



for **R**eproducible data wrangling

What happens before the stats:
the power of **R** Tidyverse for
wrangling, cleaning, and exploring your data

Giulia Puntin |  @sPuntinGi

R is for “Reproducible”


Analyze the same data and obtain the same results

Data

#	t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample
1	3793116090	2020-03-11	19:21:00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D19en1
2	3793116153	2020-03-11	19:22:33	2	6	0.0	0	0.306	0.695	0.620	0.0	default_60.par	D19en2
3	3793116259	2020-03-11	19:23:55	3	6	0.0	0	0.367	1.094	0.645	0.0	default_60.par	D19en3
4	3793116380	2020-03-11	19:24:43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par	D19en4
5	3793116327	2020-03-11	19:25:27	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D19en5
6	3793116387	2020-03-11	19:26:27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D19en6
7	3793116522	2020-03-11	19:28:42	7	6	0.0	0	0.030	0.066	0.651	0.0	default_60.par	D19en1
8	3793116613	2020-03-11	19:30:13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D19en2
9	3793116699	2020-03-11	19:31:39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D19en3
10	3793116770	2020-03-11	19:32:50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D19en4
11	3793116844	2020-03-11	19:34:04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D19en5
12	3793117043	2020-03-11	19:37:23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D19en6
13	3793117112	2020-03-11	19:38:32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D19en7
14	3793117173	2020-03-11	19:39:33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D19en8
15	3793117247	2020-03-11	19:40:47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D19en1

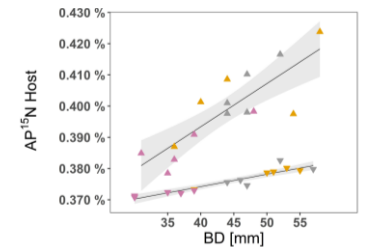


Data
analysis



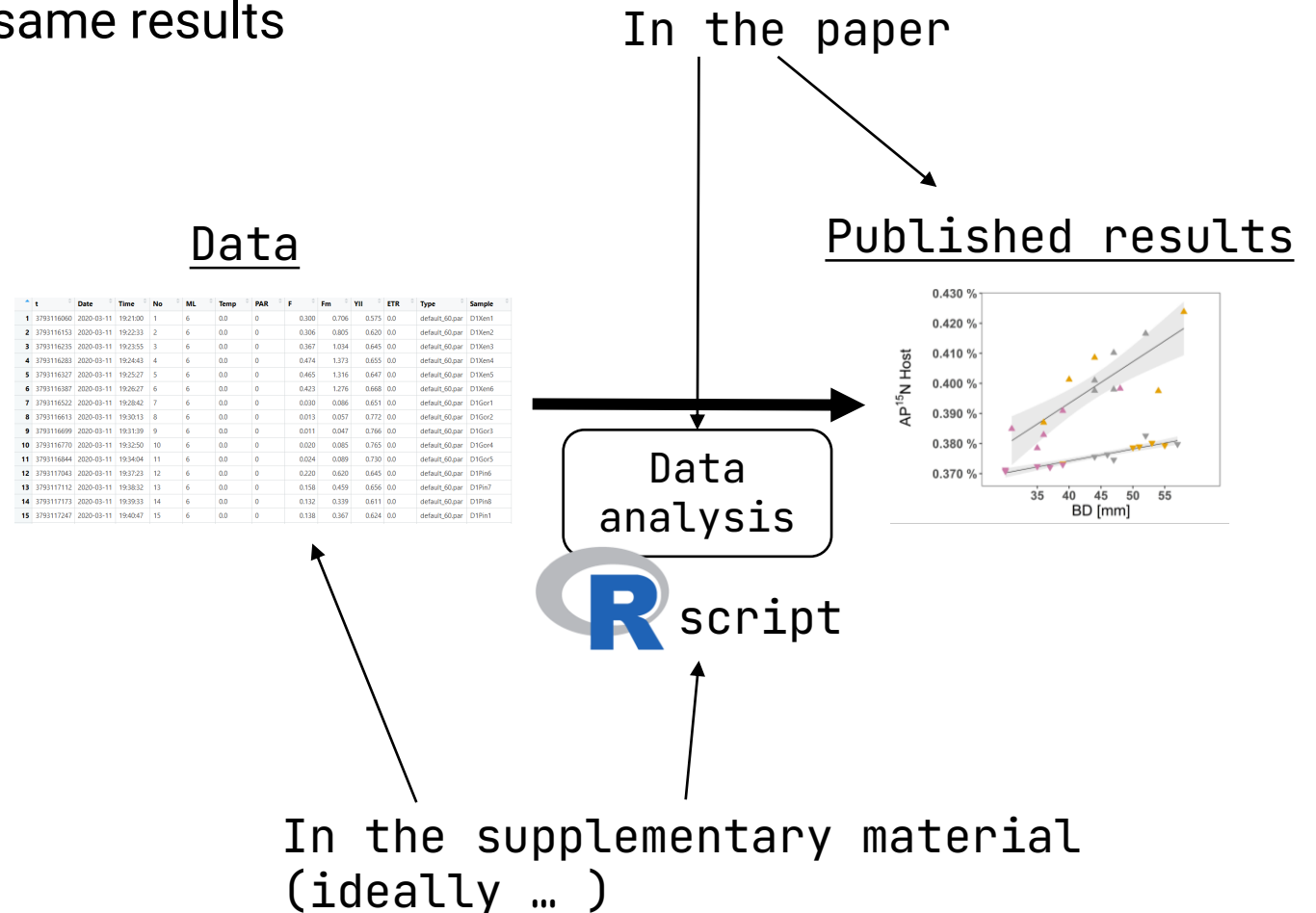
R script

Published results



R is for “Reproducible”

Analyze the same data and obtain the same results



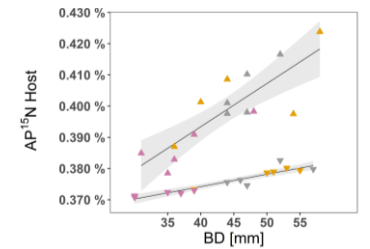
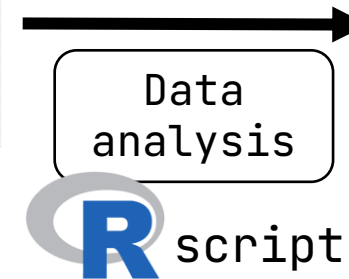
R is for “Reproducible”

But there's also that other part ...

Data

#	t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample
1	3793116090	2020-03-11	19:21:00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D19en1
2	3793116153	2020-03-11	19:22:33	2	6	0.0	0	0.306	0.695	0.620	0.0	default_60.par	D19en2
3	3793116205	2020-03-11	19:23:55	3	6	0.0	0	0.367	1.094	0.645	0.0	default_60.par	D19en3
4	3793116280	2020-03-11	19:24:43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par	D19en4
5	3793116327	2020-03-11	19:25:07	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D19en5
6	3793116387	2020-03-11	19:26:27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D19en6
7	3793116522	2020-03-11	19:28:42	7	6	0.0	0	0.030	0.066	0.651	0.0	default_60.par	D19en1
8	3793116613	2020-03-11	19:30:13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D19en2
9	3793116699	2020-03-11	19:31:39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D19en3
10	3793116770	2020-03-11	19:32:50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D19en4
11	3793116844	2020-03-11	19:34:04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D19en5
12	3793117043	2020-03-11	19:37:23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D19en6
13	3793117112	2020-03-11	19:38:32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D19en7
14	3793117173	2020-03-11	19:39:33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D19en8
15	3793117247	2020-03-11	19:40:47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D19en1

Published results



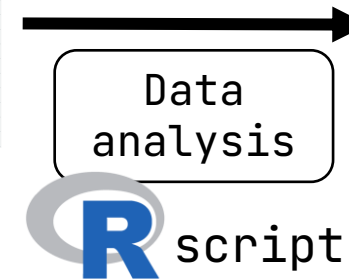
R is for “Reproducible”

But there's also that other part ...

