GG606

Workflows

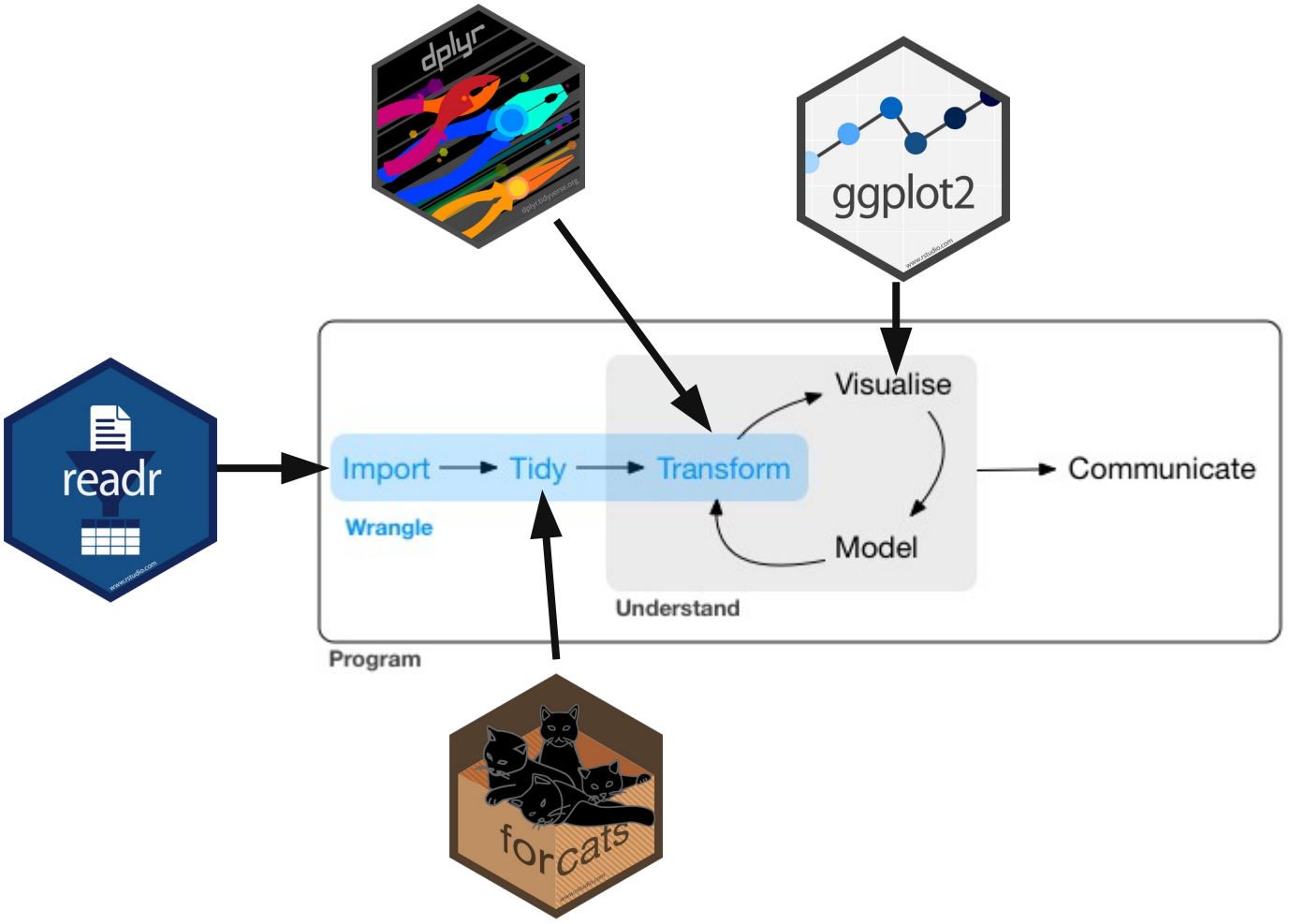
Homework

- `install.packages("palmerpenguins")`
- Make some interesting plots with new `geom_` and try out some `theme_`
- ggplot2.tidyverse.org
- Try out `ggsave`
- pangaea.de & www.frdr-dfdr.ca

