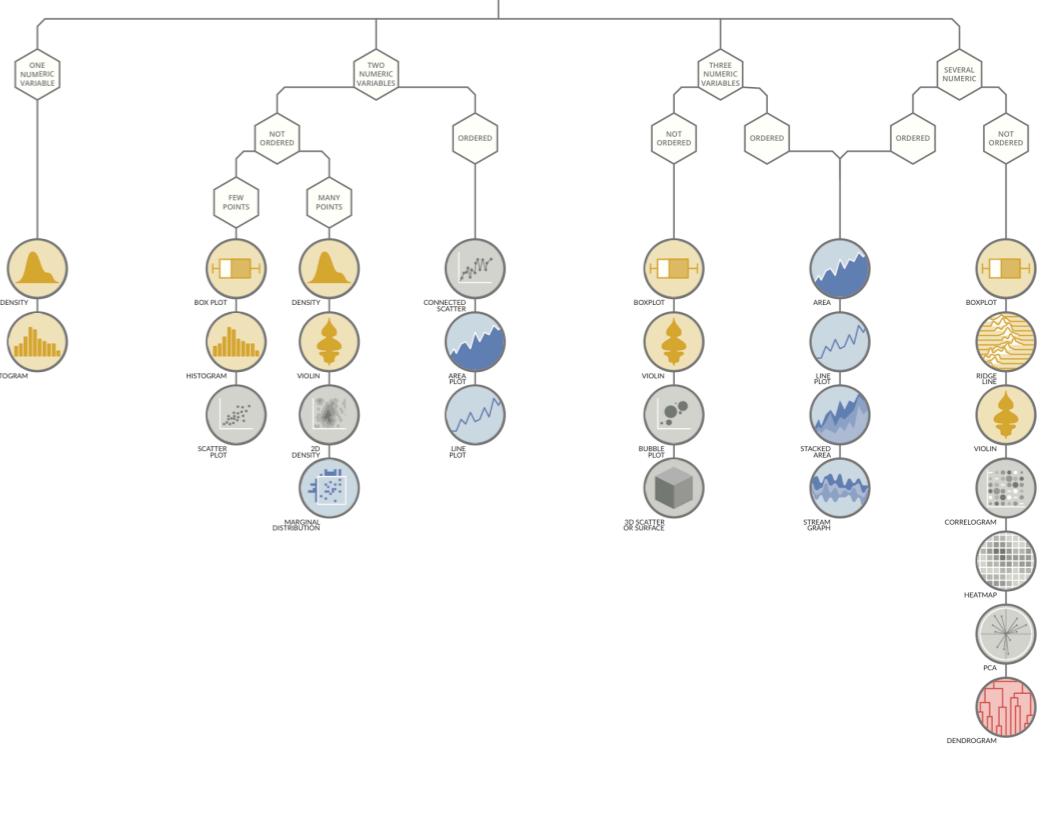
### WHAT DO YOU WANT TO SHOW ?

- |                                       |                 |                                       |           |
|---------------------------------------|-----------------|---------------------------------------|-----------|
| <span style="color: yellow;">●</span> | Distribution    | <span style="color: blue;">●</span>   | Evolution |
| <span style="color: grey;">●</span>   | Correlation     | <span style="color: purple;">●</span> | Maps      |
| <span style="color: green;">●</span>  | Ranking         | <span style="color: red;">●</span>    | Flow      |
| <span style="color: red;">●</span>    | Part of a whole |                                       |           |

