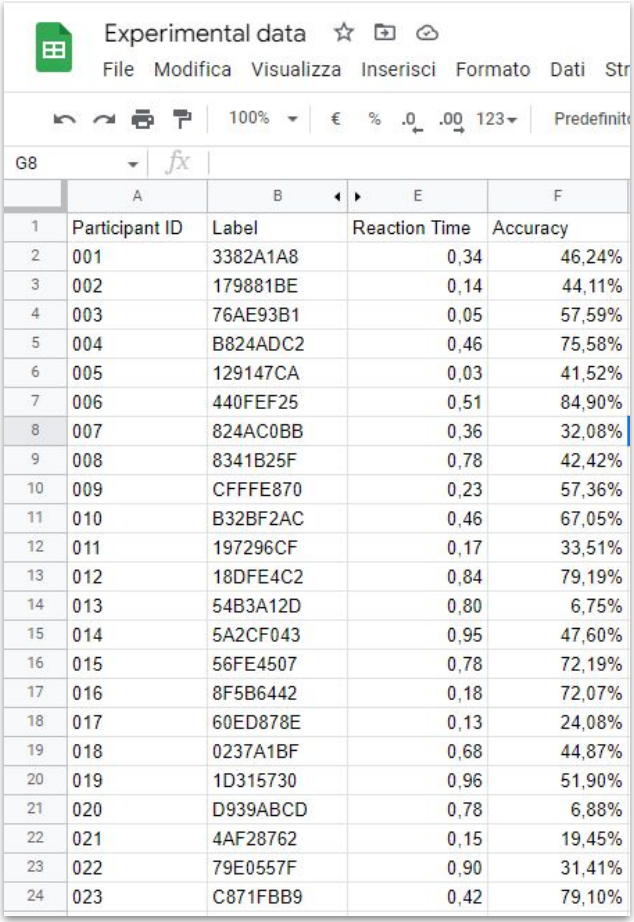
Structuring experimental data: rethinking data acquisitions to enable replications.

bit.ly/20230201-Verona-StructuringData

Vittorio Iacovella
@V_Iacovella / @v_iacovella@qoto.org
Verona, February 2nd 2023


Experimental data

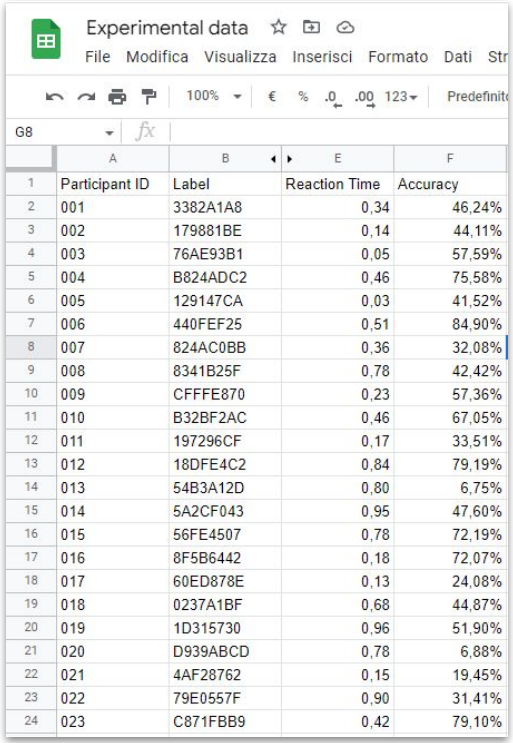
File Modifica Visualizza Inserisci Formato Dati Str

100% € % .0 .00 123 Predefinit

G8

	A	B	E	F
1	Participant ID	Label	Reaction Time	Accuracy
2	001	3382A1A8	0,34	46,24%
3	002	179881BE	0,14	44,11%
4	003	76AE93B1	0,05	57,59%
5	004	B824ADC2	0,46	75,58%
6	005	129147CA	0,03	41,52%
7	006	440FEF25	0,51	84,90%
8	007	824AC0BB	0,36	32,08%
9	008	8341B25F	0,78	42,42%
10	009	CFFFE870	0,23	57,36%
11	010	B32BF2AC	0,46	67,05%
12	011	197296CF	0,17	33,51%
13	012	18DFE4C2	0,84	79,19%
14	013	54B3A12D	0,80	6,75%
15	014	5A2CF043	0,95	47,60%
16	015	56FE4507	0,78	72,19%
17	016	8F5B6442	0,18	72,07%
18	017	60ED878E	0,13	24,08%
19	018	0237A1BF	0,68	44,87%
20	019	1D315730	0,96	51,90%
21	020	D939ABCD	0,78	6,88%
22	021	4AF28762	0,15	19,45%
23	022	79E0557F	0,90	31,41%
24	023	C871FBB9	0,42	79,10%

Why am I collecting and reporting these data?


Experimental data

File Modifica Visualizza Inserisci Formato Dati Str

100% € % .0 .00 123 Predefinit

G8

	A	B	E	F
1	Participant ID	Label	Reaction Time	Accuracy
2	001	3382A1A8	0,34	46,24%
3	002	179881BE	0,14	44,11%
4	003	76AE93B1	0,05	57,59%
5	004	B824ADC2	0,46	75,58%
6	005	129147CA	0,03	41,52%
7	006	440FEF25	0,51	84,90%
8	007	824AC0BB	0,36	32,08%
9	008	8341B25F	0,78	42,42%
10	009	CFFFE870	0,23	57,36%
11	010	B32BF2AC	0,46	67,05%
12	011	197296CF	0,17	33,51%
13	012	18DFE4C2	0,84	79,19%
14	013	54B3A12D	0,80	6,75%
15	014	5A2CF043	0,95	47,60%
16	015	56FE4507	0,78	72,19%
17	016	8F5B6442	0,18	72,07%
18	017	60ED878E	0,13	24,08%
19	018	0237A1BF	0,68	44,87%
20	019	1D315730	0,96	51,90%
21	020	D939ABCD	0,78	6,88%
22	021	4AF28762	0,15	19,45%
23	022	79E0557F	0,90	31,41%
24	023	C871FBB9	0,42	79,10%

Why am I collecting and reporting these data?