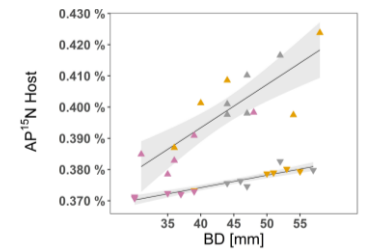
Cleaned data
ready for stats

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	D19en1
2	3793116153	2020-03-11	19:22:33	2	6	0.0	0	0.306	0.805	0.620	0.0	default_60.par	D19en2
3	3793116295	2020-03-11	19:23:55	3	6	0.0	0	0.367	1.094	0.845	0.0	default_60.par	D19en3
4	3793116380	2020-03-11	19:24:43	4	6	0.0	0	0.474	1.373	0.853	0.0	default_60.par	D19en4
5	3793116327	2020-03-11	19:25:07	5	6	0.0	0	0.465	1.316	0.847	0.0	default_60.par	D19en5
6	3793116387	2020-03-11	19:26:27	6	6	0.0	0	0.423	1.276	0.868	0.0	default_60.par	D19en6
7	3793116522	2020-03-11	19:28:42	7	6	0.0	0	0.030	0.086	0.651	0.0	default_60.par	D19er1
8	3793116613	2020-03-11	19:30:13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D19er2
9	3793116699	2020-03-11	19:31:39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D19er3
10	3793116770	2020-03-11	19:32:50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D19er4
11	3793116844	2020-03-11	19:34:04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D19er5
12	3793117043	2020-03-11	19:37:23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D19H6
13	3793117112	2020-03-11	19:38:32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D19H7
14	3793117173	2020-03-11	19:39:33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D19H8
15	3793117247	2020-03-11	19:40:47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D19H9



Published results



R is for “Reproducible”

Original data



+

T	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Time	Temp	ML	WEL	Temp	PAR	F	Fm	YII	ETR			
2	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
3	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
4	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
5	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
6	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
7	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
8	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
9	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
10	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
11	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
12	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
13	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
14	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
15	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
16	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
17	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
18	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
19	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
20	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
21	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
22	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
23	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
24	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
25	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
26	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
27	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
28	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
29	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
30	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
31	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
32	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
33	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
34	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
35	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
36	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
37	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
38	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
39	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
40	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
41	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
42	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
43	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
44	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
45	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
46	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
47	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
48	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
49	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
50	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
51	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
52	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
53	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
54	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
55	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
56	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
57	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
58	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
59	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			
60	19:00:00	19.01	0.00	0.00	6	0	0	0.0	0.706	0.575			

original
outputs

hand notes

digitalize


																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																				</
--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	----

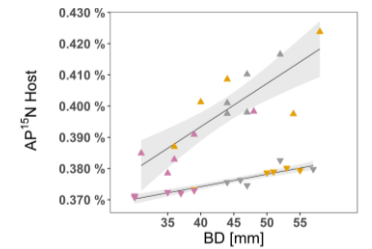
Data
wrangling

Cleaned data
ready for stats

t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample
1	2020-03-11	1921.00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D19m1
2	2020-03-11	1922.33	2	6	0.0	0	0.306	0.805	0.620	0.0	default_60.par	D19m2
3	2020-03-11	1923.55	3	6	0.0	0	0.367	1.094	0.845	0.0	default_60.par	D19m3
4	2020-03-11	1924.43	4	6	0.0	0	0.474	1.373	0.853	0.0	default_60.par	D19m4
5	2020-03-11	1925.07	5	6	0.0	0	0.465	1.316	0.847	0.0	default_60.par	D19m5
6	2020-03-11	1926.27	6	6	0.0	0	0.423	1.276	0.868	0.0	default_60.par	D19m6
7	2020-03-11	1928.42	7	6	0.0	0	0.030	0.086	0.651	0.0	default_60.par	D19m7
8	2020-03-11	1930.13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D19m8
9	2020-03-11	1931.39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D19m9
10	2020-03-11	1932.50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D19m10
11	2020-03-11	1934.04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D19m11
12	2020-03-11	1937.23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D19m12
13	2020-03-11	1938.32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D19m13
14	2020-03-11	1939.33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D19m14
15	2020-03-11	1940.47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D19m15

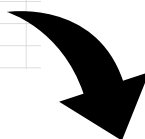
Published results

Data
analysis
 script



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



R is for “Reproducible”

Original data



+

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	0.695	0.620	0.0	default_60.par	D19en2
3	2020-03-11	19:23:55	3	6	0.0	0	0.367	1.094	0.645	0.0	default_60.par	D19en3
4	2020-03-11	19:24:43	4	6	0.0	0	0.474	1.373	0.655	0.0	default_60.par	D19en4
5	2020-03-11	19:25:27	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

original
outputs

hand notes

digitalize

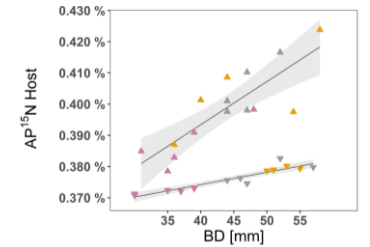
Digitalized data			
Date	Time	No	Value
2020-03-11	19:21:00	1	0.0
2020-03-11	19:22:33	2	0.0
2020-03-11	19:23:55	3	0.0
2020-03-11	19:24:43	4	0.0
2020-03-11	19:25:27	5	0.0
2020-03-11	19:26:27	6	0.0
2020-03-11	19:28:42	7	0.0
2020-03-11	19:30:13	8	0.0
2020-03-11	19:31:39	9	0.0
2020-03-11	19:32:50	10	0.0
2020-03-11	19:34:04	11	0.0
2020-03-11	19:37:23	12	0.0
2020-03-11	19:38:32	13	0.0
2020-03-11	19:39:33	14	0.0
2020-03-11	19:40:47	15	0.0

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	0.695	0.620	0.0	default_60.par	D19en2
3	2020-03-11	19:23:55	3	6	0.0	0	0.367	1.094	0.645	0.0	default_60.par	D19en3
4	2020-03-11	19:24:43	4	6	0.0	0	0.474	1.373	0.655	0.0	default_60.par	D19en4
5	2020-03-11	19:25:27	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



Data
analysis

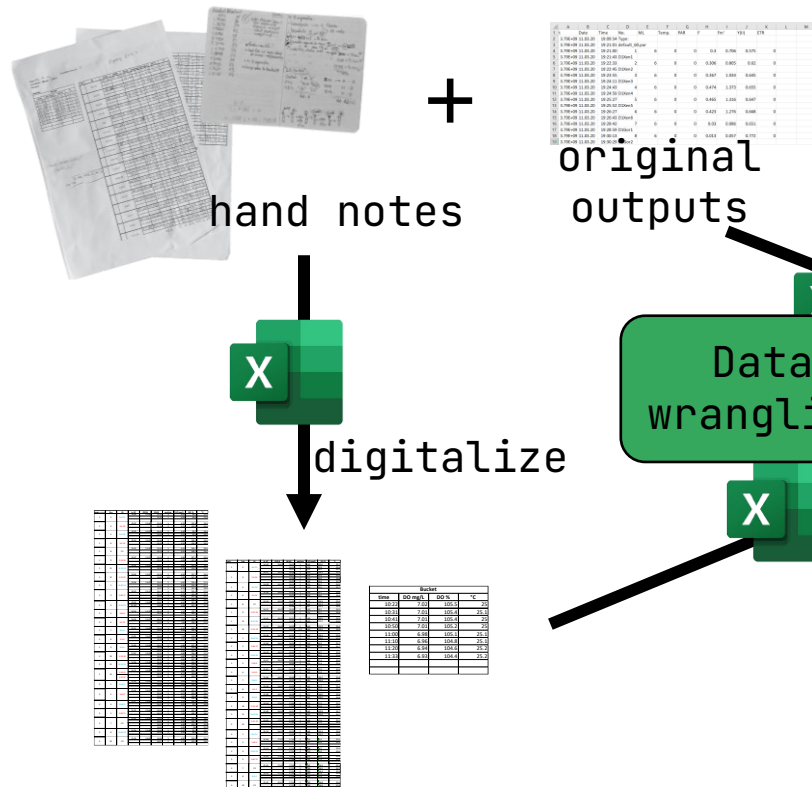
R script

Often overlooked/underestimated:

- Not systematically taught (missing from typical R courses)
⇒ badly done (= NOT reproducible)
- Time consuming (can take up as more time than stat testing)
⇒ **sensitive step** → room for improvement!

