



**ICRS**

International  
Coral Reef Society

# ICRS Student Chapter

02.12.2021

**Giulia Puntin**



@sPuntinGi

## An intro to for reproducible research

what should happen  
between data collection and data analysis  
(that a lot of people do wrong)



# Reproducible research

Analyze the same data and obtain the same results

## Data

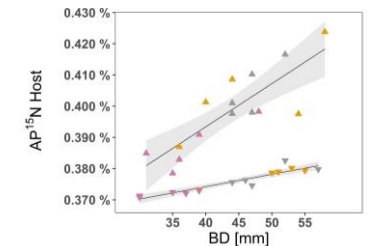
t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample	
1	3793116000	2020-03-11	1921.00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D13en1
2	3793116153	2020-03-11	1922.33	2	6	0.0	0	0.306	0.805	0.620	0.0	default_60.par	D13en2
3	3793116295	2020-03-11	1923.55	3	6	0.0	0	0.367	1.094	0.645	0.0	default_60.par	D13en3
4	3793116380	2020-03-11	1924.43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par	D13en4
5	3793116327	2020-03-11	1925.07	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D13en5
6	3793116387	2020-03-11	1926.27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D13en6
7	3793116522	2020-03-11	1928.42	7	6	0.0	0	0.030	0.086	0.651	0.0	default_60.par	D13en7
8	3793116613	2020-03-11	1930.13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D13en8
9	3793116699	2020-03-11	1931.39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D13en9
10	3793116770	2020-03-11	1932.50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D13en10
11	3793116844	2020-03-11	1934.04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D13en11
12	3793117043	2020-03-11	1937.23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D13en12
13	3793117112	2020-03-11	1938.32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D13en13
14	3793117173	2020-03-11	1939.33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D13en14
15	3793117247	2020-03-11	1940.47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D13en15



Data analysis



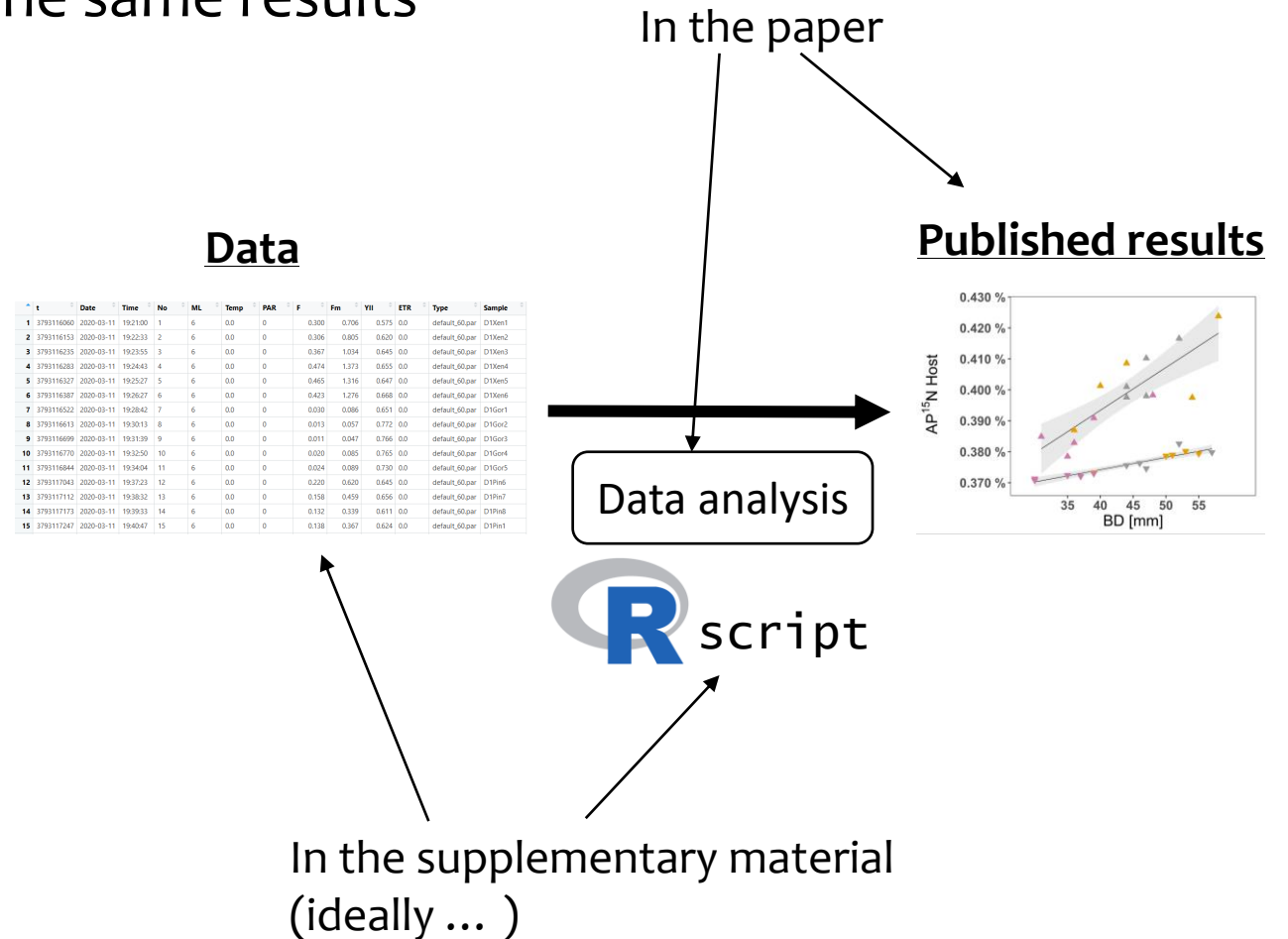
## Published results



Connected, but not the same  
as **replicability**: by repeating  
the same experiment reach  
the same conclusions

# Reproducible research

Analyze the same data and obtain the same results



# Reproducible research

But there's also that other part ...

Cleaned data  
ready for stats

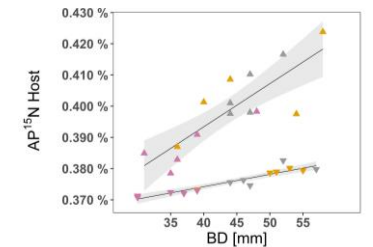
Data

#	t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample
1	3793116000	2020-03-11	19:21:00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D13en1
2	3793116153	2020-03-11	19:22:33	2	6	0.0	0	0.306	0.695	0.620	0.0	default_60.par	D13en2
3	3793116295	2020-03-11	19:23:55	3	6	0.0	0	0.367	1.094	0.645	0.0	default_60.par	D13en3
4	3793116380	2020-03-11	19:24:43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par	D13en4
5	3793116327	2020-03-11	19:25:07	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D13en5
6	3793116387	2020-03-11	19:26:27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D13en6
7	3793116552	2020-03-11	19:28:42	7	6	0.0	0	0.030	0.066	0.651	0.0	default_60.par	D13en7
8	3793116613	2020-03-11	19:30:13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D13en8
9	3793116699	2020-03-11	19:31:39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D13en9
10	3793116770	2020-03-11	19:32:50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D13en10
11	3793116844	2020-03-11	19:34:04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D13en11
12	3793117043	2020-03-11	19:37:23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D13en12
13	3793117112	2020-03-11	19:38:32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D13en13
14	3793117173	2020-03-11	19:39:33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D13en14
15	3793117247	2020-03-11	19:40:47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D13en15

