

Seminario introduttivo su: “Principi, tecniche e strumenti per l’analisi riproducibile dei dati”

Presidenza del Consiglio dei Ministri,
Dip. Politiche di Coesione, UdM PNRR

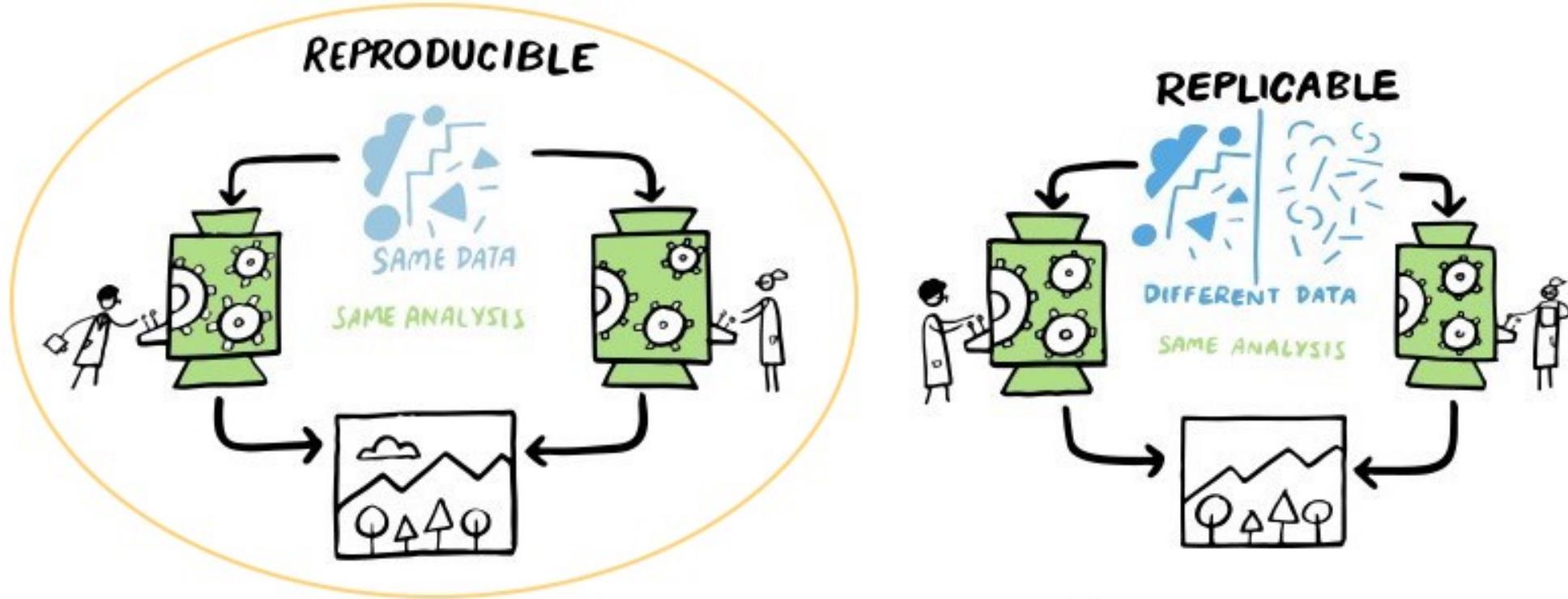
Luisa M. Mimmi

July 14, 2023

Temi del Seminario

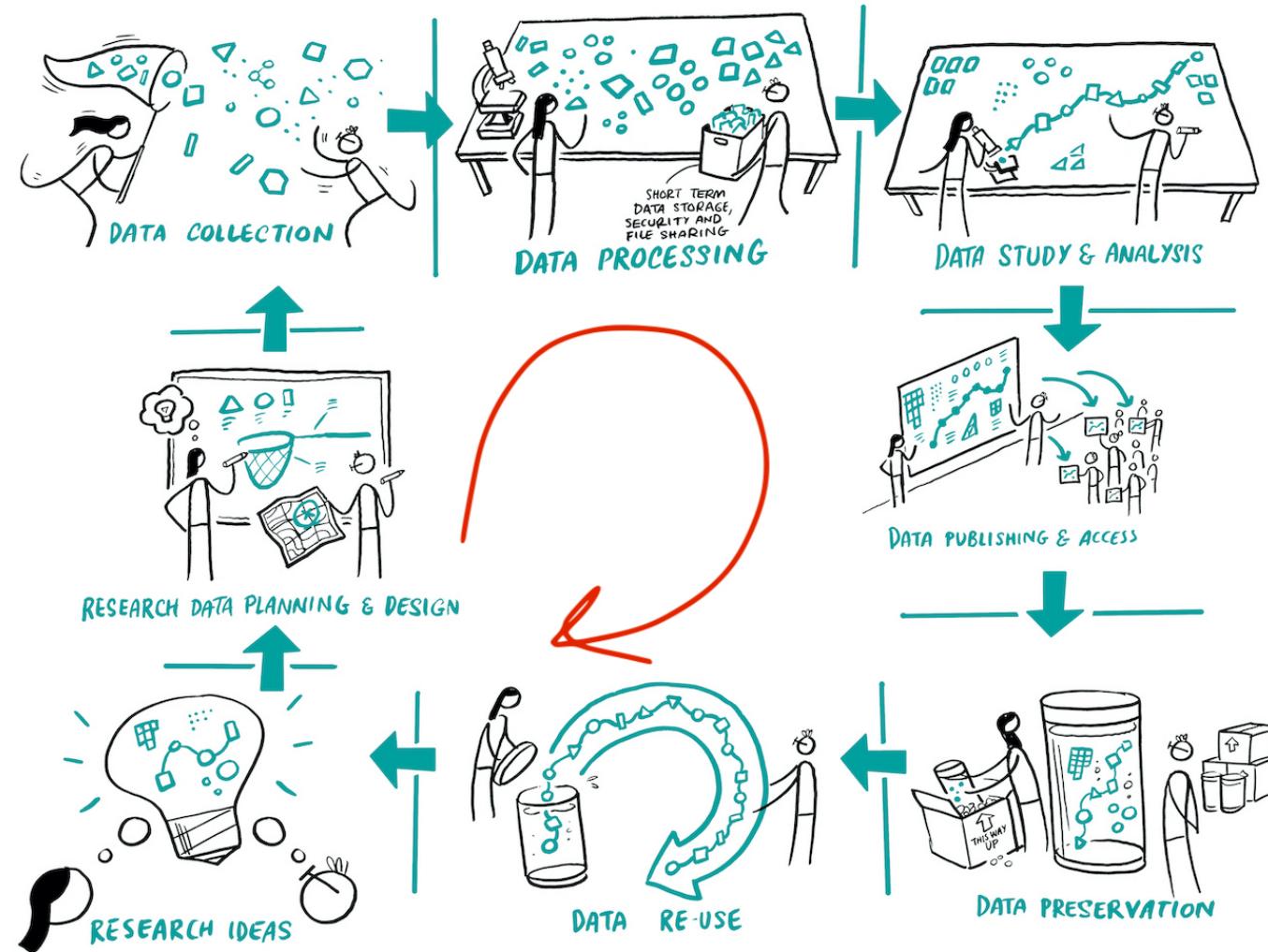
1. Cosa s'intende per “riproducibilità”?
2. Perché condurre analisi con approccio riproducibile?
3. Buone pratiche da seguire (e/o errori comuni)
 - organizzazione dei dati
 - strumenti di analisi (cenni)
4. Conclusioni e considerazioni

Cos'è la “riproducibilità”?



- **RIPRODUCIBILE:** STESSO risultato dalla STESSA analisi sugli **STESSI dati**
- **REPLICABILE:** STESSO risultato dalla STESSA analisi su **DATI DIVERSI**

Quando adottare tale approccio?



Scriberia

Perché adottare un'approccio riproducibile?

Chi ne beneficia?