R is for “Reproducible”

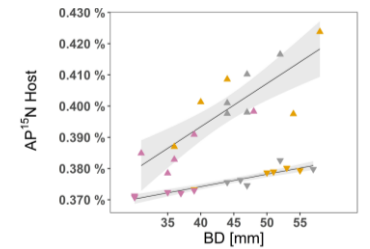
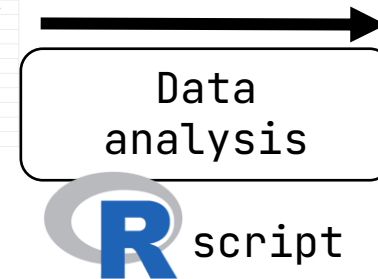
Original data



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	3793116382	2020-03-11	1924.43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par D19en4
5	3793116387	2020-03-11	1925.07	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	3793116552	2020-03-11	1928.42	7	6	0.0	0	0.030	0.066	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

Published results



Beware of spreadsheets ...

Problems derived from working with spreadsheets:

- Messy (many files) ...
- **Error prone** (e.g. genuine mistakes + autocorrection ...)
- **Not scalable** (it just doesn't work with large data sets)
- **Not reproducible** (hard to keep track of every action)



John Feminella @jxxf · 23h

Optimist: The glass is ½ full.

Pessimist: The glass is ½ empty.

Excel: The glass is January 2nd.

70

4,241

40K



nature

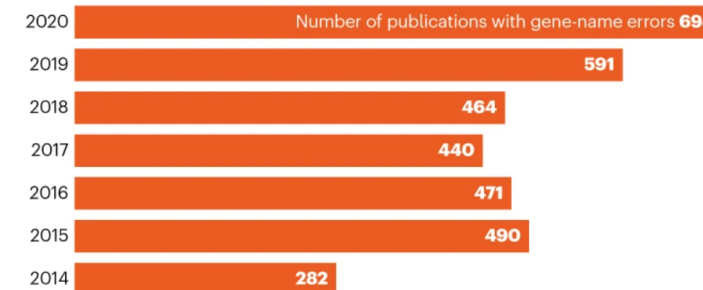
NEWS | 13 August 2021 | Correction [25 August 2021](#)

Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

A GROWING PROBLEM

A 2016 analysis found that 20% of papers featuring gene names had errors created by spreadsheet autocorrect functions, but a bigger survey now finds the proportion is up to 30%. Since 2014, the number of papers with errors has increased significantly.



©nature

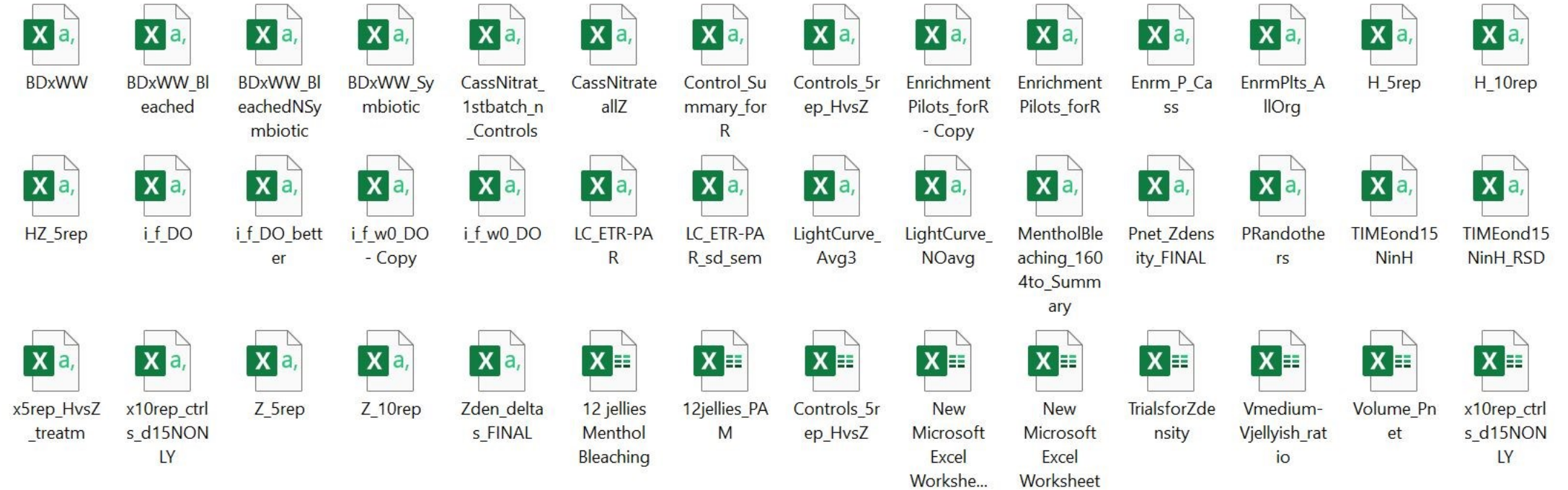
Beware of spreadsheets ...

Problems derived from working with spreadsheets:

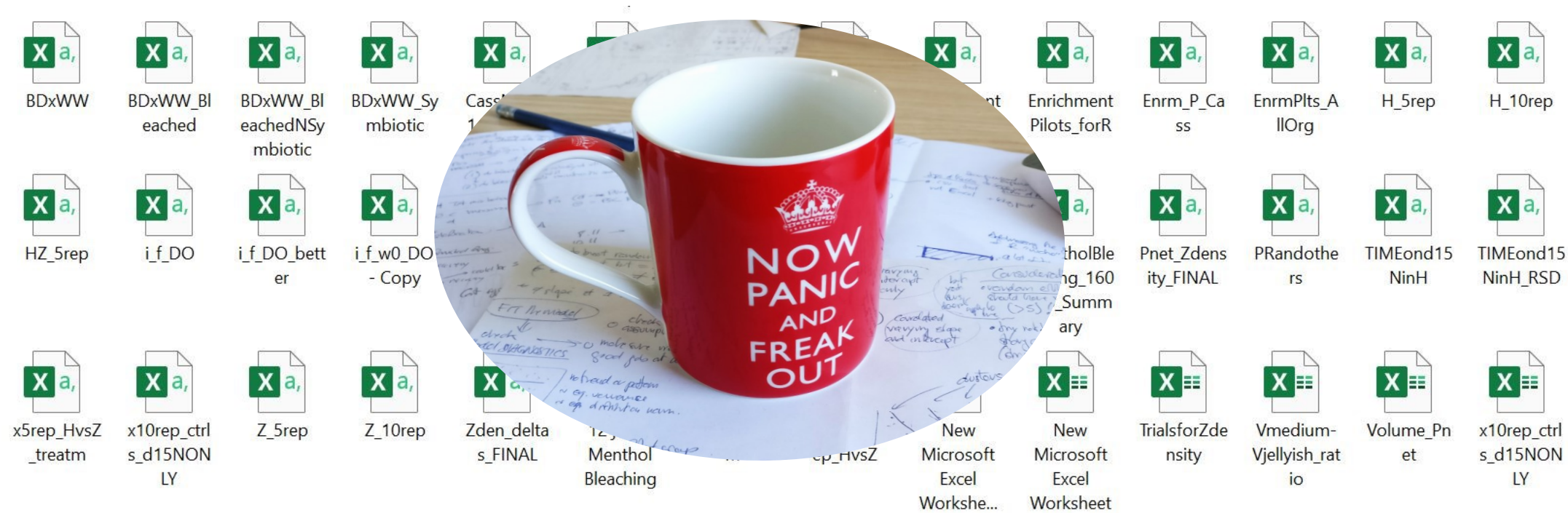
- Messy (many files) ...
- **Error prone** (e.g. genuine mistakes + autocorrection ...)
- **Not scalable** (it just doesn't work with large data sets)
- **Not reproducible** (hard to keep track of every action)

On the contrary, 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)

Old me before discovering Tidyverse ...