Data analysis

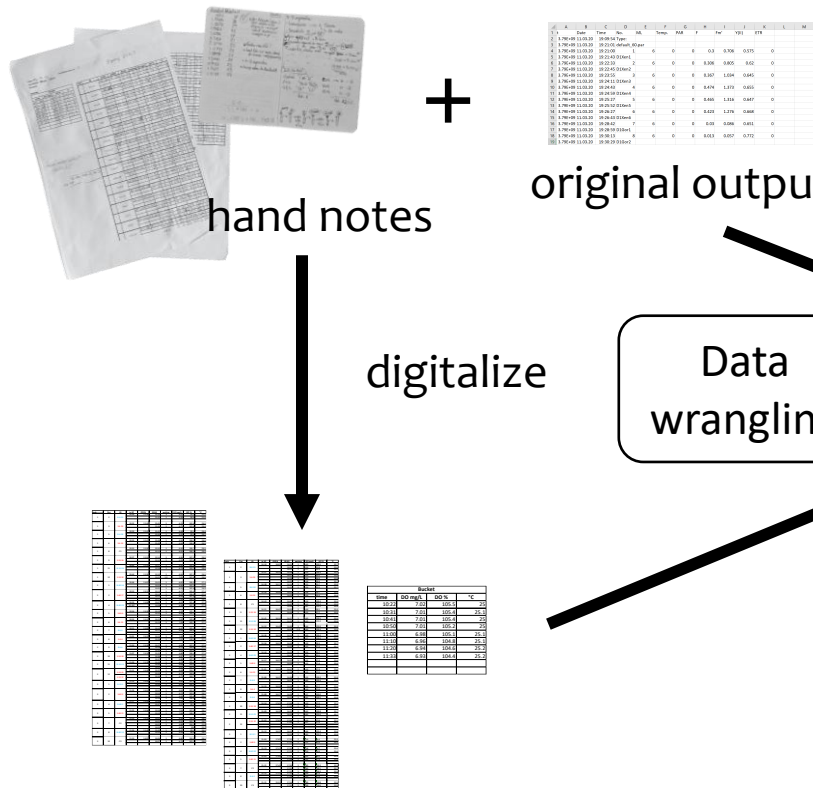


Published results



# Reproducible research

## Original data



original outputs

Data  
wrangling

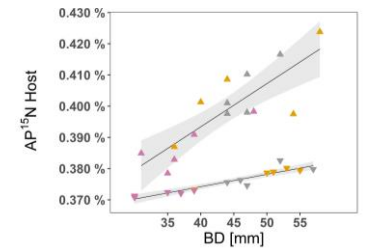
## Cleaned data ready for stats

t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample
1	2020-03-11	19:21:00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D19en1
2	2020-03-11	19:22:33	2	6	0.0	0	0.306	0.805	0.620	0.0	default_60.par	D19en2
3	2020-03-11	19:23:55	3	6	0.0	0	0.367	1.054	0.645	0.0	default_60.par	D19en3
4	2020-03-11	19:24:43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par	D19en4
5	2020-03-11	19:25:07	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D19en5
6	2020-03-11	19:26:27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D19en6
7	2020-03-11	19:28:42	7	6	0.0	0	0.030	0.086	0.651	0.0	default_60.par	D19en7
8	2020-03-11	19:30:13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D19en8
9	2020-03-11	19:31:39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D19en9
10	2020-03-11	19:32:50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D19en10
11	2020-03-11	19:34:04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D19en11
12	2020-03-11	19:37:23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D19en12
13	2020-03-11	19:38:32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D19en13
14	2020-03-11	19:39:33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D19en14
15	2020-03-11	19:40:47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D19en15

Data analysis

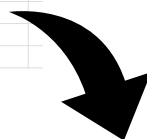


## Published results



## Unusable format ...

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	t	Date	Time	No.	ML	Temp.	PAR	F	Fm'	Y(II)	ETR		
2	3.79E+09	11.03.20	19:09:54	Type:									
3	3.79E+09	11.03.20	19:21:01	default_60.par									
4	3.79E+09	11.03.20	19:21:00	1	6	0	0	0.3	0.706	0.575	0		
5	3.79E+09	11.03.20	19:21:43	D1Xen1									
6	3.79E+09	11.03.20	19:22:33	2	6	0	0	0.306	0.805	0.62	0		
7	3.79E+09	11.03.20	19:22:45	D1Xen2									
8	3.79E+09	11.03.20	19:23:55	3	6	0	0	0.367	1.034	0.645	0		
9	3.79E+09	11.03.20	19:24:11	D1Xen3									
10	3.79E+09	11.03.20	19:24:43	4	6	0	0	0.474	1.373	0.655	0		
11	3.79E+09	11.03.20	19:24:59	D1Xen4									
12	3.79E+09	11.03.20	19:25:27	5	6	0	0	0.465	1.316	0.647	0		
13	3.79E+09	11.03.20	19:25:52	D1Xen5									
14	3.79E+09	11.03.20	19:26:27	6	6	0	0	0.423	1.276	0.668	0		
15	3.79E+09	11.03.20	19:26:43	D1Xen6									
16	3.79E+09	11.03.20	19:28:42	7	6	0	0	0.03	0.086	0.651	0		
17	3.79E+09	11.03.20	19:28:59	D1Gor1									
18	3.79E+09	11.03.20	19:30:13	8	6	0	0	0.013	0.057	0.772	0		
19	3.79E+09	11.03.20	19:30:29	D1Gor2									