Homework

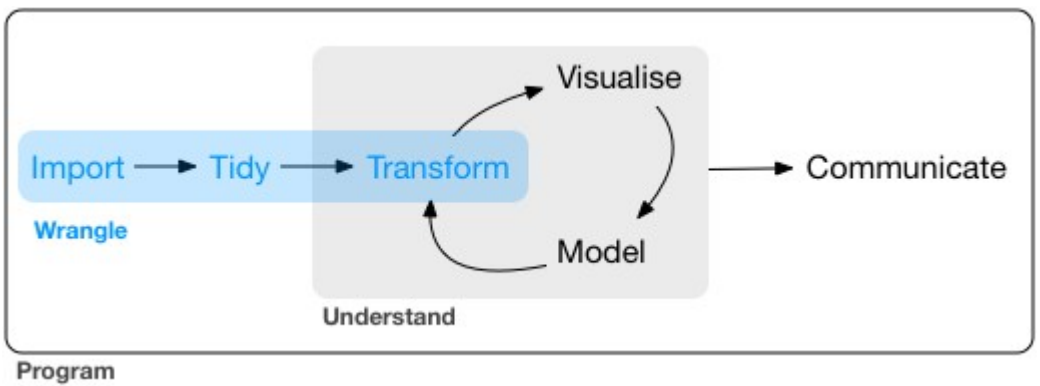
Workflows

- Organisation
- Reproducible
- Inputs & Outputs
- R-Script vs R-Markdown vs Function


dplyr
ggplot2
readr
Import
Tidy
Transform
Visualise
Model
Communicate
Wrangle
Understand
Program
forcats

R-Script

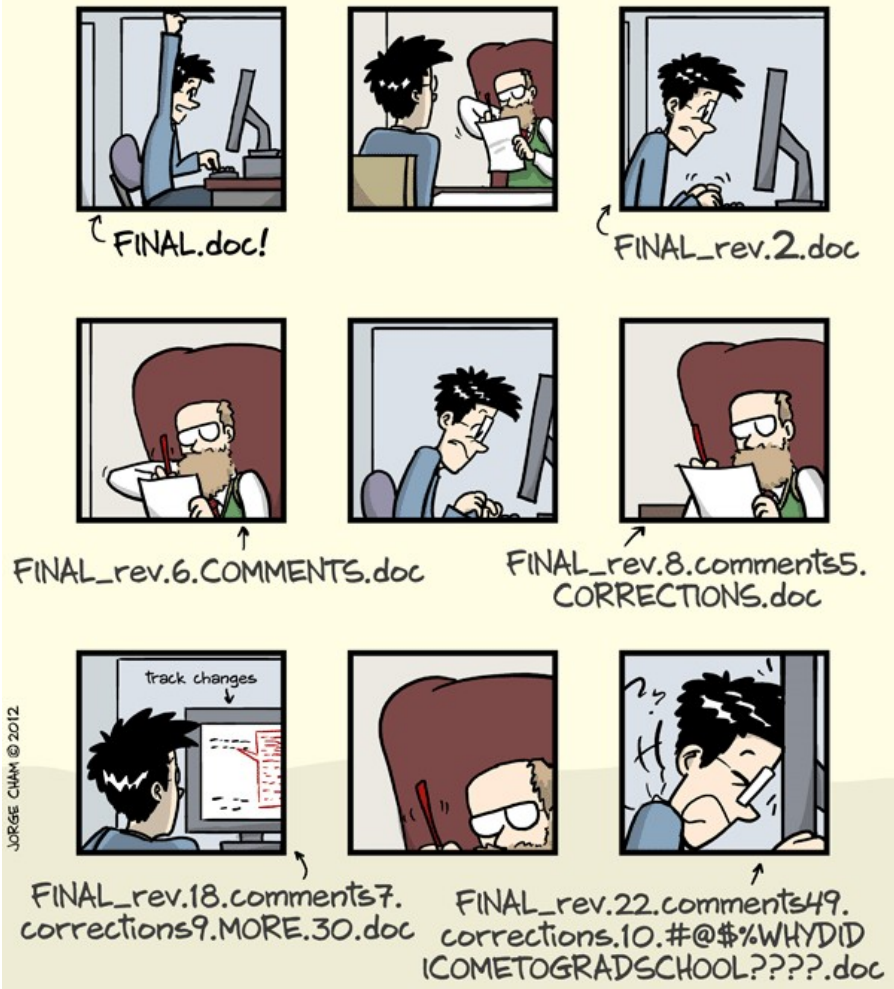
- File > New File > R Script
- Series of command *and comments*
- Sequence & Story