Old me before discovering Tidyverse ...



R is for “Reproducible”

Original data



+

1	A	B	C	D	E	F	G	H	I	J	K	L	M
2	Time	Temp	ML	CO ₂	Temp	PAR	F	Fm	Fv	YII	ETR		
3	19:00:00	18.00	0	1921.00	18.00	0	0	0	0.4	0.706	0.575		
4	19:00:00	18.00	2	1922.33	18.00	0	0	0	0.306	0.805	0.620		
5	19:00:00	18.00	3	1923.55	18.00	0	0	0	0.367	1.054	0.845		
6	19:00:00	18.00	4	1924.43	18.00	0	0	0	0.474	1.373	0.853		
7	19:00:00	18.00	5	1925.07	18.00	0	0	0	0.465	1.316	0.847		
8	19:00:00	18.00	6	1926.27	18.00	0	0	0	0.423	1.276	0.868		
9	19:00:00	18.00	7	1928.42	18.00	0	0	0	0.030	0.086	0.651		
10	19:00:00	18.00	8	1930.13	18.00	0	0	0	0.013	0.057	0.772		
11	19:00:00	18.00	9	1931.39	18.00	0	0	0	0.011	0.047	0.766		
12	19:00:00	18.00	10	1932.50	18.00	0	0	0	0.020	0.085	0.765		
13	19:00:00	18.00	11	1934.04	18.00	0	0	0	0.024	0.089	0.730		
14	19:00:00	18.00	12	1937.23	18.00	0	0	0	0.220	0.620	0.645		
15	19:00:00	18.00	13	1938.32	18.00	0	0	0	0.158	0.459	0.656		
16	19:00:00	18.00	14	1939.33	18.00	0	0	0	0.132	0.339	0.611		
17	19:00:00	18.00	15	1940.47	18.00	0	0	0	0.138	0.367	0.624		

original
outputs

hand notes

digitalize

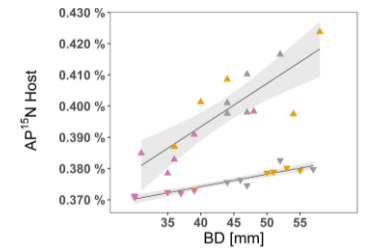
Data
wrangling

Cleaned data
ready for stats

Published results

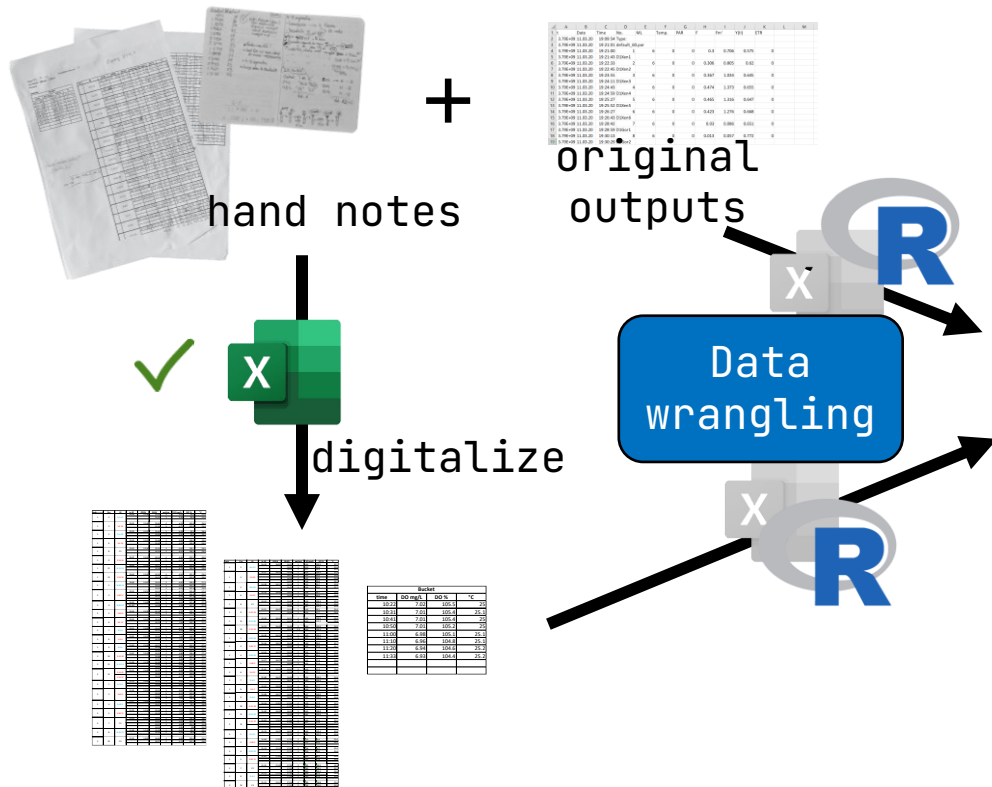
t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample	
1	1793116000	2020-03-11	1921.00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D19m1
2	1793116153	2020-03-11	1922.33	2	6	0.0	0	0.306	0.805	0.620	0.0	default_60.par	D19m2
3	1793116295	2020-03-11	1923.55	3	6	0.0	0	0.367	1.054	0.845	0.0	default_60.par	D19m3
4	1793116380	2020-03-11	1924.43	4	6	0.0	0	0.474	1.373	0.853	0.0	default_60.par	D19m4
5	1793116387	2020-03-11	1925.07	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D19m5
6	1793116387	2020-03-11	1926.27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D19m6
7	1793116522	2020-03-11	1928.42	7	6	0.0	0	0.030	0.086	0.651	0.0	default_60.par	D19m7
8	1793116613	2020-03-11	1930.13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D19m8
9	1793116699	2020-03-11	1931.39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D19m9
10	1793116770	2020-03-11	1932.50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D19m10
11	1793116844	2020-03-11	1934.04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D19m11
12	1793117043	2020-03-11	1937.23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D19m12
13	1793117112	2020-03-11	1938.32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D19m13
14	1793117173	2020-03-11	1939.33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D19m14
15	1793117247	2020-03-11	1940.47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D19m15

Data
analysis



R is for “Reproducible”

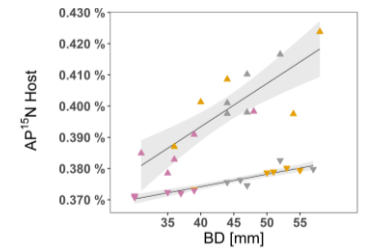
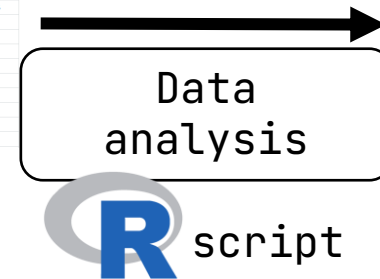
Original data