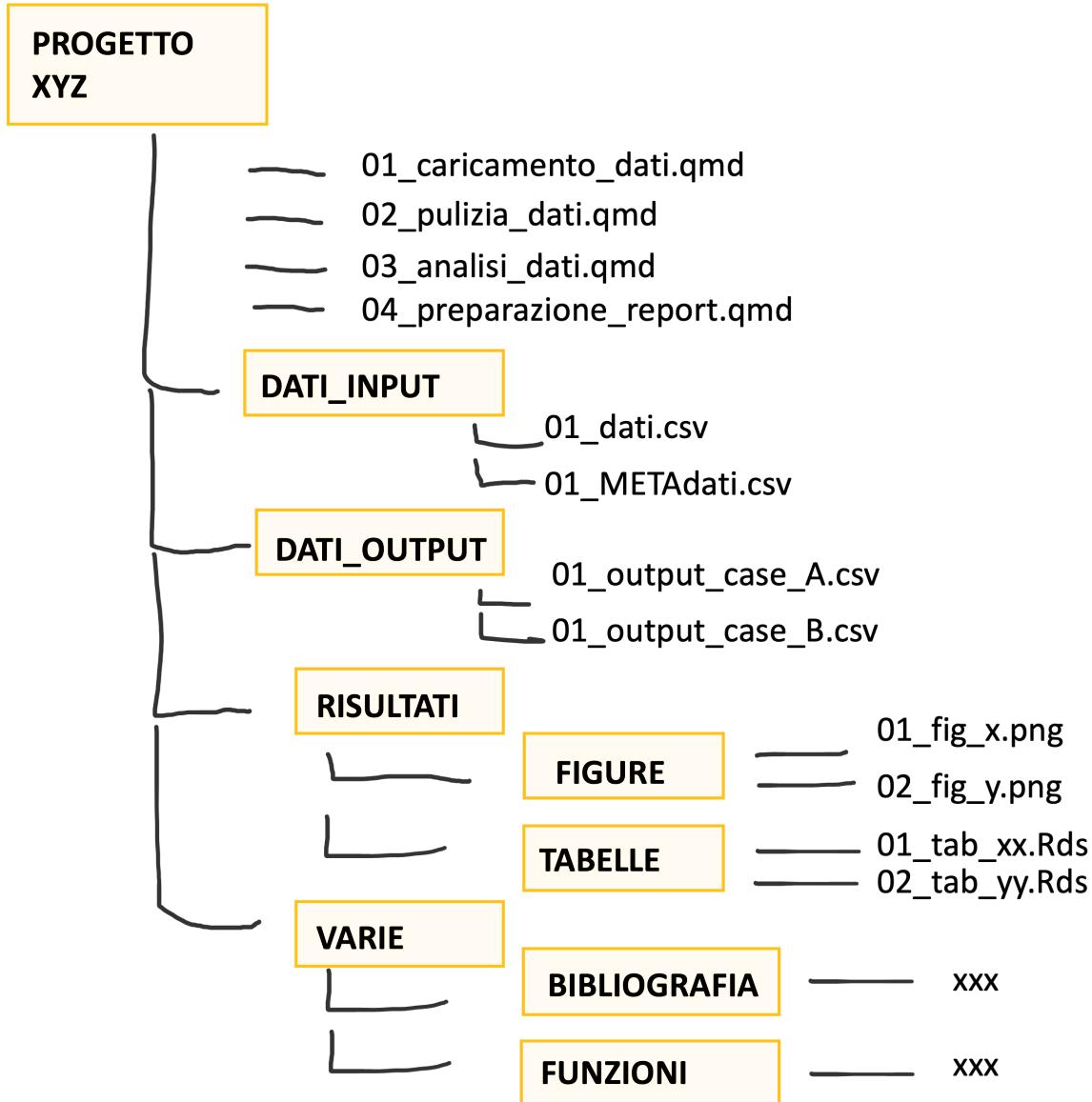
1. IO, soprattutto nel futuro
2. Colleghi/ co-autori
3. La comunità scientifica:
 - potrà studiare/ replicare/ estendere...
4. Editori e peer-reviewers
5. Revisori/ Controllori
6. Media/ Divulgatori

... se la mia analisi diventa:

1. RI-UTILIZZABILE
2. COLLABORATIVA
3. TRASPARENTE / CITABILE
4. DIMOSTRABILE
5. CREDIBILE
6. ACCESSIBILE

Buone PRATICHE da seguire (e/o errori classici da evitare)

Organizzazione file



- cartella DATI_INPUT in modalità “**read only**”
- struttura in **sequenza logica** (“01_caricamento”, “02_pulizia”)
- nomi “**machine readable**”:
 - (“Pessim@ .Nom€ X FILE. 17 LuGlio 2023.pdf”)
 - brevi, senza troppi spazi o simboli (SOLO 1 “.”)
 - numerati con “**left-padding**” (“01_”, “02_”, “03_”, ... “10_”)
 - convenzioni x estrarre informazioni (...**_fig_**..., ...**_tab_**..., **case_**...)