## CATEGORIC AND NUMERIC



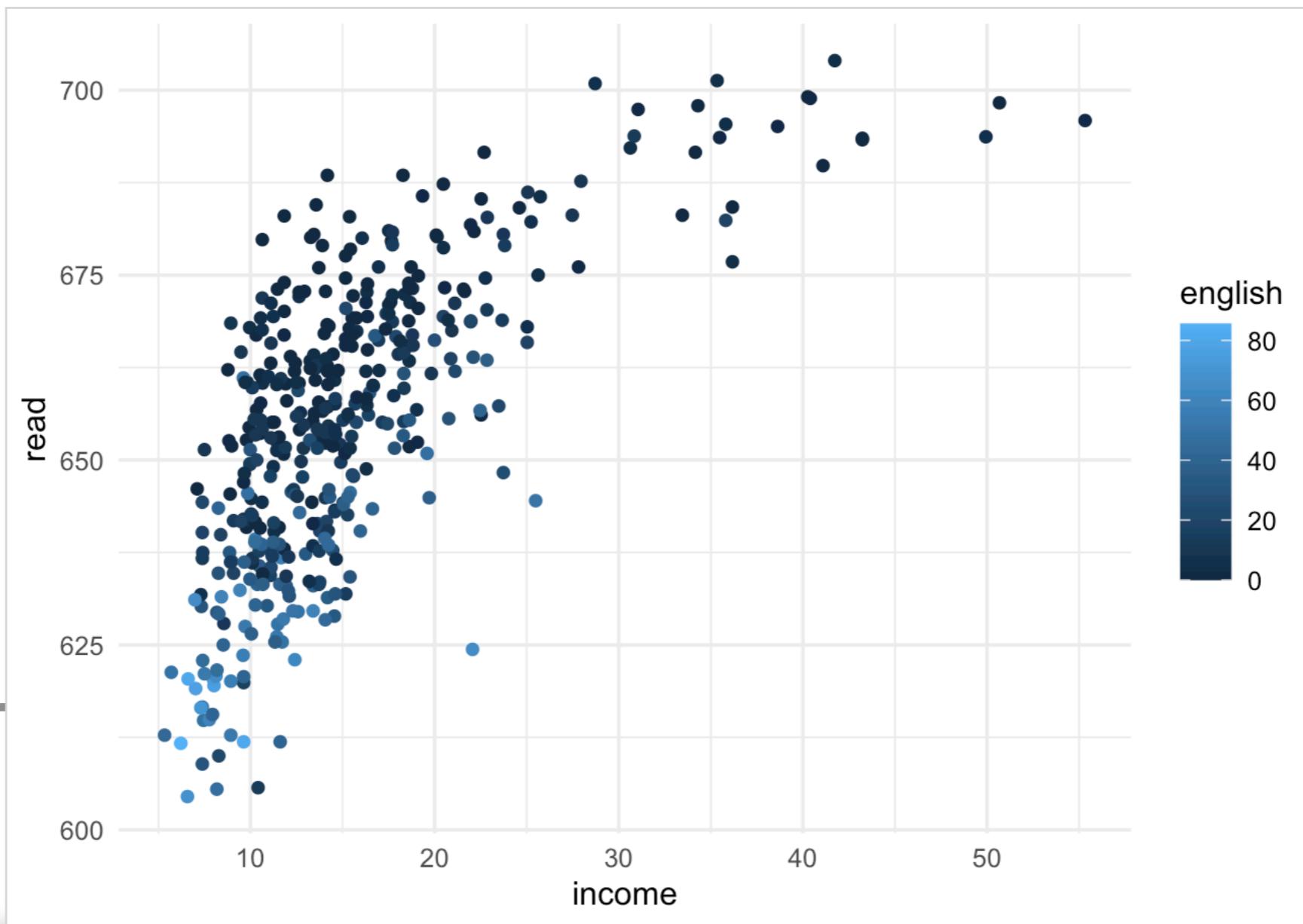
## NUMERIC



# Uscire dalla bidimensionalità

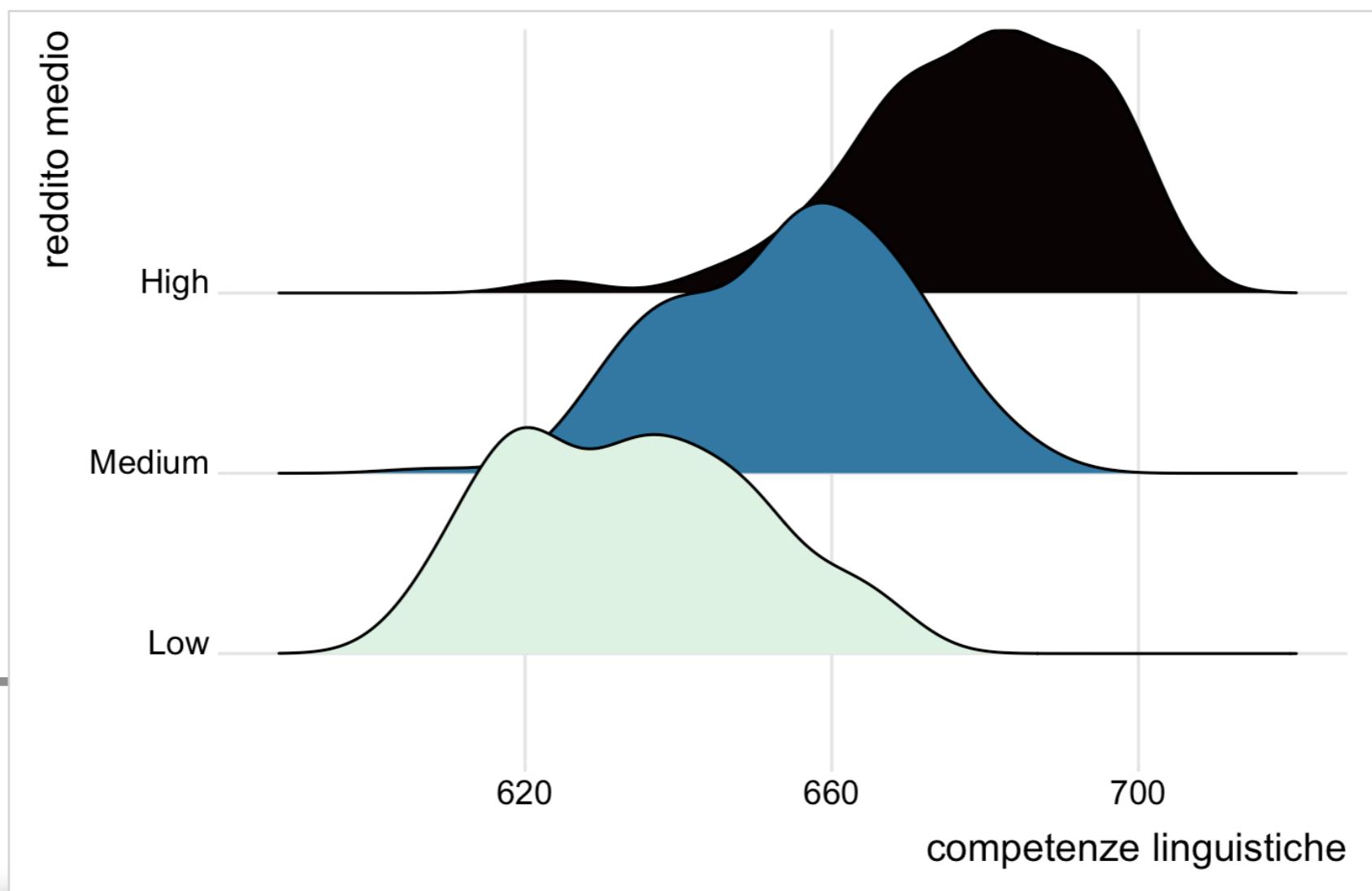
Colore a conferma della covarianza

---



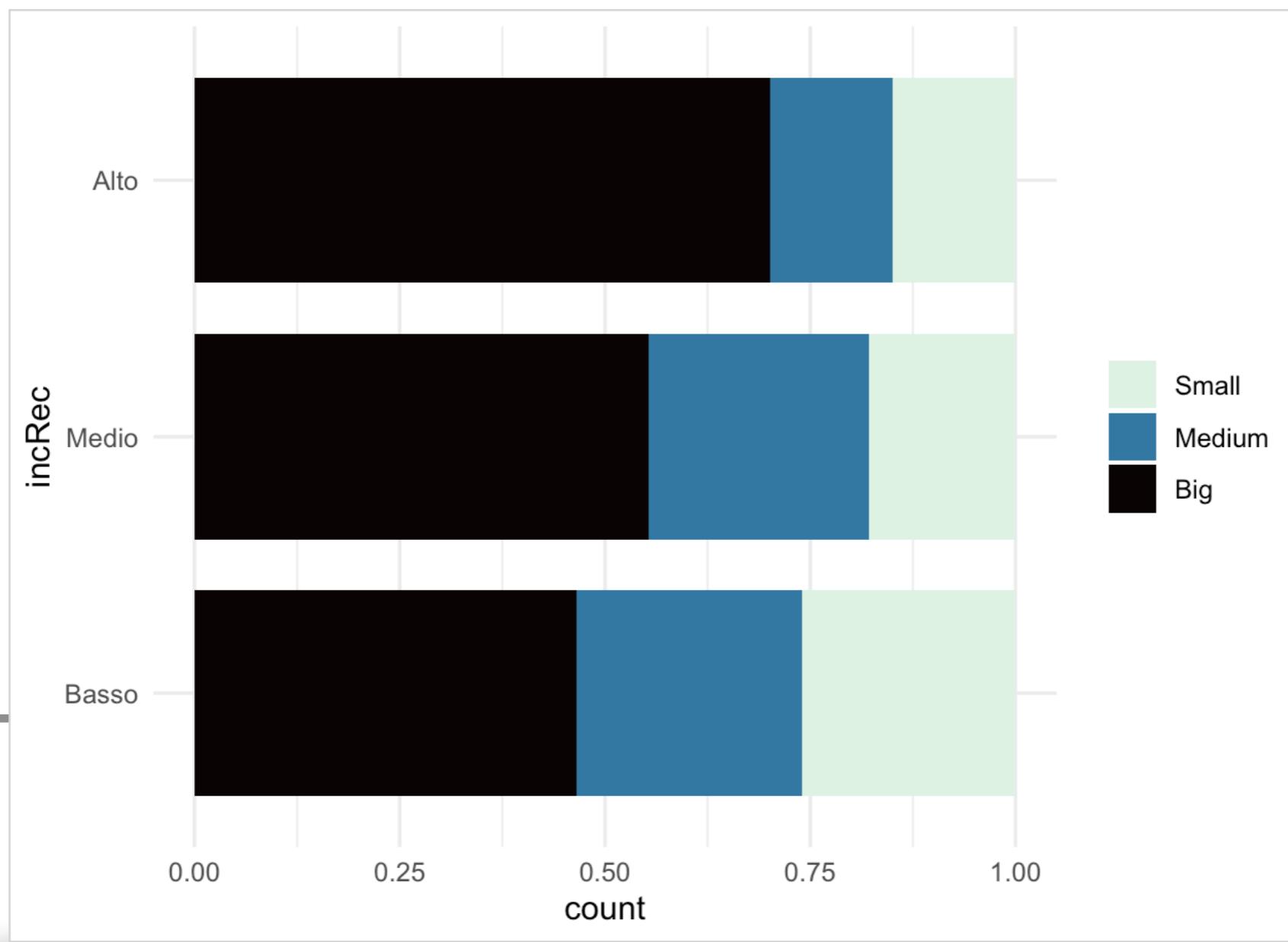