Data are quantitative tools to support a scientific claim.

After data collection, you rigorously analyze your data and build the scientific narration

Your final goal is to position your story as an official contribution to the general progress in knowledge

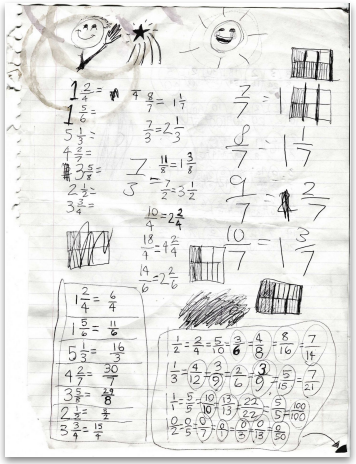
Why am I collecting and reporting these data?

Data are quantitative tools to support a scientific claim.

After data collection, you rigorously analyze your data and build the scientific narration

Your final goal is to position your story as an official contribution to the general progress in knowledge

readability, rigour, standardization, quality



1.https://www.publicdomainpictures.net/it/view-image.php?image=158321&picture=scribble-carta-nota
2.https://phdcomics.com/comics/archive_print.php?comicid=719 3. self-explanatory headlines

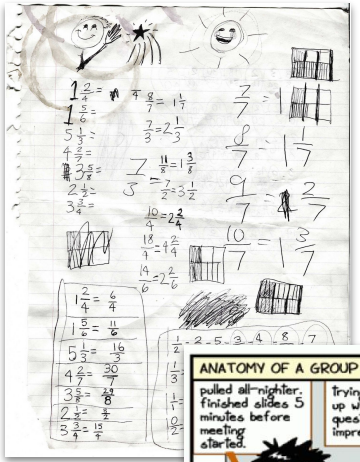
Why am I collecting and reporting these data?

Data are quantitative tools to support a scientific claim.

After data collection, you rigorously analyze your data and build the scientific narration

Your final goal is to position your story as an official contribution to the general progress in knowledge

readability, rigour, standardization, quality



1.https://www.publicdomainpictures.net/it/view-image.php?image=158321&picture=scribble-carta-nota
2.https://phdcomics.com/comics/archive_print.php?comicid=719 3. self-explanatory headlines

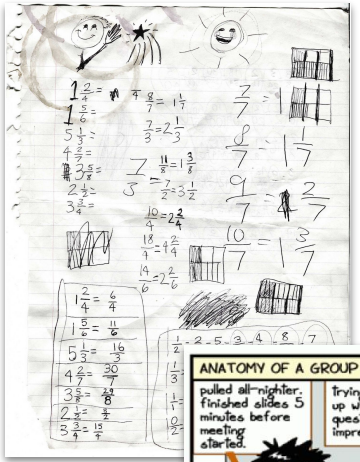
Why am I collecting and reporting these data?

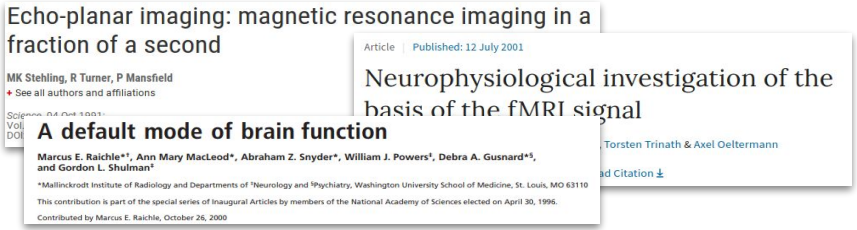
Data are quantitative tools to support a scientific claim.

After data collection, you rigorously analyze your data and build the scientific narration

Your final goal is to position your story as an official contribution to the general progress in knowledge

readability, rigour, standardization, quality





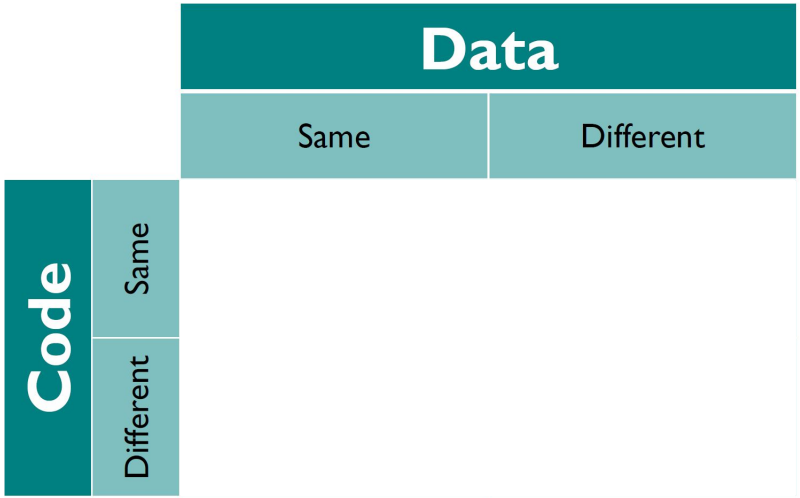
1.https://www.publicdomainpictures.net/it/view-image.php?image=158321&picture=scribble-carta-nota
2.https://phdcomics.com/comics/archive_print.php?comicid=719 3. self-explanatory headlines

Why am I collecting and reporting these data?