Contenuto di un file (da incubo)

game	date	team		losing	city	number
XXXVIII	2/1/2004	New England Patriots	32 to 29	Carolina Panthers	Houston, TX	70 thousand
XXXIX	6-Feb-05		24–21	Eagles	Jacksonville FLORIDA	78.125
XL	february 5 2006	Pittsburgh Steelers	21–10	Seattle Seahawks	DETROIT MI	68.206
XLI	4 feb. 2007	Indianapolis Colts	29 to 17	bears	Miami Gardens Florida	74.512
XLII	2008 3 february	giants	17–14	New England Patriots	Glendale, AZ	71.101,00
fourty 3	1/2/09	Pittsburgh Steelers	27–23	Arizona Cardinals	tampa	70.774
					total	362.718
XLIV	7-feb-10	New Orleans Saints	31–17	Indianapolis Colts	Miami Gardens, Florida	74.059
fourty-five	2/6/2011	packers	31–25	Pittsburgh Steelers	Arlington, Texas	one hundred three thousand
XLVI	5 feb 12	New York Giants	21–17	New England Patriots	Indianapolis, Indiana	68.658
47	3-feb-13	Baltimore Ravens	34–31		Nola	71024
			109K	109 000	\$109.000 total spending	
XLVIII	2 feb '14	Seattle Seahawks	43–8	broncos	East Rutherford, NJ	1999 thousand dollars
XLIX	1-feb-15	New England Patriots	28–24	Seattle Seahawks	Glendale, Arizona	\$101.099,00
50	2/7/2016	Denver Broncos	24–10	Carolina Panthers	Santa Clara, California	20 thousand, but there might be more... need to check on this
LI	5-feb-17	patriots	34–28 (OT)	atl	tx	7,08E+04

- informazioni nascoste nella struttura
 - info colore o sfondo
 - righe/celle raggruppate o nascoste
 - colonne con mix dati, formule, o commenti
 - righe vuote
- format specifici x localizzazione:
 - 12-31-2023 o 31-12-2023?
 - \$ 100,000.5 o € 100.000,5?
 - ecc
- cella vuota usata come “0” (e non “NULL”)
- valori nulli problematici (-999, -1, 0)
- link o formule NON esportabili

Come strutturare i dati

“Tidy datasets are all alike, but every messy dataset is messy in its own way.”

~ Hadley Wickham

Come strutturare i dati (principi)

1. Ogni **entità** (tipo osservazioni) occupa una tabella separata

2. Tabella “*tidy*” ≈ “*RETTOANGOLARE*”

- ogni colonna una singola variabile
- ogni riga una singola osservazione
- ogni cella un valore

3. Ogni tabella ha una **chiave primaria** UNIVOCA x identificare le osservazioni

4. Ogni tabella contiene **chiavi esterne** per costruire **relazioni** tra tabelle

Tab 1 Progetti

prog_id	misura	sogg_att
00000000010001209472023	IS	COMUNE DI MOTTA SAN GIOVANNI
00000000010001951252023	FR	AGENZIA PER LA COESIONE TERRITORIALE
00000000010001197172022	BC	COMUNE DI VITULAZIO
00000000010001963492023	PE	ICARO CONSORZIO COOPERATIVE SOCIALI SOCI

Tab 2 Province

regione	provincia	prog_id
CAMPANIA	Caserta	00000000010001197172022
CALABRIA	Reggio di Calabria	00000000010001209472023

Importanza dei metadati

DATI



METADATI

- **Data:** 9 luglio 2023
- **Luogo:** Monviso
- **Risoluzione:** 4032x3024
- **Dimensione file:** 3,5 MB
- **Formato file:** .jpeg
- ...