Cleaned data
ready for stats

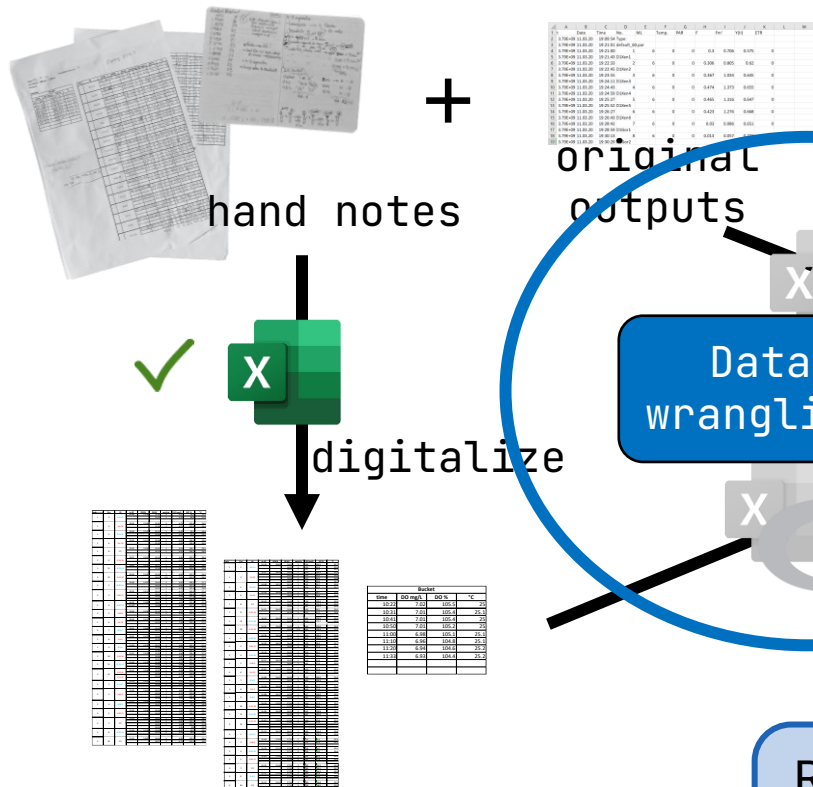
t	Date	Time	No	ML	Temp	PAR	F	Fm	YII	ETR	Type	Sample	
1	1793116000	2020-03-11	1921.00	1	6	0.0	0	0.300	0.706	0.575	0.0	default_60.par	D19en1
2	1793116153	2020-03-11	1922.33	2	6	0.0	0	0.306	0.695	0.620	0.0	default_60.par	D19en2
3	1793116295	2020-03-11	1923.55	3	6	0.0	0	0.367	1.054	0.645	0.0	default_60.par	D19en3
4	1793116380	2020-03-11	1924.43	4	6	0.0	0	0.474	1.373	0.653	0.0	default_60.par	D19en4
5	1793116387	2020-03-11	1925.07	5	6	0.0	0	0.465	1.316	0.647	0.0	default_60.par	D19en5
6	1793116387	2020-03-11	1926.27	6	6	0.0	0	0.423	1.276	0.668	0.0	default_60.par	D19en6
7	1793116522	2020-03-11	1928.42	7	6	0.0	0	0.030	0.066	0.651	0.0	default_60.par	D19en7
8	1793116613	2020-03-11	1930.13	8	6	0.0	0	0.013	0.057	0.772	0.0	default_60.par	D19en8
9	1793116699	2020-03-11	1931.39	9	6	0.0	0	0.011	0.047	0.766	0.0	default_60.par	D19en9
10	1793116770	2020-03-11	1932.50	10	6	0.0	0	0.020	0.085	0.765	0.0	default_60.par	D19en10
11	1793116844	2020-03-11	1934.04	11	6	0.0	0	0.024	0.089	0.730	0.0	default_60.par	D19en11
12	1793117043	2020-03-11	1937.23	12	6	0.0	0	0.220	0.620	0.645	0.0	default_60.par	D19en12
13	1793117112	2020-03-11	1938.32	13	6	0.0	0	0.158	0.459	0.656	0.0	default_60.par	D19en13
14	1793117173	2020-03-11	1939.33	14	6	0.0	0	0.132	0.339	0.611	0.0	default_60.par	D19en14
15	1793117247	2020-03-11	1940.47	15	6	0.0	0	0.138	0.367	0.624	0.0	default_60.par	D19en15

Published results



R is for “Reproducible”

Original data



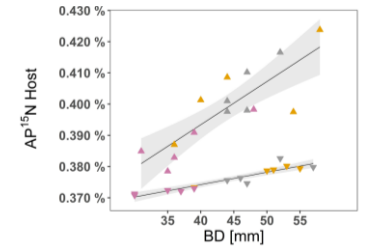
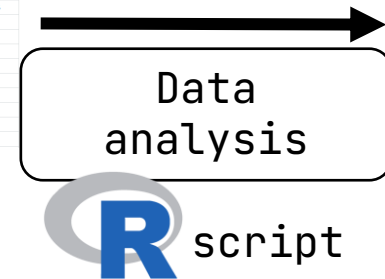
original
outputs

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	1.094	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	0.066	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



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

Tidyverse



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

“A gateway drug.”
- my friend James

Noteworthy aspects:

1. Concept of “**Tidy data**”
2. The **pipe** (`%>%`) (more human readable than nested functions)



1. Tidy data

Simple rules:

- Every **column** is a **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	127201272
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	127201272
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	127201272
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 a **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

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

Leo Tolstoy

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

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

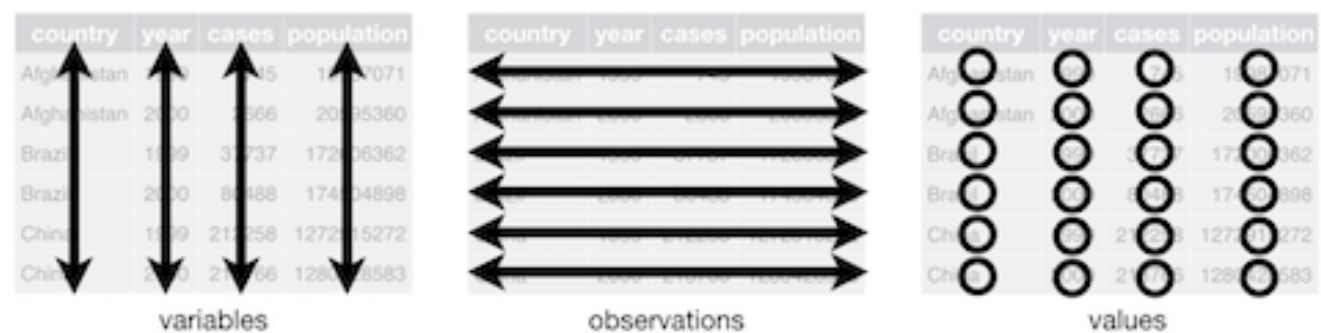
Happy families are all alike; every
unhappy family is unhappy in its own
way.

Leo Tolstoy

1. Tidy data

Simple rules:

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



[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 a **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

	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

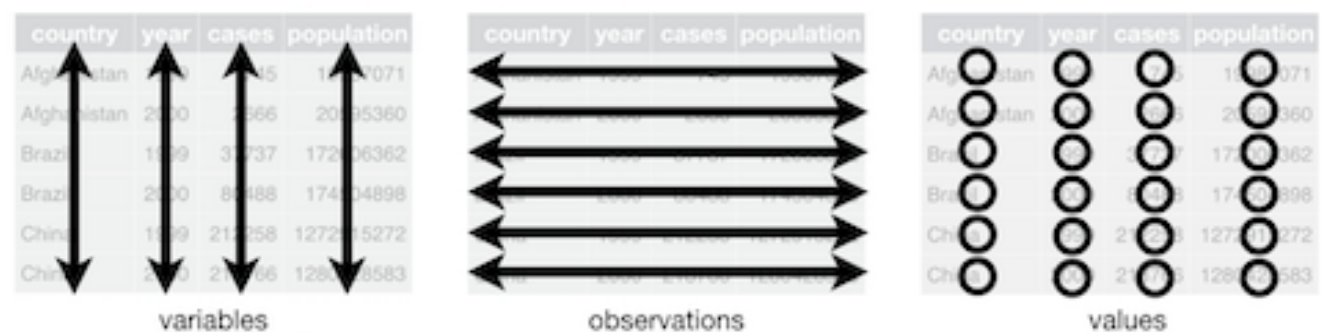
Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy

1. Tidy data

Simple rules:

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



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”

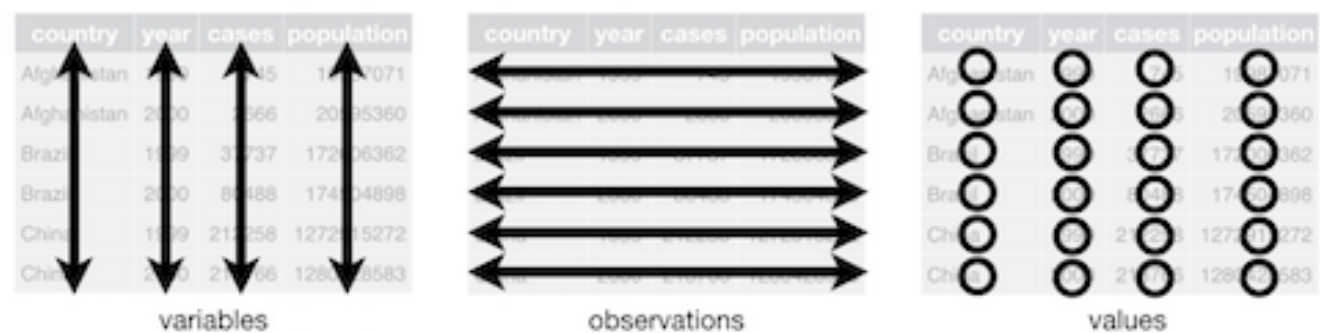
Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy

1. Tidy data

Simple rules:

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



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

Only column names

Row names

Column names

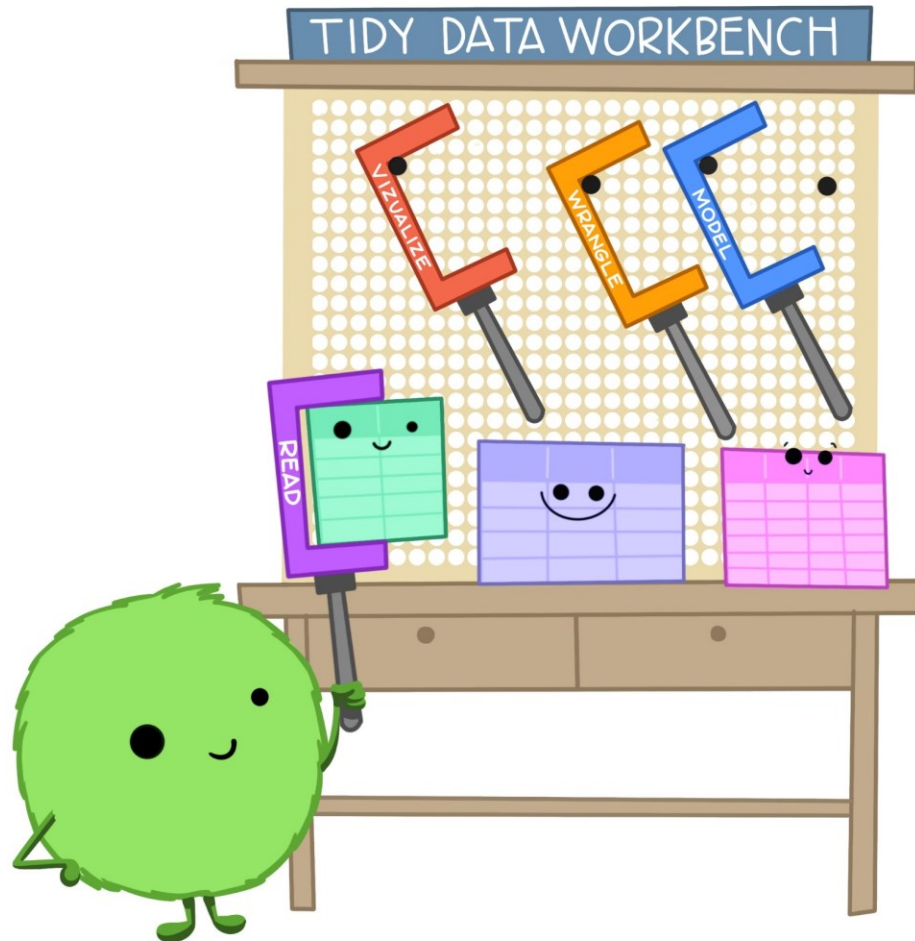
	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Not tidy

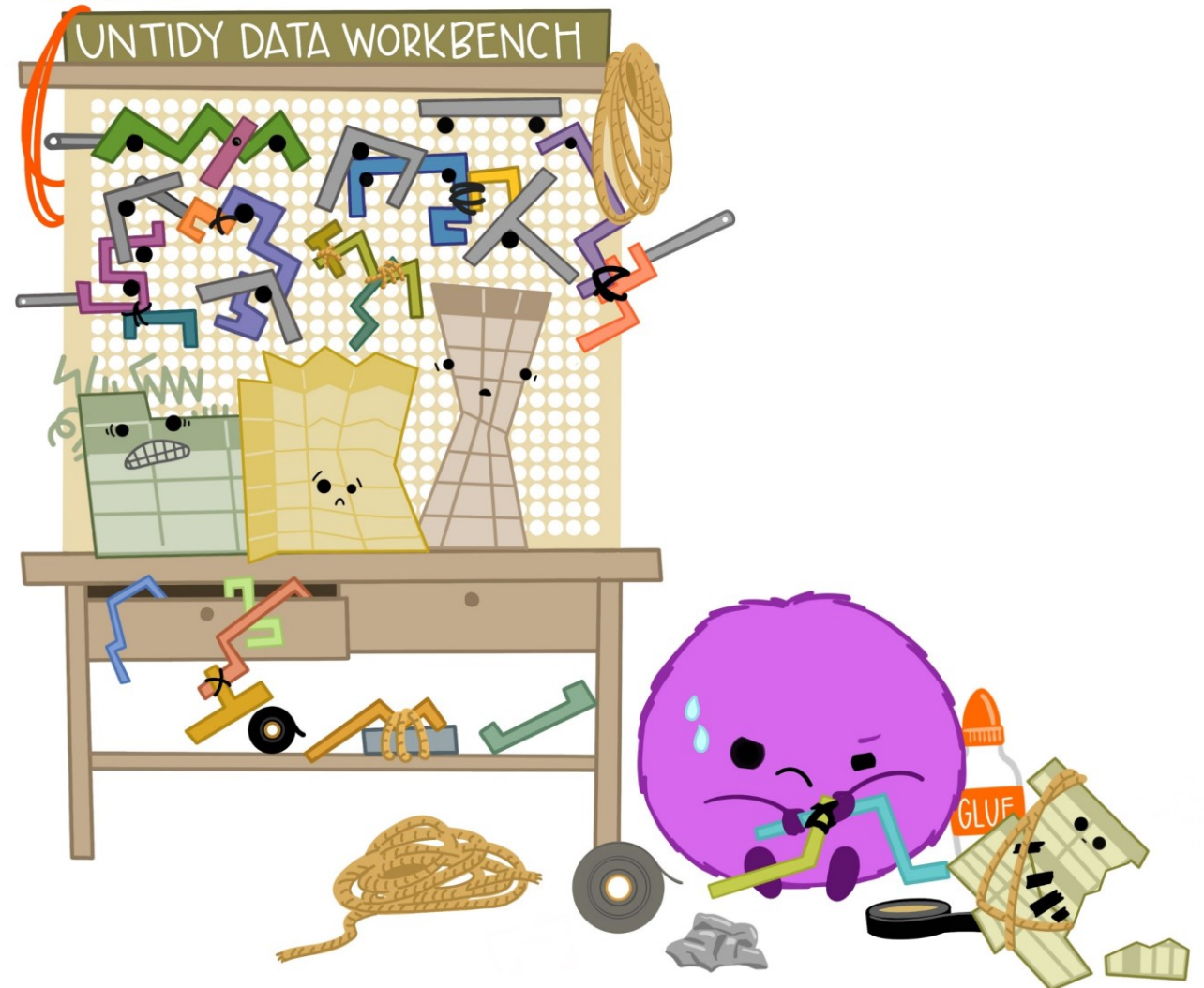
person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Tidy