Data are quantitative tools to support a scientific claim.

After data collection, you rigorously analyze your data and build the scientific narration

Your final goal is to position your story as an official contribution to the general progress in knowledge



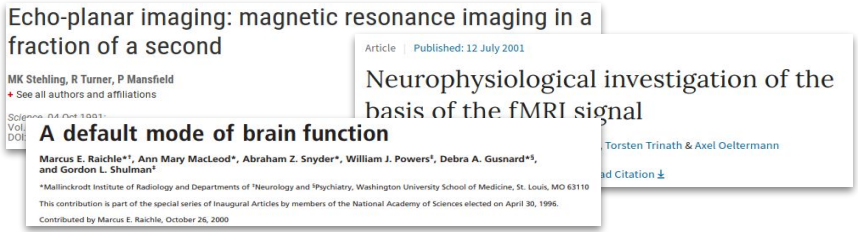
Whitaker, Kirstie (2017): Publishing a reproducible paper. figshare. Presentation.
https://doi.org/10.6084/m9.figshare.5440621.v2
https://experimentology.io/

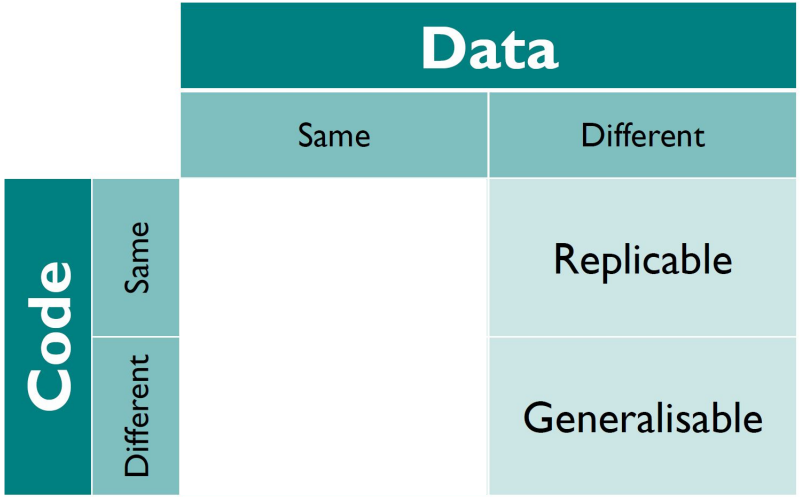
Why am I collecting and reporting these data?

Data are quantitative tools to support a scientific claim.

After data collection, you rigorously analyze your data and build the scientific narration

Your final goal is to position your story as an official contribution to the general progress in knowledge

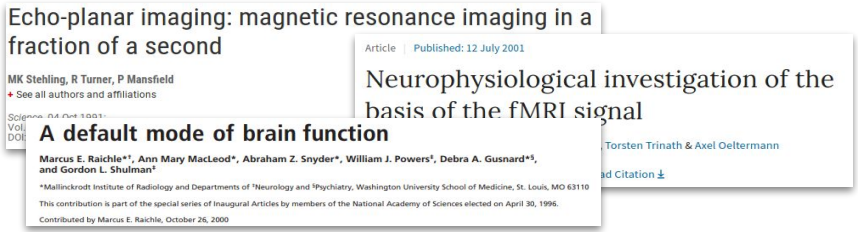


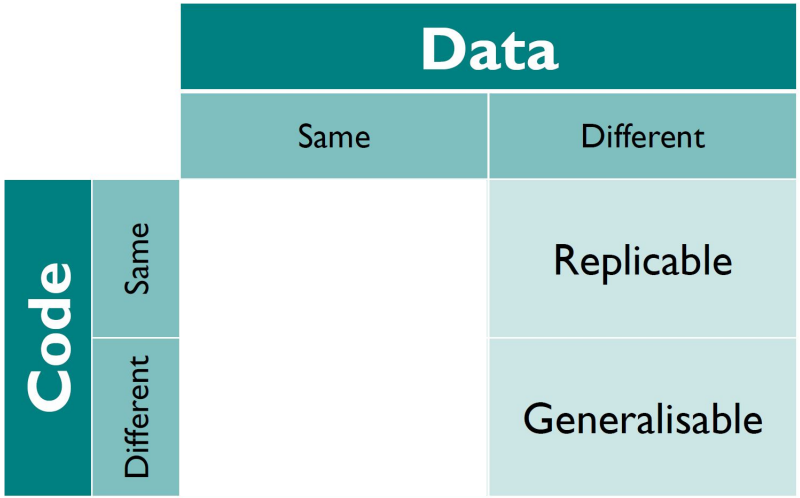


Whitaker, Kirstie (2017): Publishing a reproducible paper. figshare. Presentation.
https://doi.org/10.6084/m9.figshare.5440621.v2
https://experimentology.io/

What if

The common way to position scientific stories as official contributions to the general progress in knowledge turn out to have flaws?





Whitaker, Kirstie (2017): Publishing a reproducible paper. figshare. Presentation.
https://doi.org/10.6084/m9.figshare.5440621.v2
https://experimentology.io/

What if

The common way to position scientific stories as official contributions to the general progress in knowledge turn out to have flaws?