“Data Codebook” (Codice dei dati)

NOME CAMPO SERVIZIO	DESCRIZIONE	K/O/C	FORMATO			TAB. CONT.	NOTE
			TIPO	DIMENS.			
				INT.	DEC.		
IdIntervento	Codice della Misura/SubMisura	O	Char	24		Vedi Tabella INM_INITIATIVE	Inserire Codice Misura/Submisura (del PNRR)
IdFondo	ID Fondo	O	Char	10			Inserire il Fondo (es: RRF per il PNRR)
TipoProcedura	Tipo procedura	O	Char	4		Vedi Tabella ZPTIPO_PRATT	(i.e. Bando, Circolare, etc.)
FlagAiuti	Flag Aiuti	C	Char	1			Valore booleano ('X' = sì, " = No)
CodiceRNA	Codice RNA	C	Char	10			'Se inserito il "Flag aiuti" il codice RNA diventa obbligatorio.
TipoResponsabile	Tipologia Responsabile	O	Char	2		Vedi tabella ZPRES_PRAATT	
DenomResponsabile	Denom. Responsabile	O	Char	255			Descrizione del responsabile della procedura

CONTIENE

- **Nome delle variabili:** [...]
- **Tipo:** [Numero, testo, data, ...]
- **Formato:** [...]
- **Range di valori:** [Vedi Tabella ZPTIPO_PRATT]

E.g. PROCEDURE DI ATTIVAZIONE (Fonte: PUC)

Altre tecniche per analisi riproducibile (cenni)

1. Modificare i file di dati tramite “script” (codici) —invece che a mano
 - R, Python, Stata, Excel macro
2. Automatizzare operazioni ripetute
 - “DRY” (Don’t Repeat Yourself!)
 - Organizzare procedure in funzioni dedicate (x pulire i dati, creare grafici...)
3. Adottare **controllo di versione** per i file
 - Git, Github, OSF
4. Utilizzare software *open source* (ove possibile)
5. Usare e creare **open data** (ove possibile)

Automatizzazione (esempio)

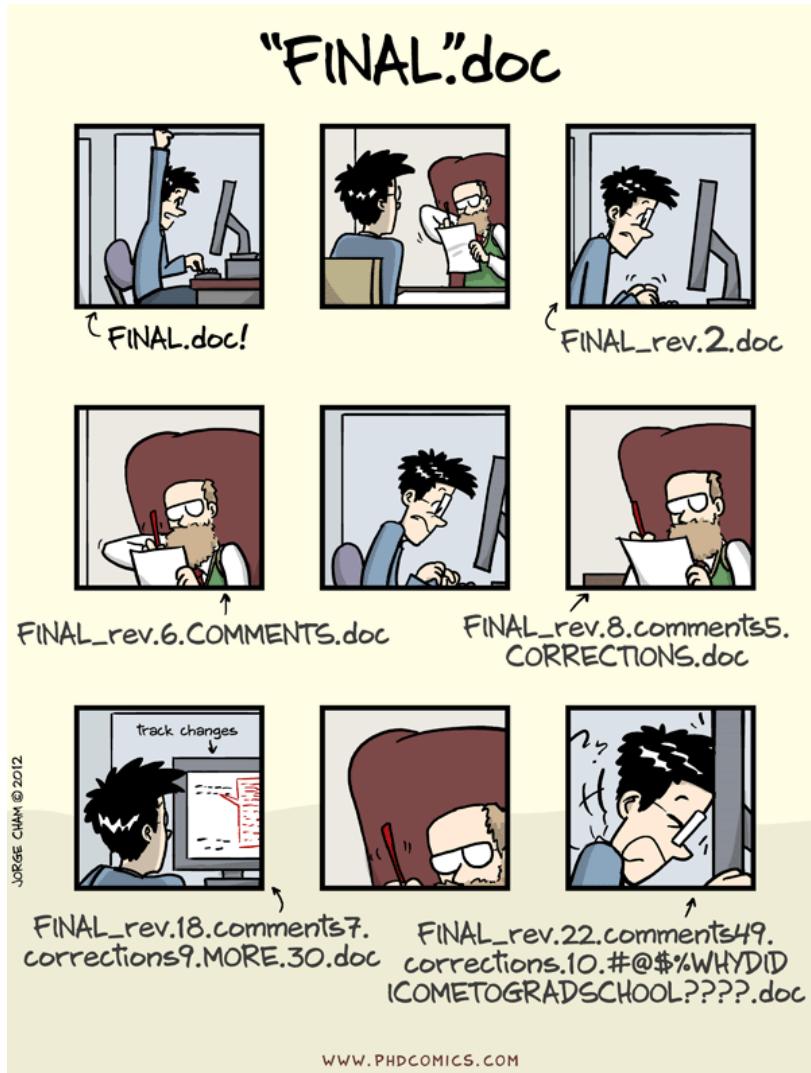
PROBLEMA: Riga sfalsata esportando da REGIS