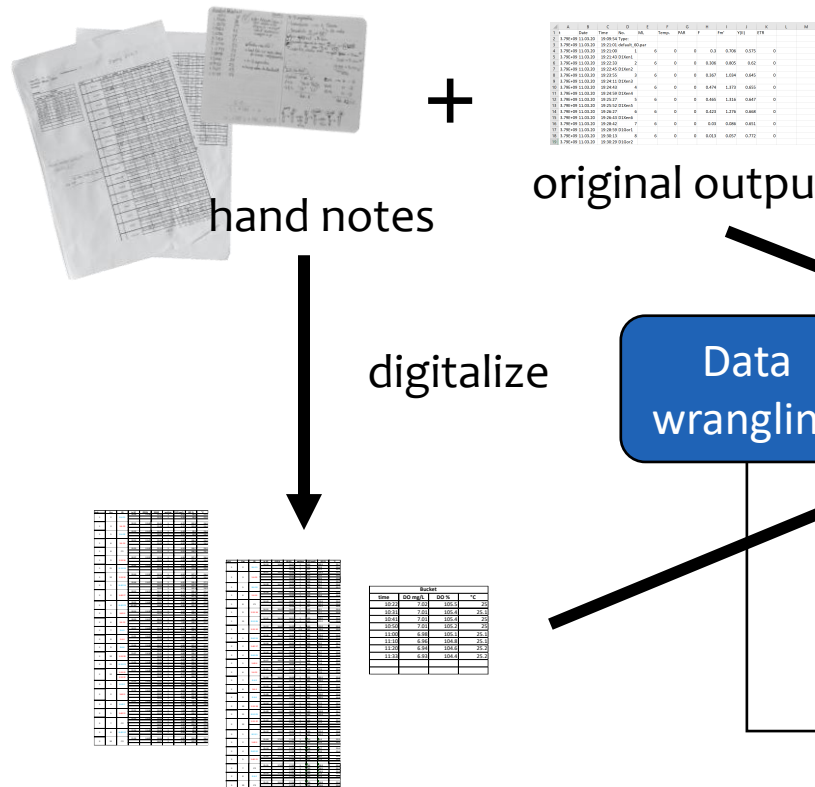
... ready to be analyzed ☺

	t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample
1	3793116060	2020-03-11	19:21:00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D1Xen1
2	3793116153	2020-03-11	19:22:33	2	6	0.0	0	0.306	0.805	0.620	0.0	default_60.par	D1Xen2
3	3793116235	2020-03-11	19:23:55	3	6	0.0	0	0.367	1.034	0.645	0.0	default_60.par	D1Xen3
4	3793116283	2020-03-11	19:24:43	4	6	0.0	0	0.474	1.373	0.655	0.0	default_60.par	D1Xen4
5	3793116327	2020-03-11	19:25:27	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D1Xen5
6	3793116387	2020-03-11	19:26:27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D1Xen6
7	3793116522	2020-03-11	19:28:42	7	6	0.0	0	0.030	0.086	0.651	0.0	default_60.par	D1Gor1
8	3793116613	2020-03-11	19:30:13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D1Gor2
9	3793116699	2020-03-11	19:31:39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D1Gor3
10	3793116770	2020-03-11	19:32:50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D1Gor4
11	3793116844	2020-03-11	19:34:04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D1Gor5
12	3793117043	2020-03-11	19:37:23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D1Pin6
13	3793117112	2020-03-11	19:38:32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D1Pin7
14	3793117173	2020-03-11	19:39:33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D1Pin8
15	3793117247	2020-03-11	19:40:47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D1Pin1



# Reproducible research

## Original data



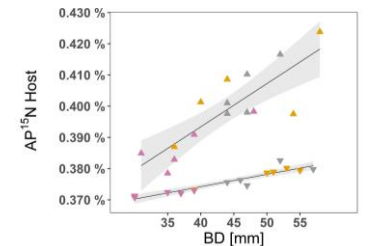
## Cleaned data ready for stats

t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample	
1	3793116000	2020-03-11	1921.00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D19en1
2	3793116153	2020-03-11	1922.33	2	6	0.0	0	0.306	0.695	0.620	0.0	default_60.par	D19en2
3	3793116295	2020-03-11	1923.55	3	6	0.0	0	0.367	1.094	0.645	0.0	default_60.par	D19en3
4	3793116380	2020-03-11	1924.43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par	D19en4
5	3793116327	2020-03-11	1925.27	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D19en5
6	3793116387	2020-03-11	1926.27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D19en6
7	3793116522	2020-03-11	1928.42	7	6	0.0	0	0.030	0.086	0.651	0.0	default_60.par	D19en7
8	3793116613	2020-03-11	1930.13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D19en8
9	3793116699	2020-03-11	1931.39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D19en9
10	3793116770	2020-03-11	1932.50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D19en10
11	3793116844	2020-03-11	1934.04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D19en11
12	3793117043	2020-03-11	1937.23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D19en12
13	3793117142	2020-03-11	1938.32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D19en13
14	3793117173	2020-03-11	1939.33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D19en14
15	3793117247	2020-03-11	1940.47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D19en15

Data analysis



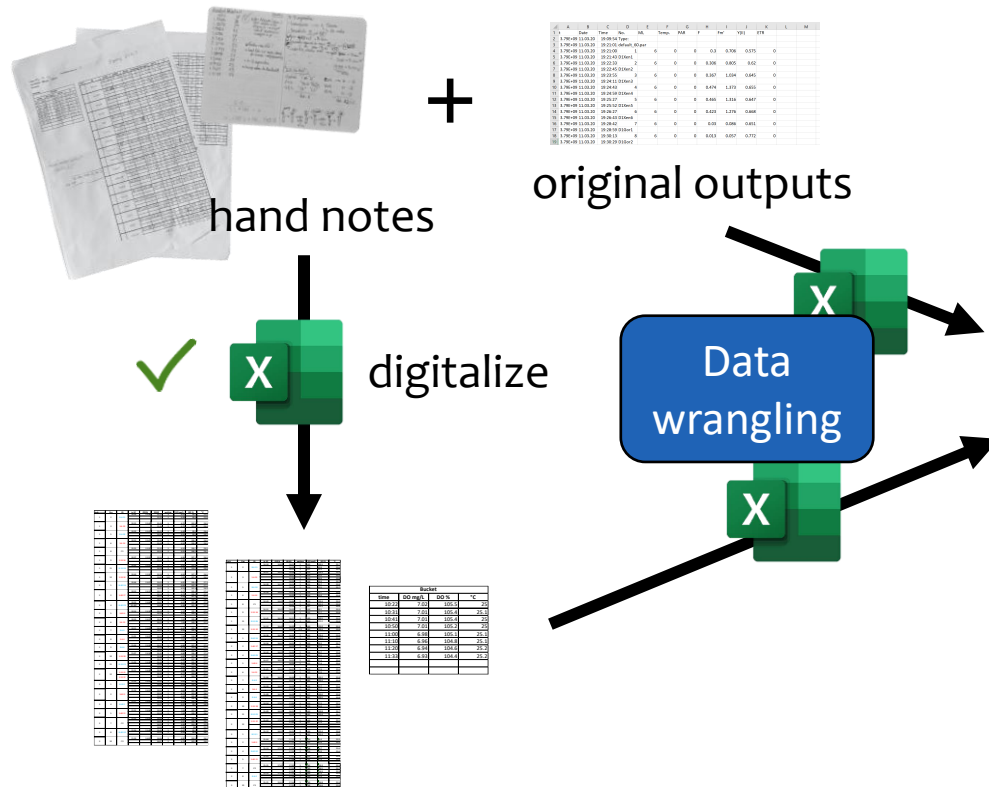
## Published results



Overlooked/underestimated  
⇒ 1. badly done (NOT reproducible)  
⇒ 2. can take up as much time as the data analysis (if not more)  
= sensitive step that can benefit a lot from improvement