# Non solo numeri

Interazioni fra variabili numeriche e categoriali



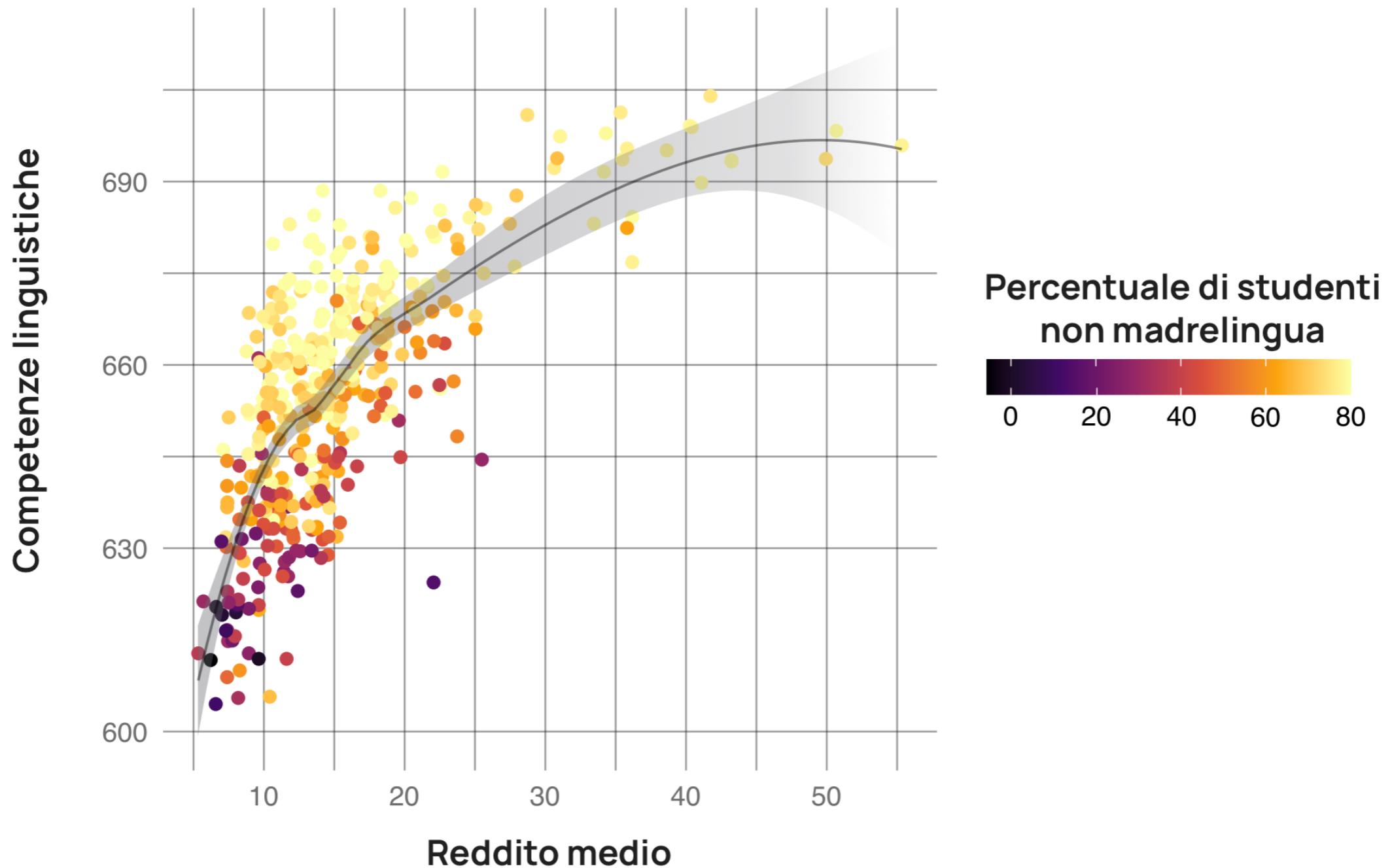
# Dare un nome alle cose

## Variabili qualitative

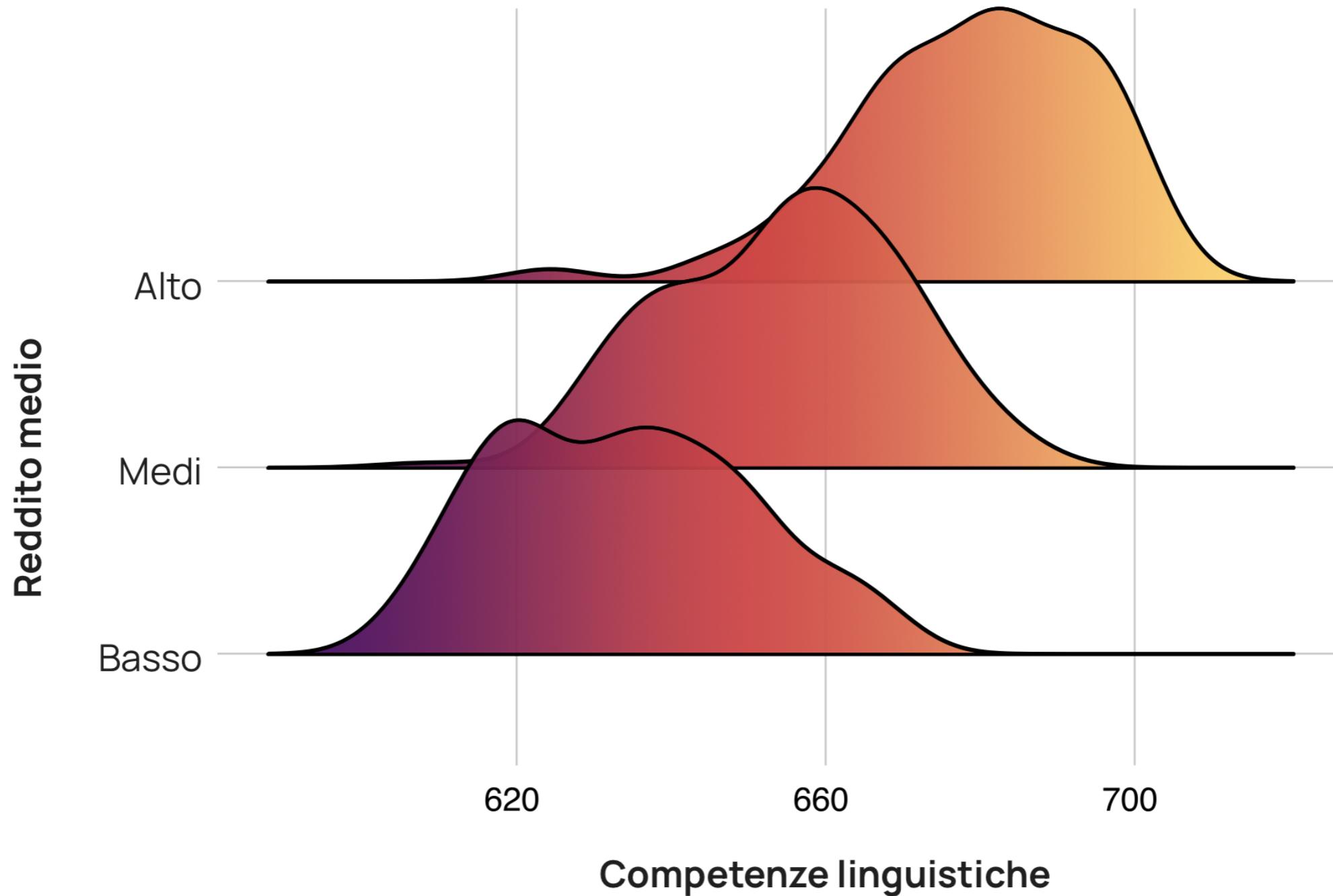


**Possiamo fare di meglio**

# Reddito medio e prima lingua degli