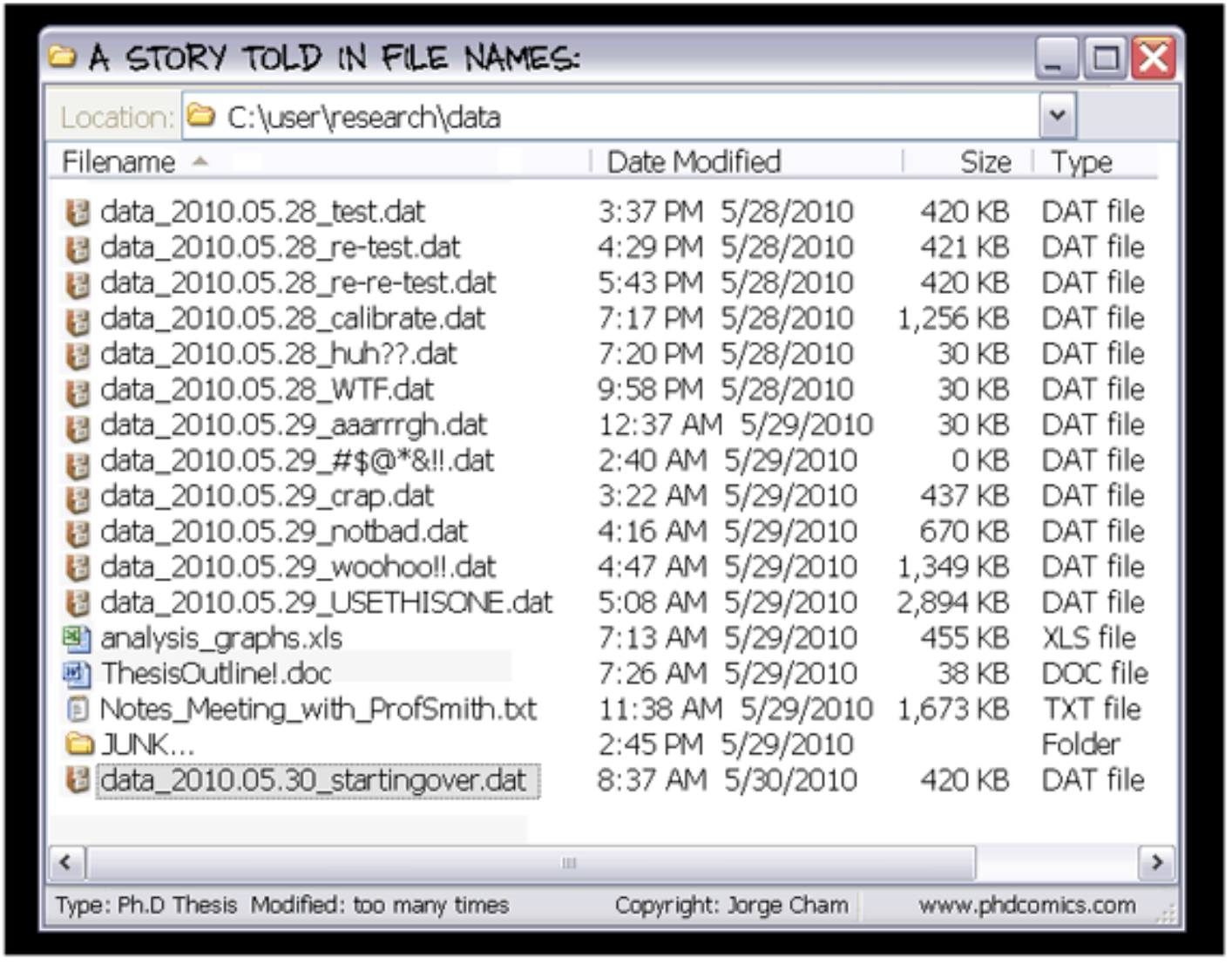


New Computer/Collaborator

- You are your own (future|past) collaborator
- *New computer?*
- *Sent files to collaborators?*

"FINAL".doc


FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012


A STORY TOLD IN FILE NAMES:
Location: C:\user\research\data
Filename
Date Modified
Size
Type
data_2010.05.28_test.dat 3:37 PM 5/28/2010 420 KB DAT file
data_2010.05.28_re-test.dat 4:29 PM 5/28/2010 421 KB DAT file
data_2010.05.28_re-re-test.dat 5:43 PM 5/28/2010 420 KB DAT file
data_2010.05.28_calibrate.dat 7:17 PM 5/28/2010 1,256 KB DAT file
data_2010.05.28_huh??.dat 7:20 PM 5/28/2010 30 KB DAT file
data_2010.05.28_WTF.dat 9:58 PM 5/28/2010 30 KB DAT file
data_2010.05.29_aaarrrgh.dat 12:37 AM 5/29/2010 30 KB DAT file
data_2010.05.29_#$@*&!!.dat 2:40 AM 5/29/2010 0 KB DAT file
data_2010.05.29_crap.dat 3:22 AM 5/29/2010 437 KB DAT file
data_2010.05.29_notbad.dat 4:16 AM 5/29/2010 670 KB DAT file
data_2010.05.29_woohoo!!.dat 4:47 AM 5/29/2010 1,349 KB DAT file
data_2010.05.29_USETHISONE.dat 5:08 AM 5/29/2010 2,894 KB DAT file
analysis_graphs.xls 7:13 AM 5/29/2010 455 KB XLS file
ThesisOutline!.doc 7:26 AM 5/29/2010 38 KB DOC file
Notes_Meeting_with_ProfSmith.txt 11:38 AM 5/29/2010 1,673 KB TXT file
JUNK... 2:45 PM 5/29/2010 Folder
data_2010.05.30_startingover.dat 8:37 AM 5/30/2010 420 KB DAT file
Type: Ph.D Thesis Modified: too many times
Copyright: Jorge Cham
www.phdcomics.com