# Reproducible research

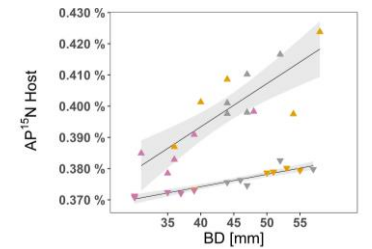
## Original data



## Cleaned data ready for stats

<i>t</i>	Date	Time	No	ML	Temp	PAR	F <sub>m</sub>	V <sub>m</sub>	ETS	default_50par	Sample
1	23/11/1660	02:10-11	192100	1	0	0.00	0.3300	0.706	0.575	0	D1x1n1
2	23/11/16153	02:10-11	192313	2	6	0.0	0.3306	0.805	0.620	0	D1x1n2
3	23/11/16235	02:10-11	192355	3	6	0.0	0.3367	1.034	0.645	0	D1x1n3
4	23/11/16307	02:10-11	192407	4	6	0.0	0.467	1.373	0.655	0	D1x1n4
5	23/11/16327	02:10-11	192527	5	6	0.0	0.445	1.316	0.647	0	D1x1n5
6	23/11/16387	02:10-11	192627	6	6	0.0	0.443	1.276	0.640	0	D1x1n6
7	23/11/16522	02:10-11	192842	7	6	0.0	0.000	0.066	0.651	0	D1x1n7
8	23/11/16613	02:10-11	193013	8	6	0.0	0.003	0.057	0.772	0	D1x1n8
9	23/11/16699	02:10-11	193139	9	6	0.0	0.011	0.047	0.766	0	D1x1n9
10	23/11/16770	02:10-11	193250	10	6	0.0	0.020	0.085	0.765	0	D1x1n10
11	23/11/16884	02:10-11	193434	11	6	0.0	0.024	0.089	0.750	0	D1x1n11
12	23/11/16944	02:10-11	193544	12	6	0.0	0.020	0.045	0.750	0	D1x1n12
13	23/11/17172	02:10-11	193632	13	6	0.0	0.118	0.619	0.656	0	D1x1n13
14	23/11/17173	02:10-11	193931	14	6	0.0	0.132	0.319	0.611	0	D1x1n14
15	23/11/17247	02:10-11	194047	15	6	0.0	0.118	0.367	0.624	0	D1x1n15

## Published results



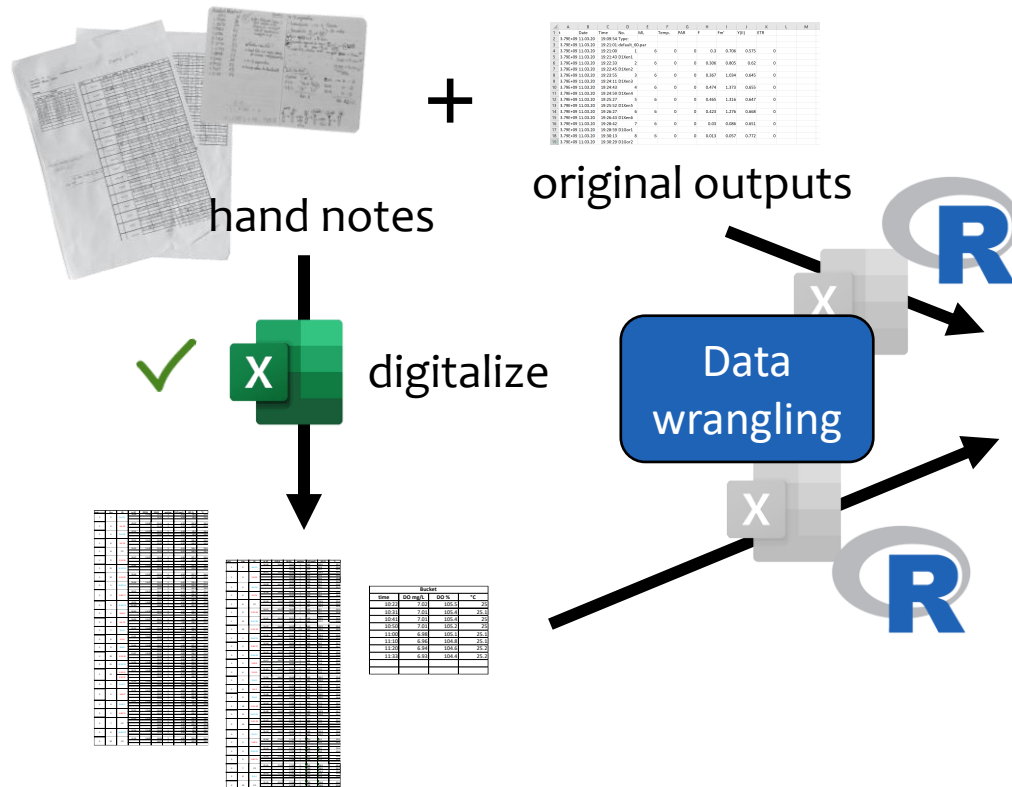
## Data analysis





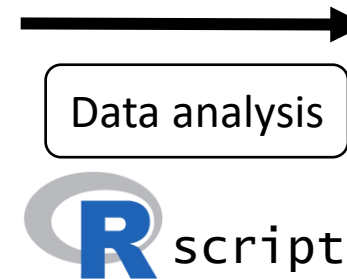
# Reproducible research

## Original data

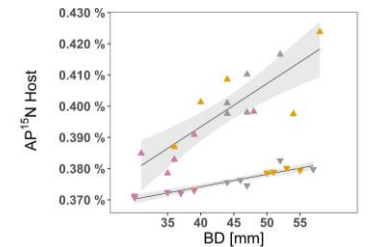


## Cleaned data ready for stats

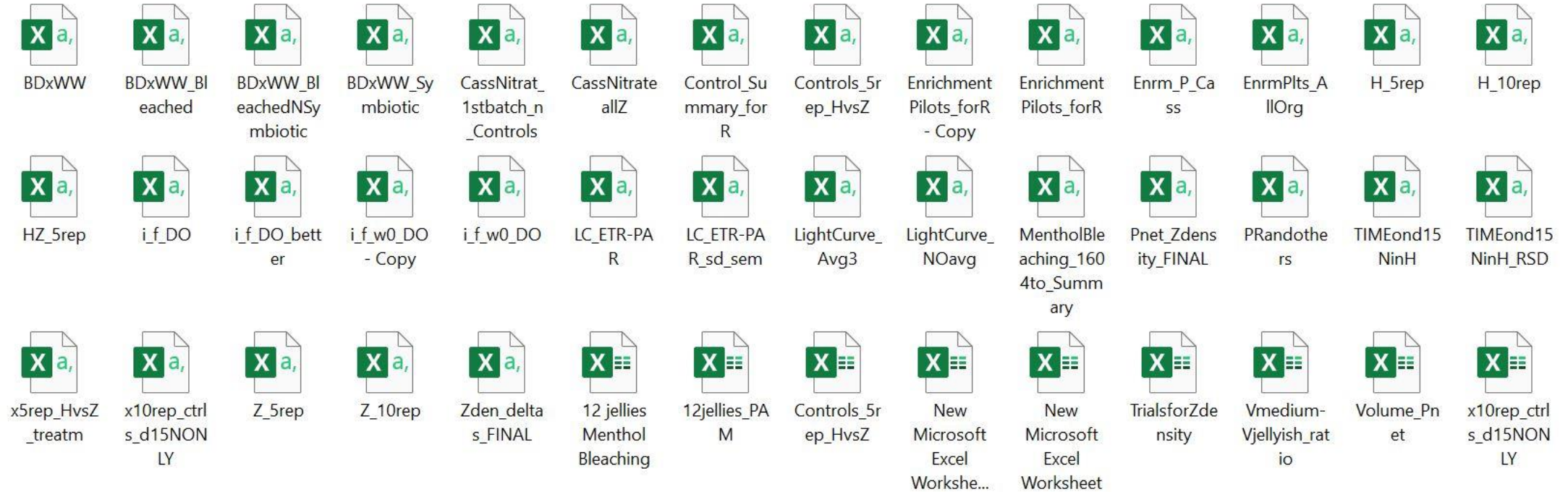
t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample
1	2020-03-11	19:21:00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D19en1
2	2020-03-11	19:22:33	2	6	0.0	0	0.306	0.805	0.620	0.0	default_60.par	D19en2
3	2020-03-11	19:23:55	3	6	0.0	0	0.367	1.054	0.645	0.0	default_60.par	D19en3
4	2020-03-11	19:24:43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par	D19en4
5	2020-03-11	19:25:07	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D19en5
6	2020-03-11	19:26:27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D19en6
7	2020-03-11	19:28:42	7	6	0.0	0	0.030	0.086	0.651	0.0	default_60.par	D19en7
8	2020-03-11	19:30:13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D19en8
9	2020-03-11	19:31:39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D19en9
10	2020-03-11	19:32:50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D19en10
11	2020-03-11	19:34:04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D19en11
12	2020-03-11	19:37:23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D19en12
13	2020-03-11	19:38:32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D19en13
14	2020-03-11	19:39:33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D19en14
15	2020-03-11	19:40:47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D19en15