“Replication Crisis” - 2015

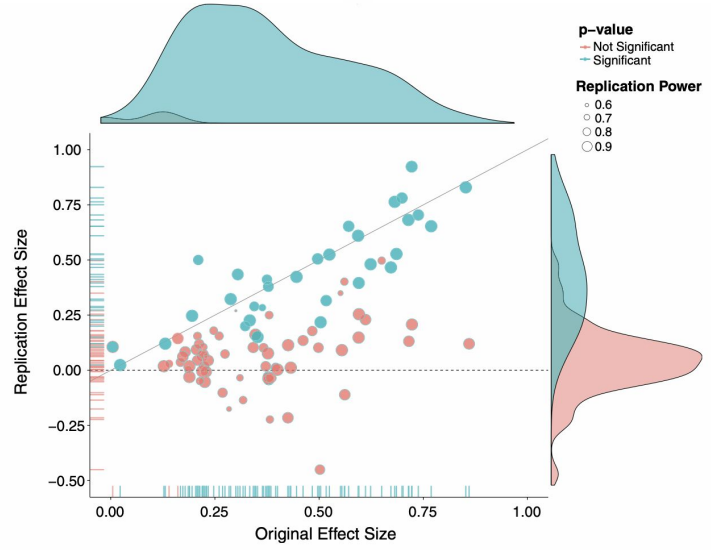
RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

replicability

Estimating the ~~reproducibility~~ of psychological science

Open Science Collaboration*



Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), 1–8.

Enable reproductions
Foster replications

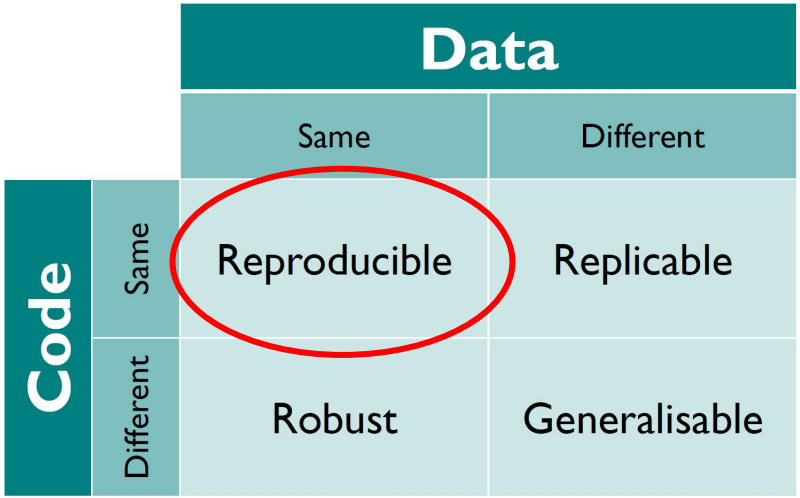
Integrate the common ways to build scientific stories

Build discoverable collections of properly arranged data and metadata, readable by both humans and machines

Take care of their preservation in time

A manifesto for reproducible science

Marcus R. Munafò[1,2*], Brian A. Nosek[3,4], Dorothy V. M. Bishop[5], Katherine S. Button[6], Christopher D. Chambers[7], Nathalie Percie du Sert[8], Uri Simonsohn[9], Eric-Jan Wagenmakers[10], Jennifer J. Ware[11] and John P. A. Ioannidis[12,13,14]



Munafò, M., Nosek, B., Bishop, D. et al. A manifesto for reproducible science. Nat Hum Behav 1, 0021 (2017). https://doi.org/10.1038/s41562-016-0021

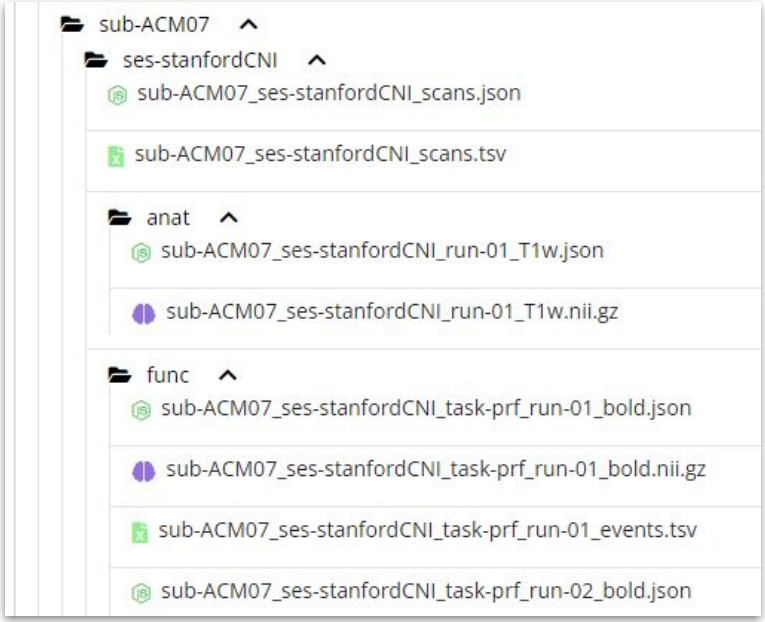
Experimental data

File Modifica Visualizza Inserisci Formato Dati Str

100% € % .0 .00 123 Predefinit

G8 fx

	A	B	E	F
1	Participant ID	Label	Reaction Time	Accuracy
2	001	3382A1A8	0,34	46,24%
3	002	179881BE	0,14	44,11%
4	003	76AE93B1	0,05	57,59%
5	004	B824ADC2	0,46	75,58%
6	005	129147CA	0,03	41,52%
7	006	440FEF25	0,51	84,90%
8	007	824AC0BB	0,36	32,08%
9	008	8341B25F	0,78	42,42%
10	009	CFFFE870	0,23	57,36%
11	010	B32BF2AC	0,46	67,05%
12	011	197296CF	0,17	33,51%
13	012	18DFE4C2	0,84	79,19%
14	013	54B3A12D	0,80	6,75%
15	014	5A2CF043	0,95	47,60%
16	015	56FE4507	0,78	72,19%
17	016	8F5B6442	0,18	72,07%
18	017	60ED878E	0,13	24,08%
19	018	0237A1BF	0,68	44,87%
20	019	1D315730	0,96	51,90%
21	020	D939ABCD	0,78	6,88%
22	021	4AF28762	0,15	19,45%
23	022	79E0557F	0,90	31,41%
24	023	C871FBB9	0,42	79,10%