studenti spiegano la variabilità di competenze linguistiche fra scuole

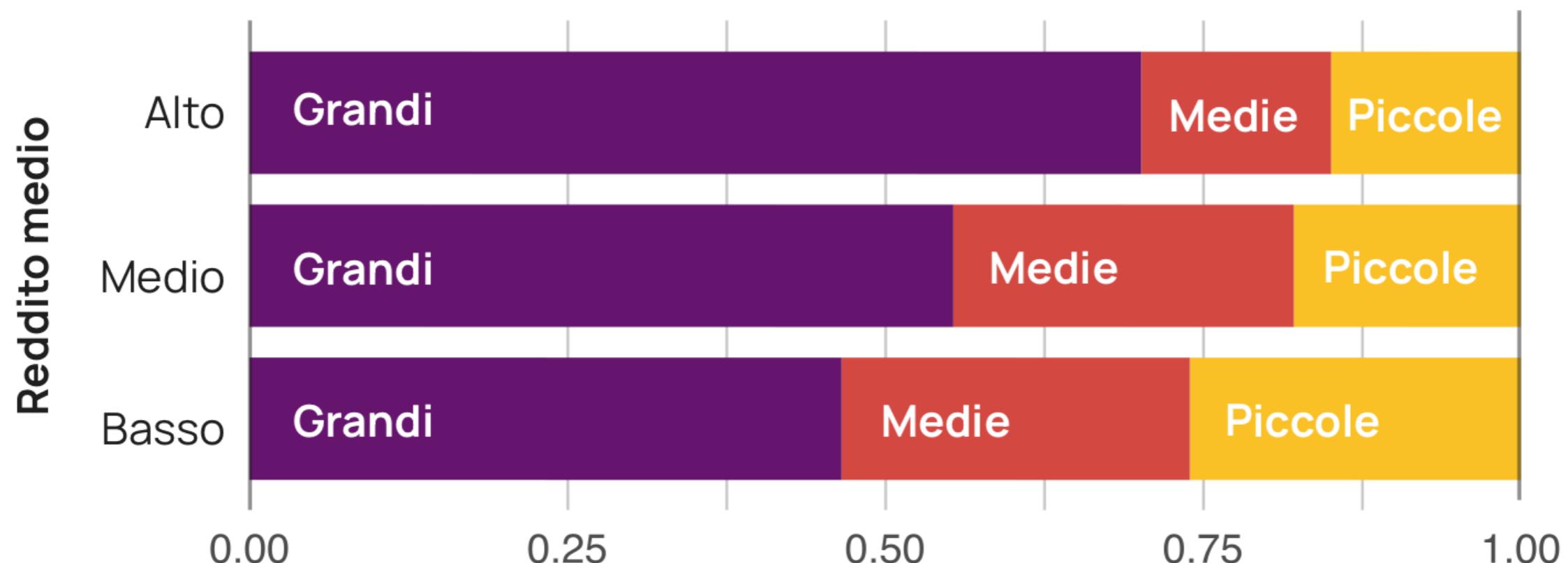


**Gli studenti di scuole ad alto reddito sono avvantaggiati nelle competenze linguistiche**



# Le scuole a reddito alto sono mediamente più grandi

Valori percentuali sul totale per categoria di reddito



dataviz-inspiration.com

About Related Subscribe

# Dataviz Inspiration

[GitHub](#) [Twitter](#) [LinkedIn](#) [Home](#)

Dataviz-inspiration.com aims at being the biggest list of chart examples available on the web. It showcases 195 of the most beautiful and impactful dataviz projects I know. The collection is a good place to visit when you're designing a new graph, together with [data-to-viz.com](#) that shares dataviz best practices.