## Published results



Does this look familiar?



Typical scenario of what happens when working in data in Excel: a new file is created for each version ...  
Naming can help only so far ...

Does this look familiar?



Typical scenario of what happens when working in data in Excel: a new file is created for each version ...  
Naming can help only so far ...

# Drop that spreadsheet already

Problems derived from working with spreadsheets:

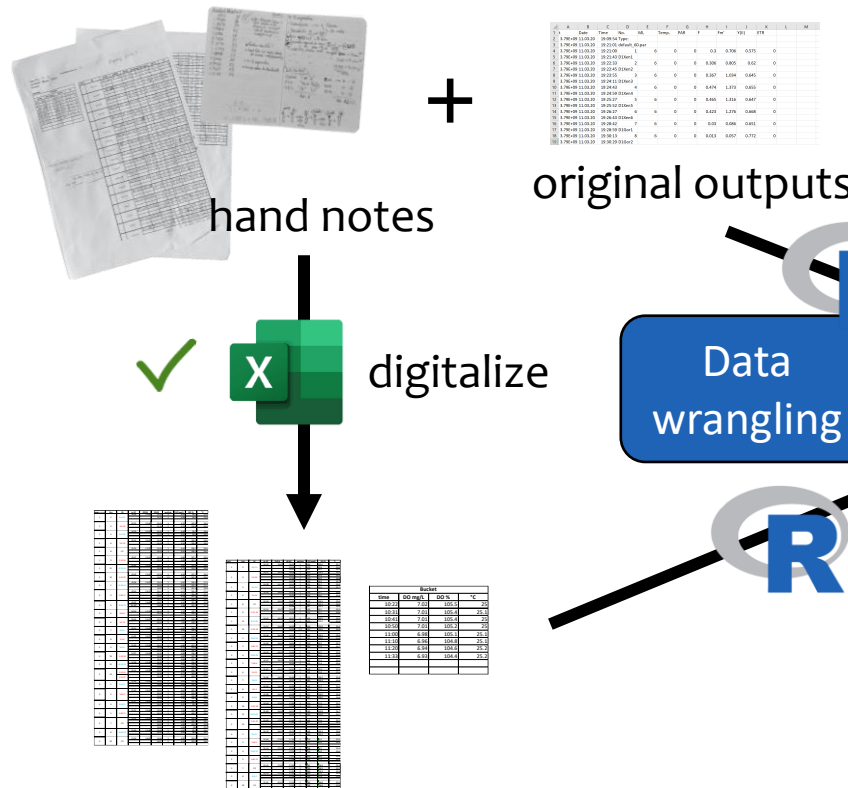
- Messy ...
- Error prone (e.g. manually copy-pasting the wrong thing in the wrong place)
- Not scalable (it just doesn't work with large data sets)
- **Not reproducible** (good luck trying to figure out what happened there ... !)

On the contrary, by working in **R**, you can do everything **without** ever **altering the original data!**  
(which also means that you can change your mind and easily un-do and re-do any operation)



# Reproducible research

## Original data



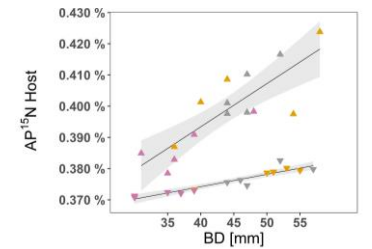
## Cleaned data ready for stats

t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample	
1	3793116000	2020-03-11	1921.00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D19en1
2	3793116153	2020-03-11	1922.33	2	6	0.0	0	0.306	0.805	0.620	0.0	default_60.par	D19en2
3	3793116235	2020-03-11	1923.55	3	6	0.0	0	0.367	1.054	0.645	0.0	default_60.par	D19en3
4	3793116282	2020-03-11	1924.43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par	D19en4
5	3793116327	2020-03-11	1925.27	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D19en5
6	3793116387	2020-03-11	1926.27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D19en6
7	3793116522	2020-03-11	1928.42	7	6	0.0	0	0.030	0.086	0.651	0.0	default_60.par	D19en7
8	3793116613	2020-03-11	1930.13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D19en8
9	3793116699	2020-03-11	1931.39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D19en9
10	3793116770	2020-03-11	1932.50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D19en10
11	3793116844	2020-03-11	1934.04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D19en11
12	3793117043	2020-03-11	1937.23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D19en12
13	3793117112	2020-03-11	1938.32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D19en13
14	3793117173	2020-03-11	1939.33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D19en14
15	3793117247	2020-03-11	1940.47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D19en15