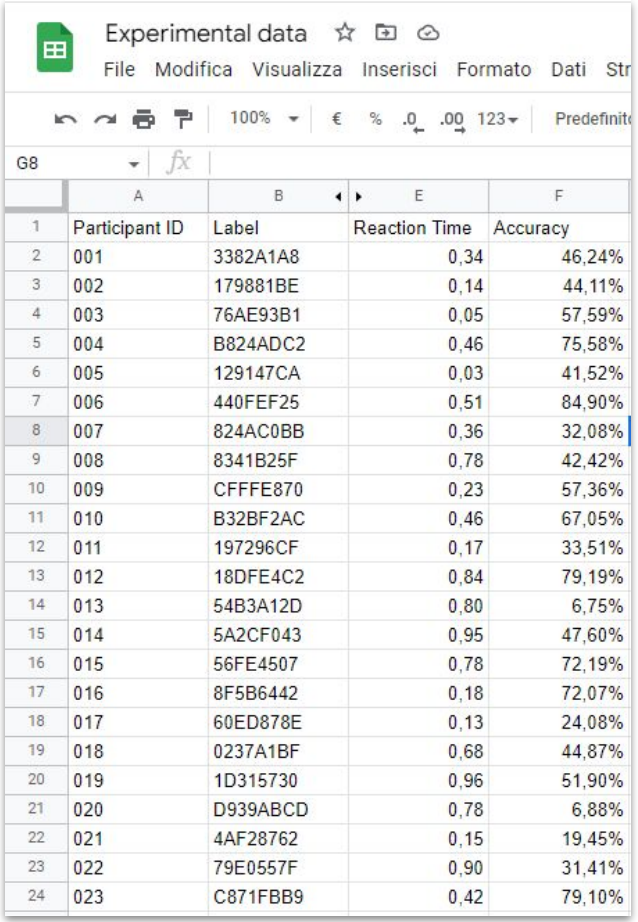


Structuring data
1. Redundancy

1. Redundancy

Do not be concise when collecting data

What you may consider useless for building a scientific narration could be crucial to enable reproducibility


Experimental data

File Modifica Visualizza Inserisci Formato Dati Str

100% € % .0 .00 123 Predefinit

G8 fx

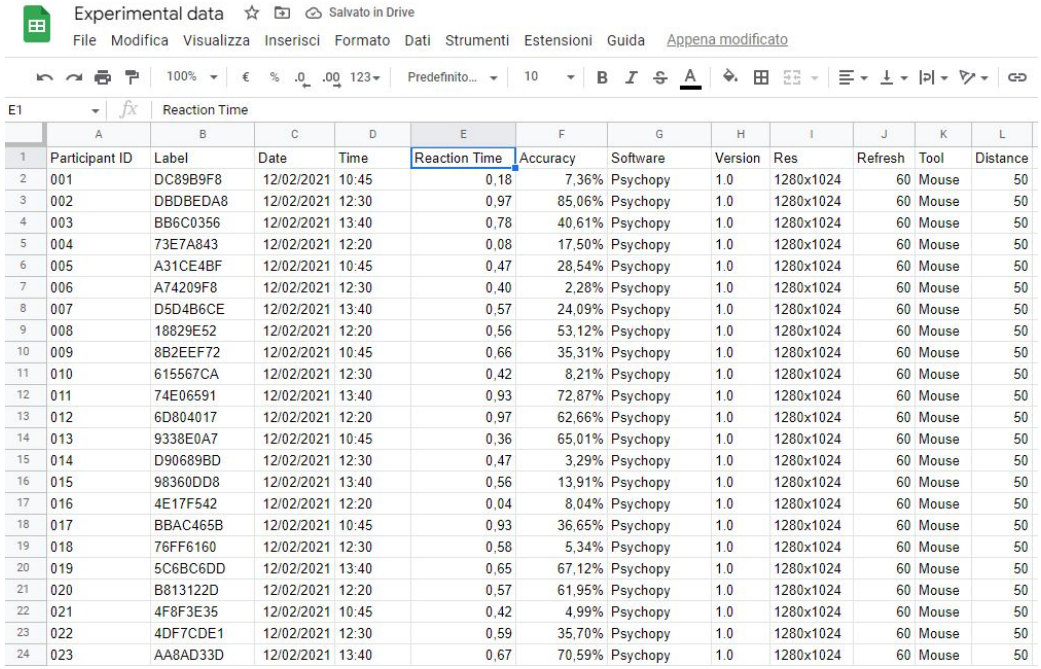
	A	B	E	F
1	Participant ID	Label	Reaction Time	Accuracy
2	001	3382A1A8	0,34	46,24%
3	002	179881BE	0,14	44,11%
4	003	76AE93B1	0,05	57,59%
5	004	B824ADC2	0,46	75,58%
6	005	129147CA	0,03	41,52%
7	006	440FEF25	0,51	84,90%
8	007	824AC0BB	0,36	32,08%
9	008	8341B25F	0,78	42,42%
10	009	CFFFE870	0,23	57,36%
11	010	B32BF2AC	0,46	67,05%
12	011	197296CF	0,17	33,51%
13	012	18DFE4C2	0,84	79,19%
14	013	54B3A12D	0,80	6,75%
15	014	5A2CF043	0,95	47,60%
16	015	56FE4507	0,78	72,19%
17	016	8F5B6442	0,18	72,07%
18	017	60ED878E	0,13	24,08%
19	018	0237A1BF	0,68	44,87%
20	019	1D315730	0,96	51,90%
21	020	D939ABCD	0,78	6,88%
22	021	4AF28762	0,15	19,45%
23	022	79E0557F	0,90	31,41%
24	023	C871FBB9	0,42	79,10%

1. Redundancy

Do not be concise when collecting data

What you may consider useless for building a scientific narration could be crucial to enable reproducibility

Systematically keep track of the software versions; monitor resolution and refresh rate, etc.