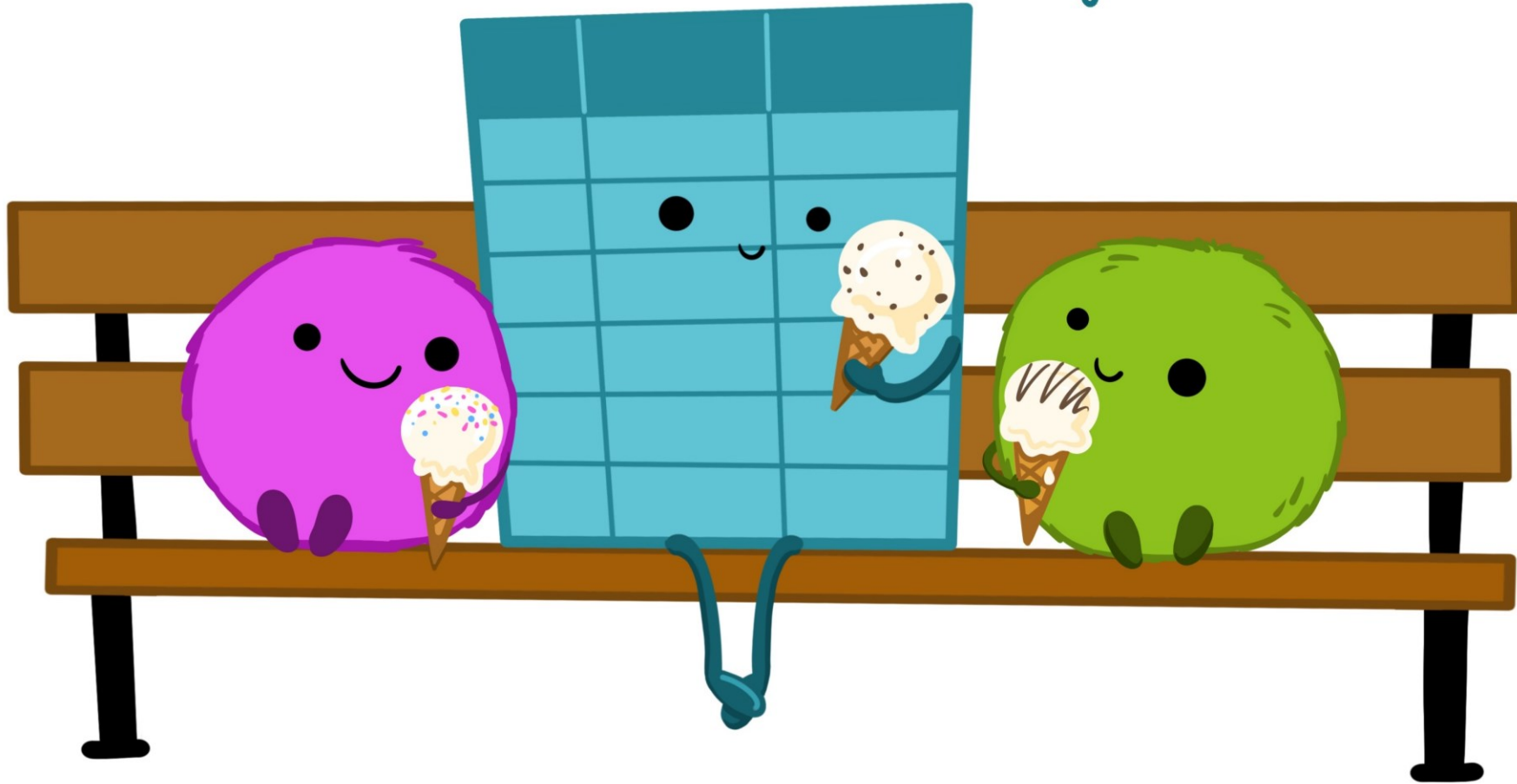
When working with tidy data, we can use the same tools in similar ways for different datasets...



...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.



make friends with tidy data.



2. Piping functions (vs nested)

Nested functions (base R)

3

2

1

```
length(unique(data$variable))
```

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

1

2

3

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



2. Piping functions (vs nested)

A more complex example of piping (from this workshop script):

```
data %>%  
  group_by(Species, Sex, Island) %>%  
  summarise(across(where(is.numeric), ~ mean(.x, na.rm = TRUE))) %>%  
  select(-Sample_Number) %>%  
  arrange(desc(Sex)) %>%  
  write_csv(., "../out/data_means.csv")
```

No need to create intermediate objects,
just pipe the outcome of one function into the next
(use "." as placeholder)

String manipulation

stringr::str_*



Work with character strings (text)

- Uses **regex** (regular expressions): a “codified” way to describe patterns in text strings, to do things with them (extract, select, replace, ...)
- Same rules as `grep` or `sed` in Unix and Bash

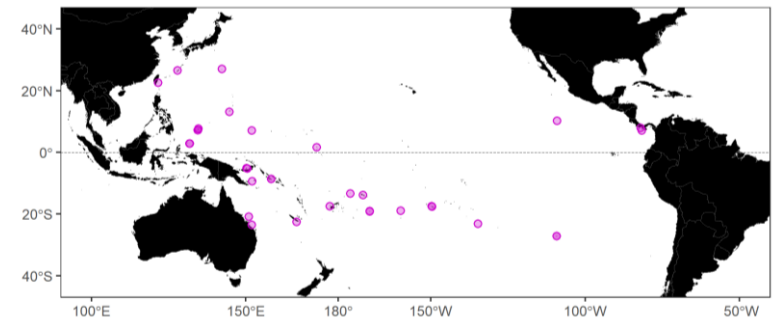
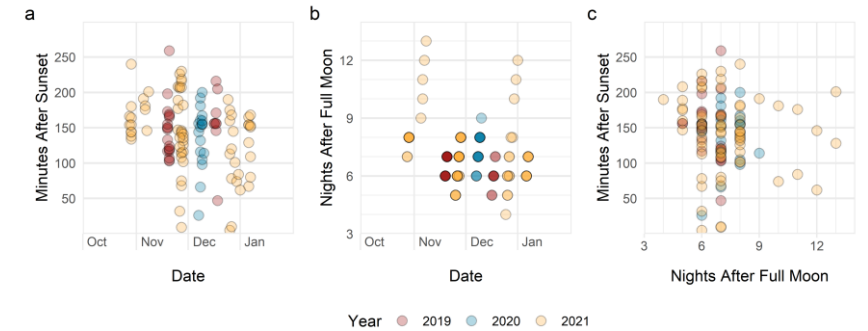
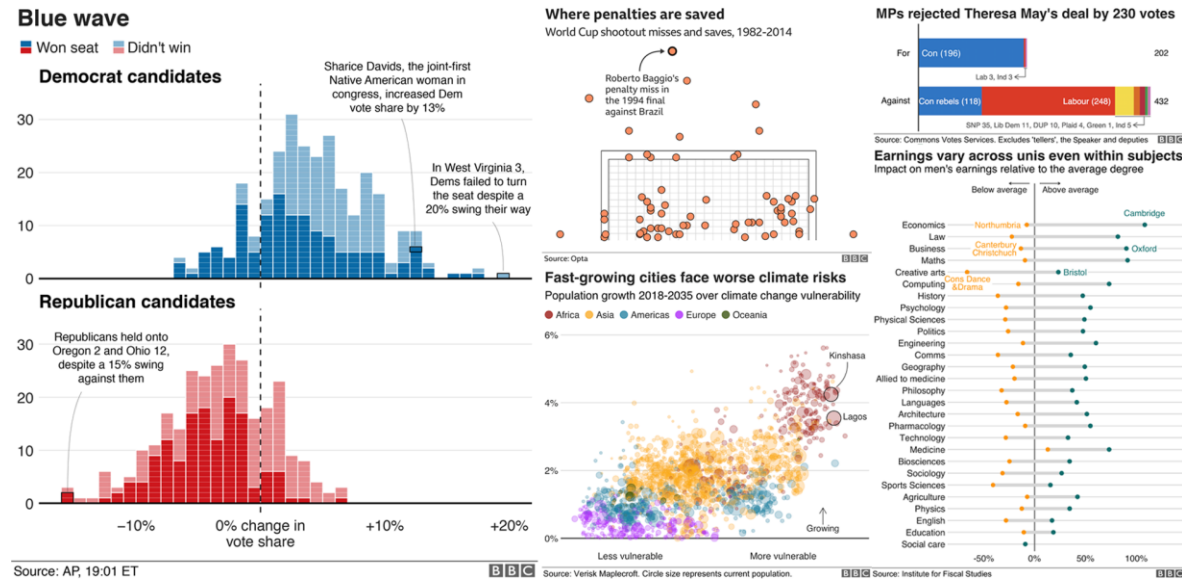
```
"d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae; g__Acinetobacter"
```

```
"d__Bacteria;/p__Proteobacteria;/c__Gammaproteobacteria;/o__Pseudomonadales;/f__Moraxellaceae;/g__Acinetobacter"
```

```
"d__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae; g__Acinetobacter"
```