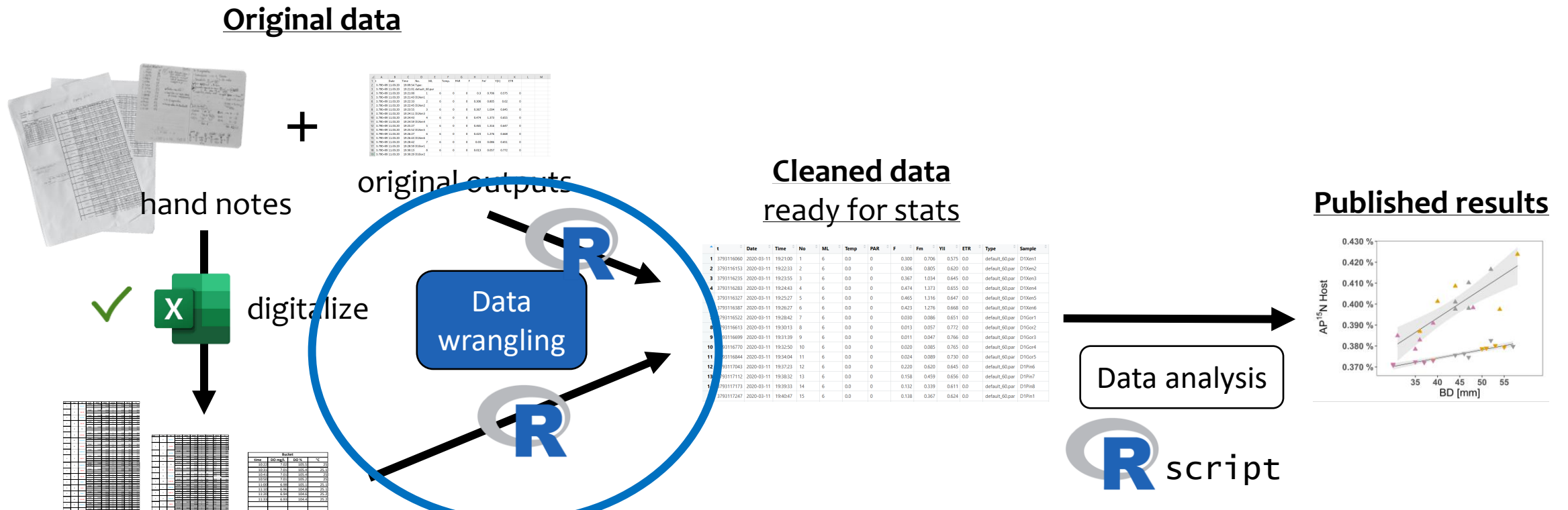
Data analysis

R script

## Published results



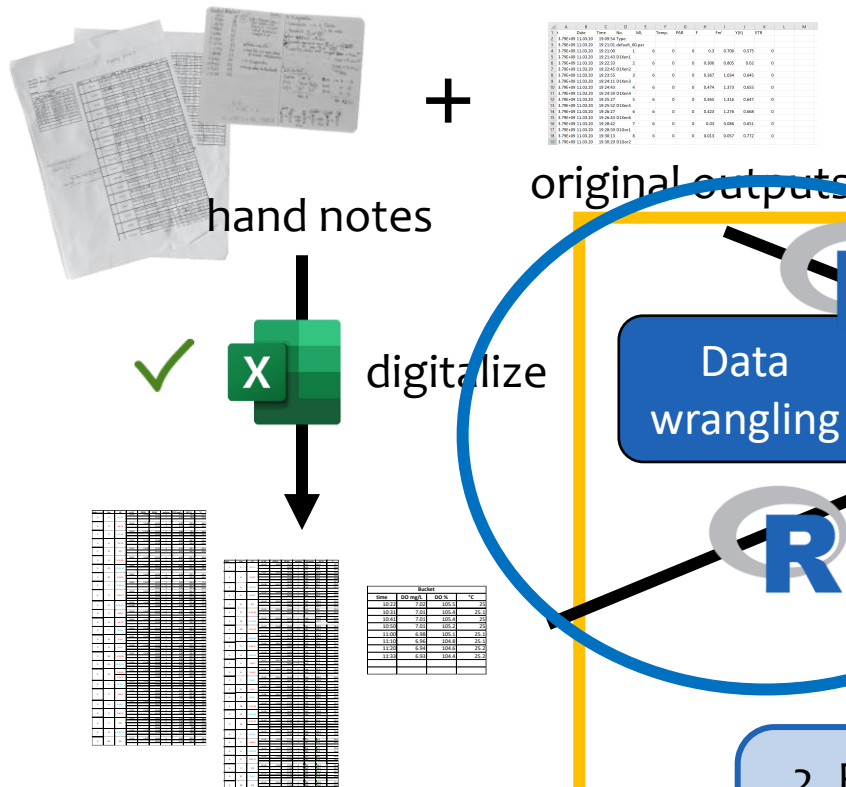
# Reproducible research



2. Reproducible data manipulation:  
use R from the very beginning of your work with data  
(not just for the stats and plots)!

# Reproducible research

## Original data



1. Fundamentals of data management:  
what happens before you even open R

## Cleaned data ready for stats

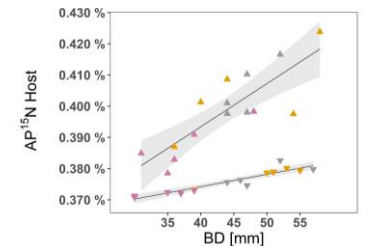
t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample	
1	2793116000	2020-03-11	1921.00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D13en1
2	2793116153	2020-03-11	1922.33	2	6	0.0	0	0.306	0.695	0.620	0.0	default_60.par	D13en2
3	2793116205	2020-03-11	1923.55	3	6	0.0	0	0.367	1.094	0.645	0.0	default_60.par	D13en3
4	2793116260	2020-03-11	1924.43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par	D13en4
5	2793116327	2020-03-11	1925.07	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D13en5
6	2793116387	2020-03-11	1926.27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D13en6
7	2793116522	2020-03-11	1928.42	7	6	0.0	0	0.030	0.066	0.651	0.0	default_60.par	D13en7
8	2793116613	2020-03-11	1930.13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D13en8
9	2793116699	2020-03-11	1931.39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D13en9
10	2793116770	2020-03-11	1932.50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D13en10
11	2793116844	2020-03-11	1934.04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D13en11
12	2793117043	2020-03-11	1937.23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D13en12
13	2793117112	2020-03-11	1938.32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D13en13
14	2793117173	2020-03-11	1939.33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D13en14
15	2793117247	2020-03-11	1940.47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D13en15

Data  
wrangling

Data analysis

R script

## Published results



2. Reproducible data manipulation:  
use R from the very beginning of your work with data  
(not just for the stats and plots)!

# 1. Fundamentals of data management

Get your data in order **before** you start working on it ...





# Why this matters

- Ever looked at an **old project** and panicked because you could not make sense of what was what?
- Ever struggled to make sense of **data** that you **received from a collaborator**?



Panic,  
Frustration,  
and **bad science** ...

# Good data management = good data analysis

Projects generate a lot of:

- data (= stuff that you measured) and
- metadata (= additional information necessary to make sense of your data).

These come in

- different shapes and formats
- from different points in time,
- from different sources
- and likely get modified in multiple occasions.

→ things can get real messy ...

Messiness can hinder your ability to fully utilize your data, vanishing the efforts you made to generate that data!

Therefore, you need to set and follow rules ...

# Practically speaking:

## How to organize your project folder

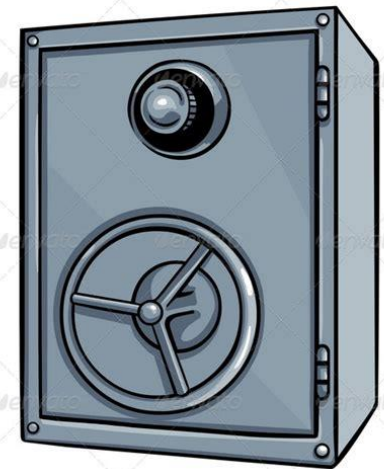
- **All** data belonging to a project must be stored **in one folder** (there you can have as many sub-folders as it is needed)
- Have a **written description** of how the data was generated and by who in a text file. It must be as clear and detailed as possible. Let's call it an “ultra-honest version of a papers' materials and methods”
- For each data set or table: provide a **clear explanation** of the meaning of **each variable** (column name) as well as of each observation (row). It must be clear what each name or measurement or attribute stands for (especially if it is an abbreviation) and in what **units** it is expressed
  - e.g. if a column named “DO”, write somewhere that it means “Dissolved Oxygen [mg/L]” ...
  - This can be for example in another sheet within the same .xlsx file, or in a separate .txt or .csv file named in a way that can be unambiguously linked to the data it refers to (e.g. “rawdata\_shrimp\_20201109.csv” and “rawdata\_shrimp\_20201109\_description.txt”)

# Practically speaking: How to organize your project folder

- **Keep the raw / original data safe!**

Meaning:

- name it clearly and unambiguously
- **never modify** it
- make a **backup** copy (or two or three ... ) saved **somewhere else**.