select chart types select tools X

New Covid-19 cases, United States

Jan. 2022 →

0 150K cases

7-day average

2021 →

2020 →

Oct. April

India

Last Updated on 01 Nov, 11:20 AM IST

Kerala 79,266 Active

Tested 60,92,01,294 As of 31 October per source

GUATEMALA

COLOMBIA

EL SALVADOR

PERU

NEPAL

ROMANIA

EGYPT

SRI LANKA

AFGHANISTAN

KAZAKHSTAN

VIETNAM

PAKISTAN

INDONESIA

MYANMAR

MOROCCO

# Cosa abbiamo fatto oggi

---

## Quando i dati sbagliano

- Errore di rappresentazione
- Errore di misura

## Analisi dati

- Esempio di analisi esplorativa
- Comunicare i risultati

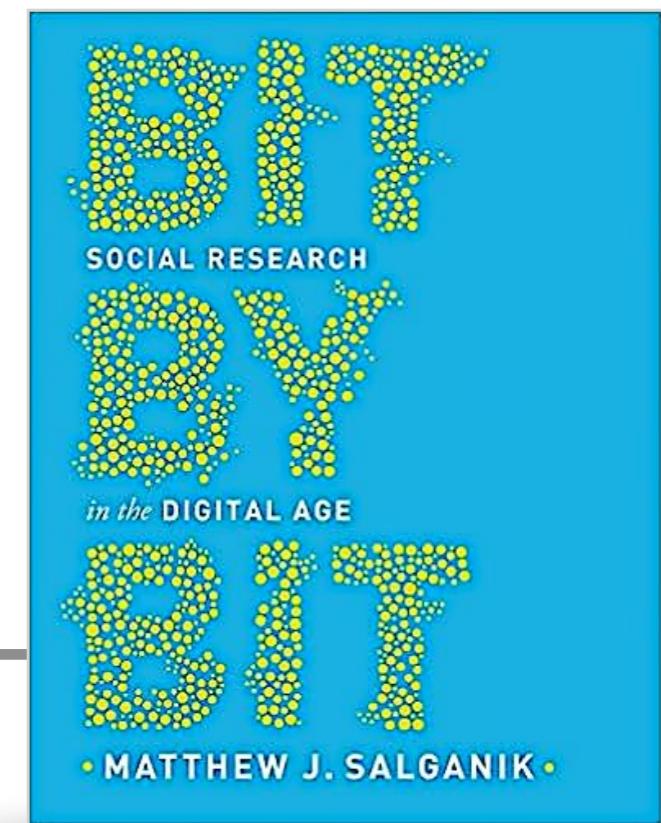
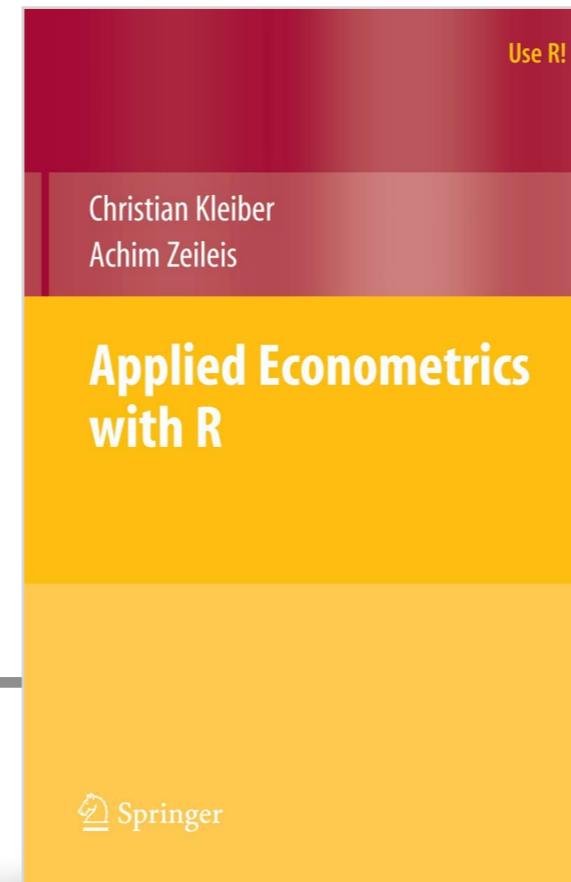
# Cosa leggere