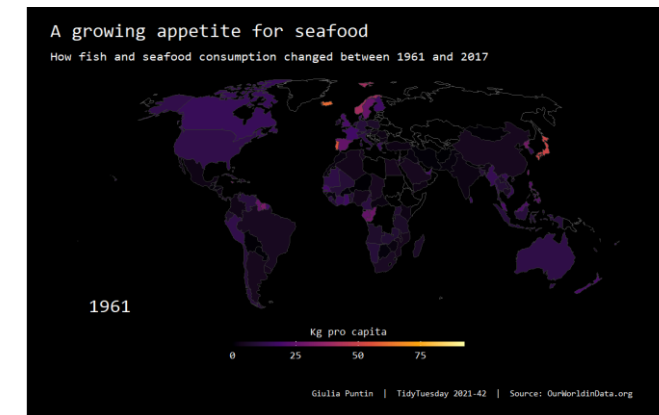
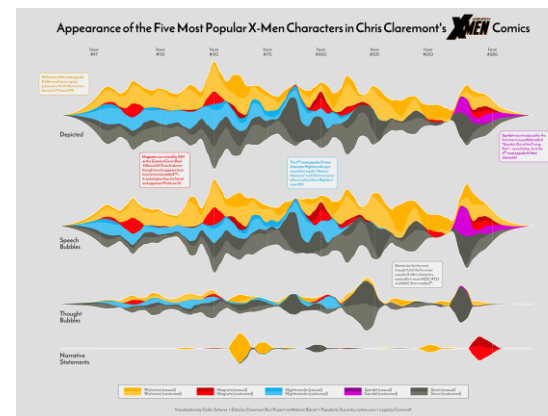
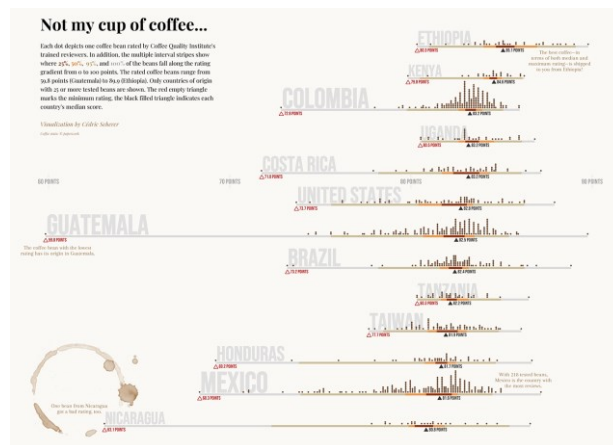
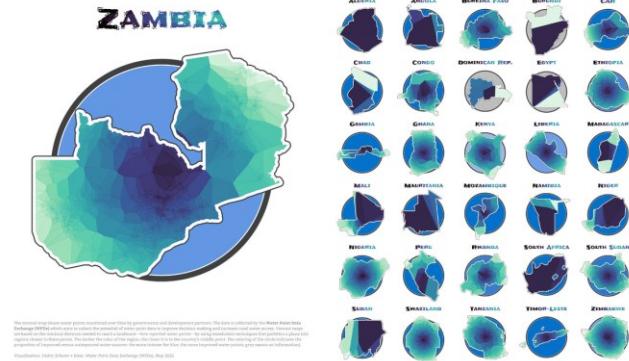
Data Viz with ggplot

- Publication quality (also BBC, Economist ...)



Data Viz with ggplot

- Publication quality (also BBC, Economist ...)
- Also just super beautiful:
 - Cedric Scherer [tutorial blog](#)
 - R graph gallery
 - #TidyTuesday



Practical part: R and Tidyverse in action!



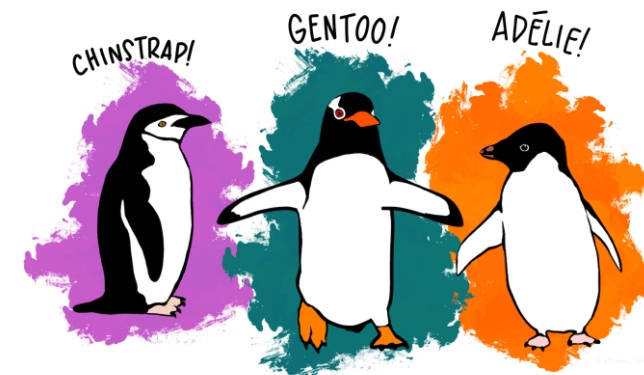
The R script is based on **dummy data**
(**modified from open data set** for didactic purposes by me)

Data on penguins' body dimensions, stable isotope measurements, and life stages by species, sex, location, etc

Chosen because of properties:

- numeric data
- character data and
- big enough that it would be annoying to do in Excel
- => good for showing my favorite Tidyverse functions

- Original data: `palmerpenguins::penguins_raw`
- For more info: `?penguins_raw`



Data wrangling

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

But briefly:

- Create/modify variables
- Subset data
- Summarize
- Re-shape
- Merge
- Correct values
- Plot like a pro

Data wrangling

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

But briefly:



Use in combination with
`if_else()`, `case_when()`, `%in%`

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

All workshop material
available at

[https://github.com/sPuntinG/**BiolPostgrad_Rworkshop2023**](https://github.com/sPuntinG/BiolPostgrad_Rworkshop2023)

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags

Go to file

Add file

<> Code

About



No description, website, or topics provided.

Readme

0 stars

1 watching

0 forks

Releases

No releases published

[Create a new release](#)

Packages

No packages published

[Publish your first package](#)

	sPuntinG Update README.md
	in completed (draft of) R script
	.gitignore changes to gitignore
	README.md Update README.md
	Rworkshop2023_Demo.R completed (draft of) R script
	Rworkshop2023_Intro.pptx completed (draft of) R script
	~\$R4ReproducibleResearch.pptx initial commit
	~\$rksp_outline.docx initial commit

Local

Codespaces

Clone



HTTPS

SSH

GitHub CLI

https://github.com/sPuntinG/GenGen_Rworkshop2



Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

... and waaaay more!

This is **just a short demonstration** of what can be done (and how easily) in R Tidyverse,
BUT there's so much more out there!

Recommend:

- Today: **ask** me/us about specific tasks/operations that you'd like to learn to execute in R
- Any time: Check out the package **cheat sheets** for inspiration (I use them a lot!)
https://posit.co/resources/cheatsheets/?type=posit-cheatsheets&_page=2/