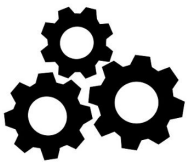
Experimental data ☆ Salvato in Drive

File Modifica Visualizza Inserisci Formato Dati Strumenti Estensioni Guida Appena modificato

100% € % .0 .00 123 Predefinito... 10

E1 fx Reaction Time

	A	B	C	D	E	F	G	H	I	J	K	L
1	Participant ID	Label	Date	Time	Reaction Time	Accuracy	Software	Version	Res	Refresh	Tool	Distance
2	001	DC89B9F8	12/02/2021	10:45	0,18	7,36%	Psychopy	1.0	1280x1024	60	Mouse	50
3	002	DBDBEDA8	12/02/2021	12:30	0,97	85,06%	Psychopy	1.0	1280x1024	60	Mouse	50
4	003	BB6C0356	12/02/2021	13:40	0,78	40,61%	Psychopy	1.0	1280x1024	60	Mouse	50
5	004	73E7A843	12/02/2021	12:20	0,08	17,50%	Psychopy	1.0	1280x1024	60	Mouse	50
6	005	A31CE4BF	12/02/2021	10:45	0,47	28,54%	Psychopy	1.0	1280x1024	60	Mouse	50
7	006	A74209F8	12/02/2021	12:30	0,40	2,28%	Psychopy	1.0	1280x1024	60	Mouse	50
8	007	D5D4B6CE	12/02/2021	13:40	0,57	24,09%	Psychopy	1.0	1280x1024	60	Mouse	50
9	008	18829E52	12/02/2021	12:20	0,56	53,12%	Psychopy	1.0	1280x1024	60	Mouse	50
10	009	8B2EEF72	12/02/2021	10:45	0,66	35,31%	Psychopy	1.0	1280x1024	60	Mouse	50
11	010	615567CA	12/02/2021	12:30	0,42	8,21%	Psychopy	1.0	1280x1024	60	Mouse	50
12	011	74E06591	12/02/2021	13:40	0,93	72,87%	Psychopy	1.0	1280x1024	60	Mouse	50
13	012	6D804017	12/02/2021	12:20	0,97	62,66%	Psychopy	1.0	1280x1024	60	Mouse	50
14	013	9338E0A7	12/02/2021	10:45	0,36	65,01%	Psychopy	1.0	1280x1024	60	Mouse	50
15	014	D90689BD	12/02/2021	12:30	0,47	3,29%	Psychopy	1.0	1280x1024	60	Mouse	50
16	015	98360DD8	12/02/2021	13:40	0,56	13,91%	Psychopy	1.0	1280x1024	60	Mouse	50
17	016	4E17F542	12/02/2021	12:20	0,04	8,04%	Psychopy	1.0	1280x1024	60	Mouse	50
18	017	BBAC465B	12/02/2021	10:45	0,93	36,65%	Psychopy	1.0	1280x1024	60	Mouse	50
19	018	76FF6160	12/02/2021	12:30	0,58	5,34%	Psychopy	1.0	1280x1024	60	Mouse	50
20	019	5C6BC6DD	12/02/2021	13:40	0,65	67,12%	Psychopy	1.0	1280x1024	60	Mouse	50
21	020	B813122D	12/02/2021	12:20	0,57	61,95%	Psychopy	1.0	1280x1024	60	Mouse	50
22	021	4F8F3E35	12/02/2021	10:45	0,42	4,99%	Psychopy	1.0	1280x1024	60	Mouse	50
23	022	4DF7CDE1	12/02/2021	12:30	0,59	35,70%	Psychopy	1.0	1280x1024	60	Mouse	50
24	023	AA8AD33D	12/02/2021	13:40	0,67	70,59%	Psychopy	1.0	1280x1024	60	Mouse	50

Structuring data
2. Interoperability 

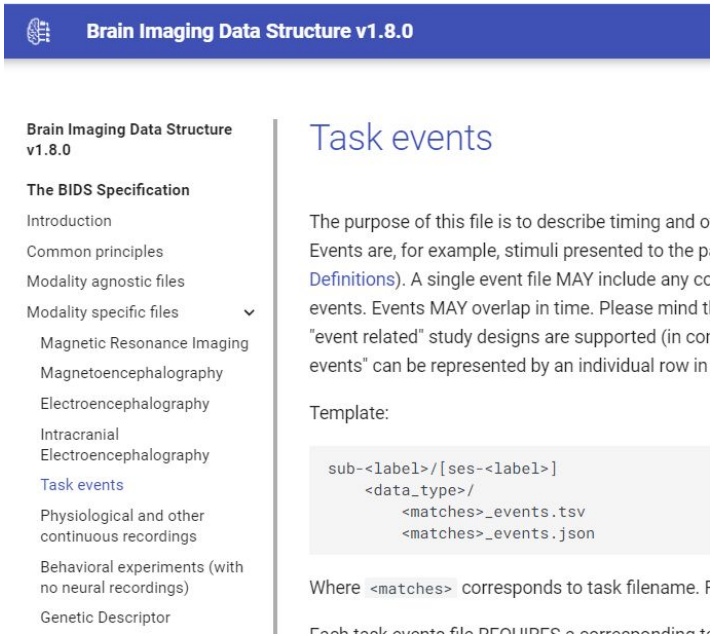
FAIR Principles https://force11.org/info/the-fair-data-principles/