# Some definitions to clarify ...

**Original data** = the very original data that was collected: your hand notes if you have manual measurements (take pictures or make a scan copy of it) or what is outputted by a machine if you have automated or digital measurements. **AS IT IS.**

**Raw data** = basically, the same as original data, but for data analysis it should be in a digital format (what you manually enter in a spreadsheet).

**Manipulations** = include “cleaning” of the data to remove wrong or incomplete information, correction of typos, calculation of statistics (mean, sd or so) or reshaping of the data for preliminary plots.

If you manipulate it, you cannot call it “raw” or “original” anymore, therefore rename it!

# How to name your files in a meaningful way

**Choose a naming system and stick to it.**

Useful information to include in the filename are:

- The **date** of creation in the filename: “cmr\_intro\_20210104.R”  
(using the format YYYY-MM-DD is particularly handy because the files will **automatically** be **sorted chronologically**)
- The **version** number: “thesis\_intro\_v1.R”, “thesis\_intro\_v2.R”, “thesis\_intro\_v2.1.R”  
(but in this way it can be tricky to keep track of the actual version when you create a new one – I prefer using the date)
- Your **initials** when you modify a file shared among collaborators: “thesis\_intro\_20210104\_GP.R”, “cmr\_intro\_20210104\_GP\_ABC.R”

In this way, even if you have a million files at the end of the project you can always **track the changes chronologically** and arrive to the source ...

(but if you work on data, you should only modify it through scripts – not in Excel!)

# How to tell if your data is well kept: a simple rule ...

Ask yourself if **another person**, by looking at your data or project folder, without any additional input, **would be able to figure out**:

- **Where** does this **data come from**: how was it generated? What did you measure and how? **Who** generated it?
- **Why**, what is the rationale and what are the reason for specifically choosing that approach
- In each data table: **what** is the **meaning of each variable** (= column name): clearly define them as explicitly as possible (even if -you think- the naming is self-explanatory!)
- The **structure** of the data set is clear: **how many** treatments, how many levels, number of replicates ...



# Additional tips to simplify your (and everyone's) life

**Characters choice** – for filenames and for their content (aka your data)

Avoid anything that can have multiple interpretations.

- Avoid **punctuation** (., : ; ? !)
- avoid **special symbols** (e.g. instead of “%”, write “perc”)
- avoid **operational symbols** (– + / \* ^ < >)
- never start with a **number** (“15thattemp.csv” is not good, rather: “attempt15.csv”)
- Also good practice to leave **no space** in the filename

(can be problem if in other OS, e.g. Linux, plus, in all coding languages I can think of, words separated by space are interpreted as separated objects)



Mario Rossi Data coral expm#3.xlsx



1st version? 02.12.2019.csv



# Additional tips to simplify your (and everyone's) life

- Try as much as possible to **generate the data directly in ENGLISH**
- Try as much as possible to use the **English numerical system** (decimals as “.” not as “,”)
- Save your original data as **.csv** (or .tsv or even plain .txt), NOT as Excel spreadsheet
  - because .csv are universally read in the same way, while Excel spreadsheet read differently depending on the version and system locale



# To sum up ...

- Keep a complete and detailed written explanation of your data set together with your data (e.g. in the same folder)
- **Never touch the raw/original data** and keep at least one backup copy of it
- Be smart and just avoid anything that might create ambiguity (naming etc.)
- Make sure your work is ALWAYS REPRODUCIBLE = **modify your data only through scripts!** (yes, even corrections to obvious mistakes like typos)



## 2. Reproducible data manipulation: data wrangling in R (Tidyverse)

Work only through scripts → R !

# Re-cap of what “data wrangling” means

Transform raw data into another format that is more suited for downstream applications (e.g., analytics).

Includes:

- Sorting/re-arranging of data (change order, transpose)
- Sub setting (separate a smaller part of a dataset from the rest -> e.g., to plot, remove outliers)
- Merging (put together data from different tables/sheets)
- Correcting errors (e.g., typos)

Typically **takes more time than the actual analysis** of the data!

# Tidyverse



A **collection of R packages** designed **for data science**, that share an underlying design philosophy, grammar, and data structures.

“A gateway drug ... “

Noteworthy aspects:

1. Concept of “**Tidy data**”
2. The **pipe** (`%>%`) (move away from nested functions)





# 1. Tidy data

Simple rules:

- Every **column** is **variable**.
- Every **row** is an **observation**.
- Every **cell** is a **single value**.

country	year	cases	population
Afghanistan	1999	15	199871
Afghanistan	2000	566	2005360
Brazil	1999	3737	17206362
Brazil	2000	8488	17404898
China	1999	21258	127215272
China	2000	21796	12802583

variables

country	year	cases	population
Afghanistan	1999	15	199871
Afghanistan	2000	566	2005360
Brazil	1999	3737	17206362
Brazil	2000	8488	17404898
China	1999	21258	127215272
China	2000	21796	12802583

observations

country	year	cases	population
Afghanistan	1999	15	199871
Afghanistan	2000	566	2005360
Brazil	1999	3737	17206362
Brazil	2000	8488	17404898
China	1999	21258	127215272
China	2000	21796	12802583

values

Happy families are all alike; every  
unhappy family is unhappy in its own  
way.

Leo Tolstoy

[12 Tidy data | R for Data Science \(had.co.nz\)](#)

# 1. Tidy data

Simple rules:

- Every **column** is **variable**.
- Every **row** is an **observation**.
- Every **cell** is a **single value**.

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	17206362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

variables

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	17206362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

observations

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	17206362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

values

Happy families are all alike; every  
unhappy family is unhappy in its own  
way.

Leo Tolstoy

country	year	rate
Afghanistan	1999	745 / 19987071
Afghanistan	2000	2666 / 20595360
Brazil	1999	37737 / 172006362
Brazil	2000	80488 / 174504898
China	1999	212258 / 1272915272
China	2000	213766 / 1280428583

[12 Tidy data | R for Data Science \(had.co.nz\)](#)

# 1. Tidy data

Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy

Simple rules:

- Every **column** is **variable**.
- Every **row** is an **observation**.
- Every **cell** is a single **value**.