A	D	E	U	V	W
1			Indicatori	PRG - Finanziamento totale (m€)	PRG - Costo ammesso progetti (m€)
PNRR - Codice Opendata Misura	PRG - Codice CUP	PRG - Codice Locale di Progetto (CLP)	PRG - Esito validazione		
M5C3I1.01	J81G22000130006	J81G22000130006	Non validato	299.143	299.143
M5C3I1.01	J88H22000460006	J88H22000460006	Validato	2.000.000	2.000.000
M5C3I1.01	J25I22005030006	J25I22005030006	Non validato	50.000	50.000
M5C3I1.01	I71G22000050006	I71G22000050006	Non validato	532.815	532.815

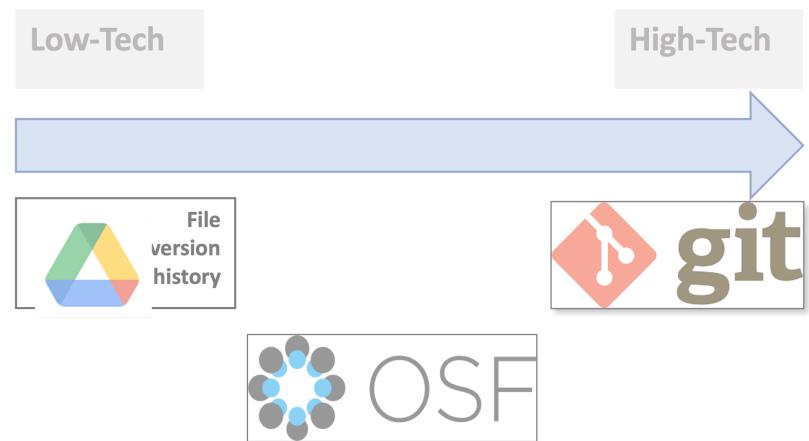
SOLUZIONE: Funzione in R per correggere file esportati

```
1 f_fix_lrow <- function(old_df_name, new_df_name){  
2   temp_1 <- old_df_name %>%  
3     dplyr::select(1:matches("Measures")) %>%  
4     janitor::row_to_names(1)  
5   temp_2 <- old_df_name %>%  
6     dplyr::select(-(1:matches("Measures"))) %>%  
7     dplyr::slice(-1)  
8   new_df_name <- dplyr::bind_cols(temp_1, temp_2)  
9 }
```

Controllo di versione



- Pratica mutuata dal mondo software, ma sempre più utilizzata
- Vari strumenti per il *controllo di versione* di file



Conclusioni

SPUNTI OFFERTI

- RIPRODUCIBILITÀ ≈ Tenere traccia (ordinata) di tutte le componenti di un progetto di analisi:
 - dati (grezzi e lavorati), metadati
 - procedure
 - risultati
- È un lavoro, ma premia in caso di collaborazione/ aggiornamento analisi/ *auditing*...
- Si può iniziare a implementare queste pratiche un po' alla volta, e a partire dagli strumenti in dotazione

DA APPROFONDIRE...

- Rapporto tra ricerca “riproducibile” & “aperta”?
 - *data ownership*
 - *privacy* (dati sensibili)
 - rischio reputazionale
- Rapporto tra analisi riproducibile e divulgazione risultati
 - visualizzazione dei dati, ecc.
- Scelta di strumenti (software e piattaforme) idonei

Riferimenti bibliografici

- The Turing Way Community. (2022). *The Turing Way: A handbook for reproducible, ethical and collaborative research*. Zenodo. <https://doi.org/10.5281/ZENODO.3233853>
- Broman, K. (2015). *Initial steps toward reproducible research*. <https://kbroman.org/steps2rr/>
- Berkeley Initiative for Transparency in the Social Sciences. (2022, ottobre 4). <https://www.bitss.org/>
- Wickham, H. (2014). *Tidy data*. The Journal of Statistical Software, vol. 59, 2014. <https://vita.had.co.nz/papers/tidy-data.pdf>
- Bryan, J. (2022). *How to name files*. Video: <https://www.youtube.com/watch?v=ES1LTlnpLMk>
- British Ecological Society. (2017). *A Guide to Reproducible Code in Ecology and Evolution*. <https://colauttilab.github.io/Readings/BES-Reproducible-Code.pdf>