---

Workshop di ricerca sociale

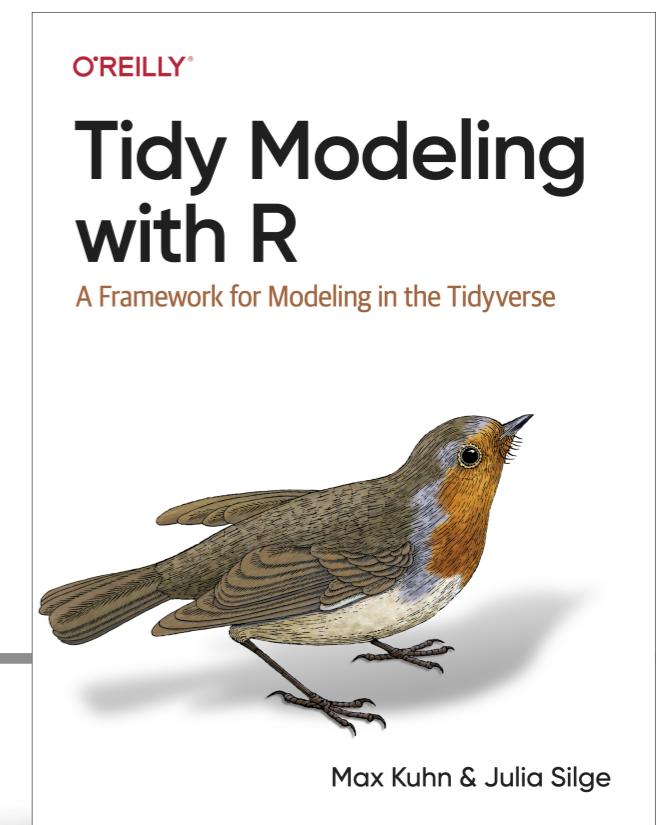
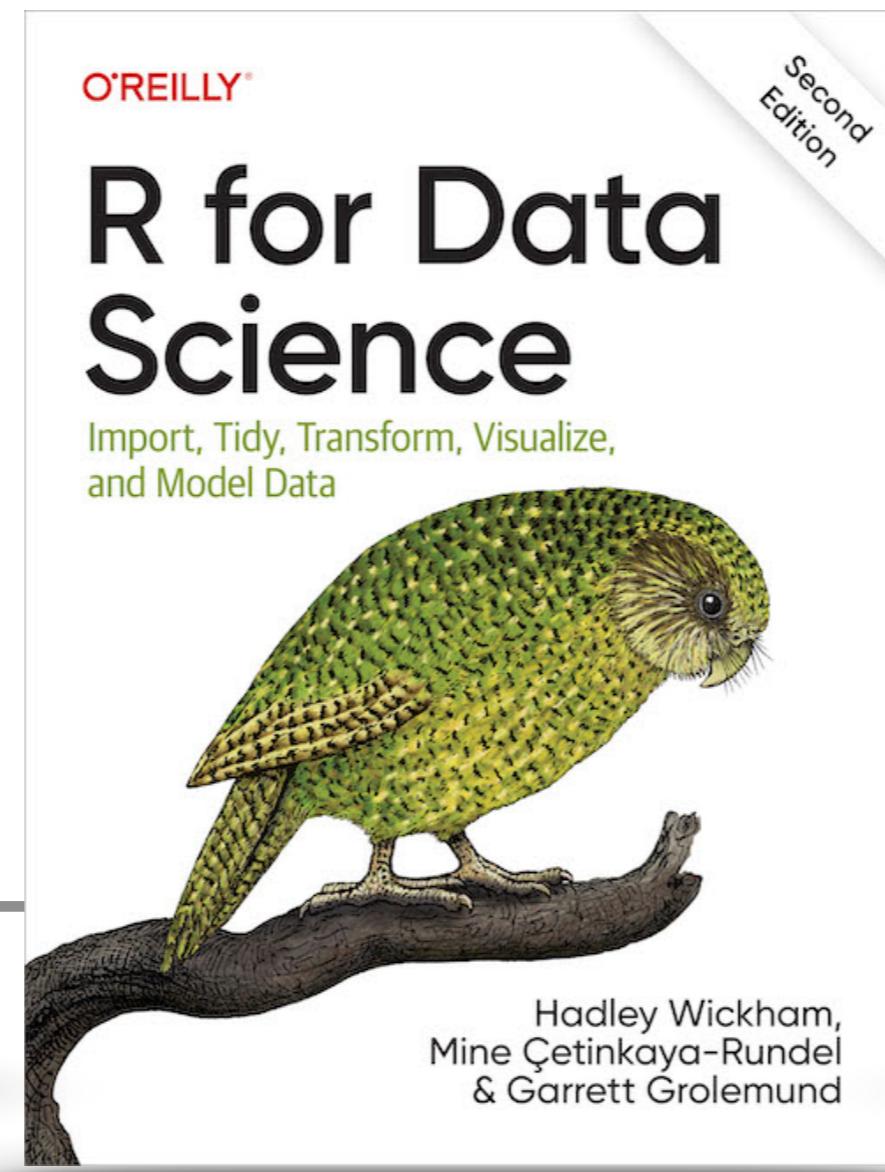
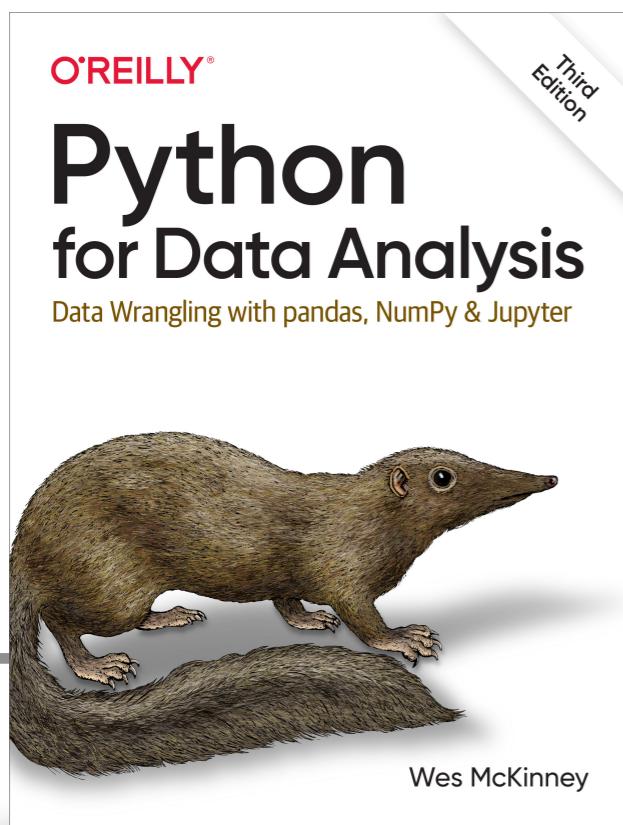
# Ricerca sociale

---



# Analisi dei dati

---



# Contatti

---

The screenshot shows a web browser window with the URL `lormatt.github.io`. The page content includes:

- A header with the name "Lorenzo Mattioli" and navigation links for "Home" and "About".
- A profile picture of a young man with glasses and a beard, sitting outdoors.
- A bio section:
  - I have a background in Economics, and am now a Master student in Politics and Social Policy at the University of Bologna. I am also involved with the association *una Regione per Restare* (RxR) as head of their Social Observatory.
  - Due to my mixed background, I can count on both solid foundations in data analysis and econometrics and extensive domain knowledge in social sciences.
  - My main research interests revolve around social stratification and material inequality, but this site will also accommodate some of my less "serious" work, as well as some teaching material.
- A link stating "My complete resume is available [here](#)".
- Social media icons for Instagram, GitHub, and Email.
- A "Cookie Preferences" link at the bottom.

**contatti:**



@ermattaraus



@lorMatt



[lorenzomattioli01@gmail.com](mailto:lorenzomattioli01@gmail.com)

# Slide

---