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

variables

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

observations

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

values

[12 Tidy data | R for Data Science \(had.co.nz\)](#)

country	year	rate
Afghanistan	1999	745 / 19987071
Afghanistan	2000	2666 / 20595360
Brazil	1999	37737 / 172006362
Brazil	2000	80488 / 174504898
China	1999	212258 / 1272915272
China	2000	213766 / 1280428583

✗

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

✓

[12 Tidy data | R for Data Science \(had.co.nz\)](#)

# 1. Tidy data

Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy

Simple rules:

- Every **column** is **variable**.
- Every **row** is an **observation**.
- Every **cell** is a single **value**.

country	year	cases	population
Afghanistan	2000	2566	20195360
Brazil	1999	30737	172096362
Brazil	2000	80488	174004898
China	1999	210258	1272015272
China	2000	210706	128025583

variables

country	year	cases	population
Afghanistan	2000	2566	20195360
Afghanistan	2000	2566	20195360
Brazil	1999	30737	172096362
Brazil	1999	30737	172096362
Brazil	2000	80488	174004898
Brazil	2000	80488	174004898
China	1999	210258	1272015272
China	1999	210258	1272015272
China	2000	210706	128025583
China	2000	210706	128025583

observations

country	year	cases	population
Afghanistan	2000	2566	20195360
Afghanistan	2000	2566	20195360
Brazil	1999	30737	172096362
Brazil	1999	30737	172096362
Brazil	2000	80488	174004898
Brazil	2000	80488	174004898
China	1999	210258	1272015272
China	1999	210258	1272015272
China	2000	210706	128025583
China	2000	210706	128025583

values

<sup>12</sup> Tidy data | R for Data Science (had.co.nz)

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

# 1. Tidy data

Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy

Simple rules:

- Every **column** is **variable**.
- Every **row** is an **observation**.
- Every **cell** is a single **value**.

country	year	cases	population
Alghanistan	2000	2566	20005360
Brazil	1999	30737	172006362
Brazil	2000	80488	174004898
China	1999	210258	1272005272
China	2000	210706	1280005583

variables

country	year	cases	population
Alghanistan	2000	2566	20005360
Alghanistan	2000	2566	20005360
Brazil	1999	30737	172006362
Brazil	1999	30737	172006362
Brazil	2000	80488	174004898
Brazil	2000	80488	174004898
China	1999	210258	1272005272
China	1999	210258	1272005272
China	2000	210706	1280005583
China	2000	210706	1280005583

observations

country	year	cases	population
Alghanistan	2000	2566	20005360
Alghanistan	2000	2566	20005360
Brazil	1999	30737	172006362
Brazil	1999	30737	172006362
Brazil	2000	80488	174004898
Brazil	2000	80488	174004898
China	1999	210258	1272005272
China	1999	210258	1272005272
China	2000	210706	1280005583
China	2000	210706	1280005583

values

[12 Tidy data | R for Data Science \(had.co.nz\)](#)

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

“wide”

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

“wide”

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

“long”

## 2. Piping functions (vs nested)

Nested functions (base R)

3

2

1

```
length(unique(data$column))
```

Using the pipe makes the code easier to write and to read (Tidyverse)

1

2

3

```
data$column %>% unique() %>% length()
```





# A set of powerful and intuitive functions

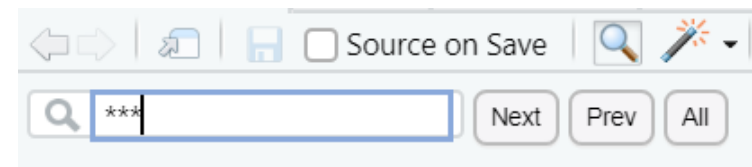
We'll see them in action in a minute ... !

But briefly:

- Create/modify variables: `mutate()`
- Subset data: `filter()`, `select()`
- Summarize: `summarise()`, `group_by()`
- Restructure: `pivot_wider()`, `pivot_longer()`
- Merge: `*_join()` ... (e.g. `left_join()` )
- Correct values: `rename()`, `replace()`
- Plot like a pro: `ggplot()`

↖ These can take you a long way ... !

# \*\*\*  
marked in the script  
-> easy to find



... insanely better than any spreadsheet



# Ok enough talking ... let's get coding!

all material available in GitHub to run at your own peace after the workshop



# Kidding, one last thing ...

We'll go through an example of exploratory data analysis (EDA\*)

- **Explore** structure of the data (understand it)
- Scout for **mistakes** (always present in real world data)
- **Correct** such mistakes **exclusively through R**
- **Re-shape** the data (tables) to our needs - Prepare for stats (e.g. ANOVA)
- Make some **plots** (show the power of exploratory data viz! now because if earlier might be confusing)

\*use this terminology in your CV to sound more profesh ;)

# Overview of the data set

The R script is based on **dummy data** (created for didactic purposes by me) that includes:

- PAM data, collected in 4 separate sessions + additional data regarding the experimental conditions → 5 files (**.csv**)
- “Metadata” with description of the dataset → text (**.txt**)
- Graphical representation of the experimental design → figure (**.png**)  
*One ~~picture~~ figure is worth a thousand words ... even for scientists!*

This is a **realistic example** of how to correctly store the raw/original data with the info necessary to understand it, and how to process/manipulate/wrangle it in R



Graphical representation of the experimental design for our dummy data (“PAM\_replication.png”)

# Also keep in mind that ...

**Learning R coding is a hands-on effort:** you **will never learn R just by watching** (a lecture, a tutorial ... ) and memorizing. The only way is to try things for yourself: write (or copy paste) the code, read packages and functions descriptions, follow similar examples.

**It's a lot of reading, and trial and error.**

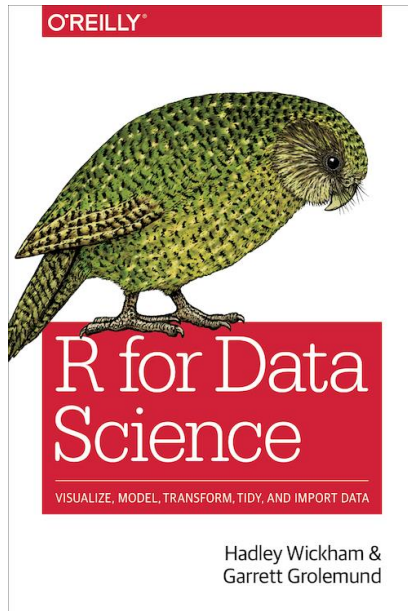
This can be discouraging at the beginning, but it is time well invested!

Therefore this script is **not** to be taken as **an explanation of how to do everything** – although it is heavily commented for ease of understanding – the idea is for **the script to be a starting point (a roadmap)** to show **examples of what R can do for you**. Then you will have to take the time to explore the functions by yourself (tip: use the “help” panel to check how the function works, what arguments it requires and in what shape ... ).



# (my favorite) Resources

THE book,  
a must-read: “R4DS”



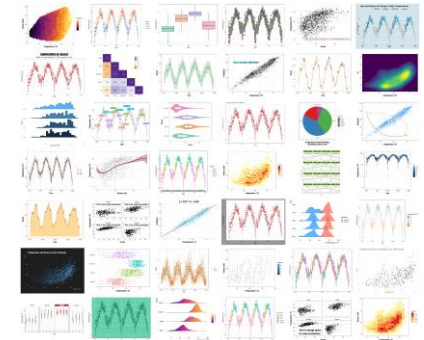
<https://r4ds.had.co.nz/>



*Riffomonas*



To follow on Twitter  
#Rstats  
#tidyverse  
@rfunctionaday (R Function A Day)



Tutorials and inspiration for ggplot2

[From data to Viz](#) | Find the graphic you need ([data-to-viz.com](http://data-to-viz.com))

[A ggplot2 Tutorial for Beautiful Plotting in R](#) - Cédric Scherer ([cedricscherer.com](http://cedricscherer.com))

[rfordatascience/tidytuesday](#): Official repo for the [#tidytuesday](#) project ([github.com](https://github.com))