2. Interoperability

Arrange the redundant information you collected using a standard schema according to the FAIR principles:

> Metadata and data use a formal, accessible, shared, and broadly applicable language for knowledge representation

BIDS
BRAIN IMAGING DATA STRUCTURE



https://bids.neuroimaging.io/

2. Interoperability

In BIDS information is arranged throughout a self-explanatory, informative directory structure

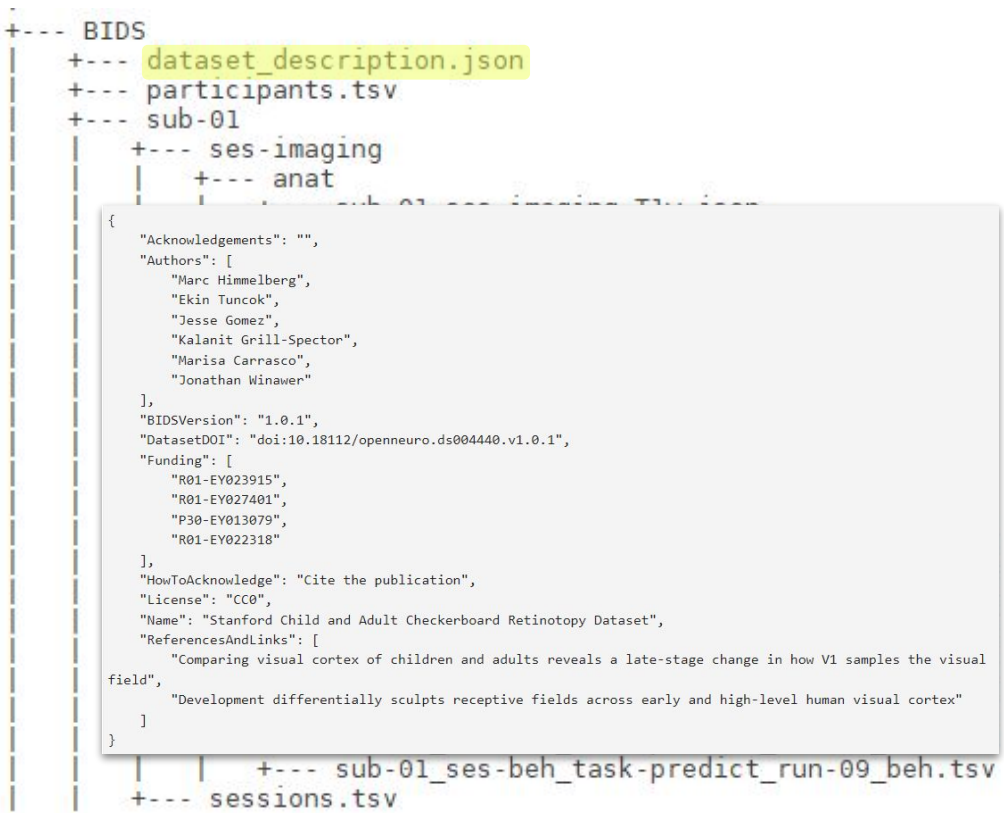
File labelling is modularly constructed in order to communicate information without exploring the internal file content

```
.
+--- BIDS
|   +--- dataset_description.json
|   +--- participants.tsv
|   +--- sub-01
|   |   +--- ses-imaging
|   |   |   +--- anat
|   |   |   |   +--- sub-01_ses-imaging_T1w.json
|   |   |   |   +--- sub-01_ses-imaging_T1w.nii.gz
|   |   |   +--- dwi
|   |   |   |   +--- sub-01_ses-imaging_dwi.bval
|   |   |   |   +--- sub-01_ses-imaging_dwi.bvec
|   |   |   |   +--- sub-01_ses-imaging_dwi.json
|   |   |   |   +--- sub-01_ses-imaging_dwi.nii.gz
|   |   |   +--- scans.tsv
|   |   +--- ses-task
|   |   |   +--- beh
|   |   |   |   +--- sub-01_ses-beh_task-predict_beh.json
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-01_beh.tsv
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-02_beh.tsv
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-03_beh.tsv
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-04_beh.tsv
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-05_beh.tsv
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-06_beh.tsv
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-07_beh.tsv
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-08_beh.tsv
|   |   |   |   +--- sub-01_ses-beh_task-predict_run-09_beh.tsv
|   |   +--- sessions.tsv
```


https://bids.neuroimaging.io/

2. Interoperability

In BIDS easy to access, human readable information is positioned within the dataset root directory in order to facilitate quick and easy dataset summary



https://bids.neuroimaging.io/

2. Interoperability

The native format for BIDS-compliant information is structured (.json / .tsv) text for metadata and data

Compilation of this files is easy, cross-platform and it does not require any specific tool

Serialization and de-serialization is native and easy to automatize

```
"trial_id":{
"LongName":"Identifier of a specific trial",
"Description":"It identifies specific trials throughout the experiment",
"Units" : "ordinal number"
},

"trial_type":{
"LongName":"Type of a specific trial",
"Description":"A number identifying the category of a trial",
"Units" : "integer number"
},

"RT":{
"LongName":"Reaction Time",
"Description":"It identifies the time when the participant responded to the instruction",
"Units" : "seconds"
},

"total_time": {
"LongName":"Trial full duration",
```

trial_id	trial_type	RT	total_time	grey_c	yellow_c
1	4	0.12983	2232.2	971	321
2	4	0.14882	2190.4	1868	272
3	3	0.049064		3286.1	64	111
4	3	0.13208	2040.3	69	176
5	5	0.1157	2786.7	1360	263
6	4	0.14808	2273.8	511	203
7	5	0.16499	3273.3	488	372
8	4	0.049056		2492	1100	367
9	5	0.18212	2558	260	228
10	3	0.26536	2758	1058	67
11	5	0.13204	3406.2	1652	216
12	5	0.14826	3822.1	800	104
13	5	0.098874		3406.2	290	342

https://bids.neuroimaging.io/

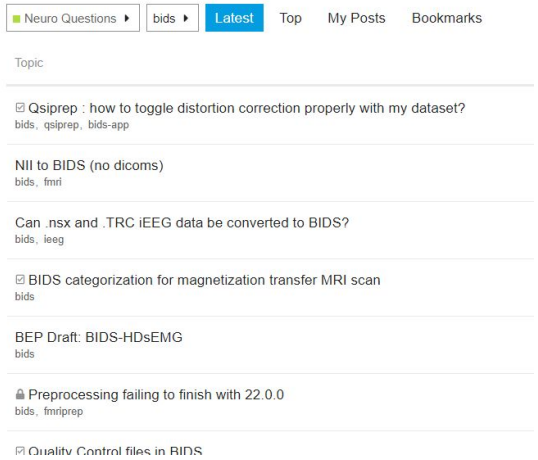
2. Interoperability

BIDS covers almost all the different angles of the neuroimaging galaxy

If you cannot find your modality, BIDS is modular, easily scalable and integrable through BIDS extension proposals

There is an active and responsive community to help you

BEP034	Computational modeling	• derivative • metadata	• Michael Schirner • Petra Ritter
BEP035	Modular extensions for individual participant data mega-analyses with non-compliant derivatives	• derivative	• Giuseppe Gallitto • Balint Kincses • Tamas Spisak
BEP036	Phenotypic Data Guidelines	• raw	• Eric Earl • Samuel Guay
BEP037	Non-Invasive Brain Stimulation (NIBS)	• raw	• Giacomo Bertazzoli • Vittorio Iacovella • Carlo Miniussi • Marta Bortoletto



https://bids.neuroimaging.io/

Structuring data
3. Findability 

FAIR Principles https://force11.org/info/the-fair-data-principles/

3. Findability

Place your structured dataset in a place where it will be assigned with a globally unique and persistent identifier

Dedicated repositories would require you to structure your data according to a standard schema

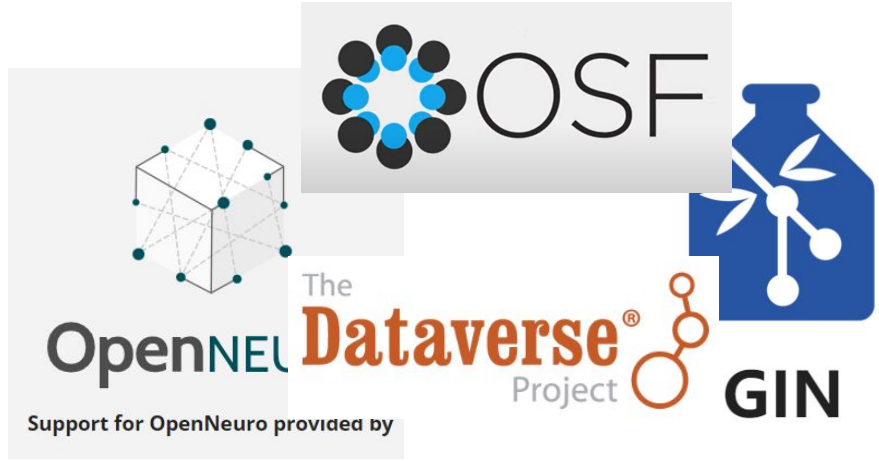
GIN

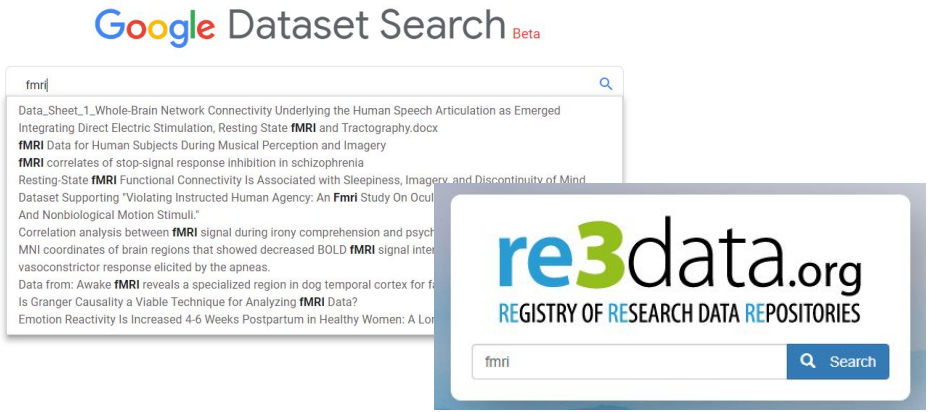
FAIR Principles https://force11.org/info/the-fair-data-principles/

3. Findability

Place your structured dataset in a place where it will be assigned with a globally unique and persistent identifier

This place should be registered or indexed in a searchable resource





FAIR Principles https://force11.org/info/the-fair-data-principles/

Take home messages

Reproducibility and replicability shape up research routines, starting from data acquisition

Building structured, machine readable data collections is a way to make your research reproducible and foster replication

Take home questions

Is your current project reproducible and / or replicable?

Can you easily describe and show to a colleague the features of a dataset you collected six months ago?

thanks for your attention

https://bit.ly/20230201_Iacovella_DataAcquisitions

Vittorio Iacovella
@V_Iacovella / @v_iacovella@qoto.org
Verona, February 2nd 2023