New Computer/Collaborator

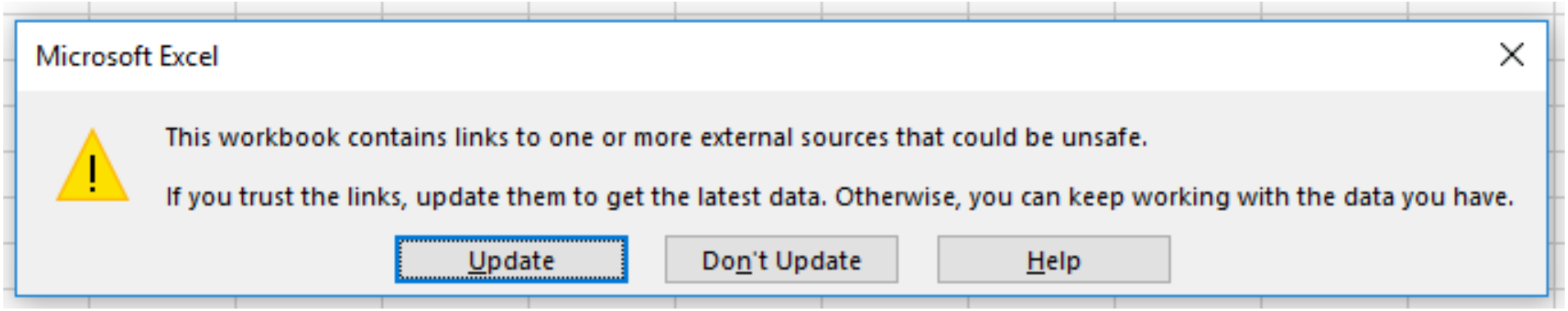
- Will your files always be here?
- Where should your data, code, outputs go?
- File paths can be a nightmare

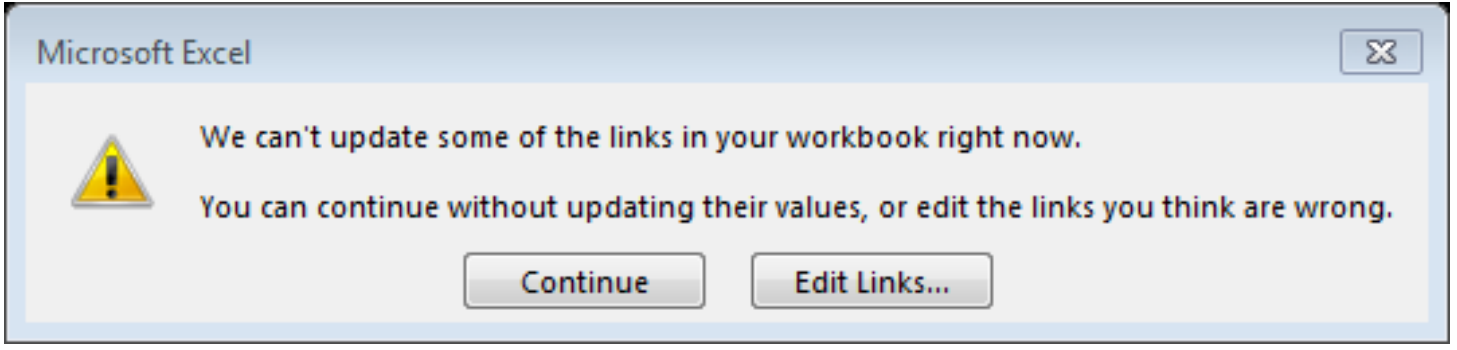
New Computer/Collaborator

- Will your files always be here?
- Where should your data, code, outputs go?
- File paths can be a nightmare
- RStudio Project helps

```
> getwd()
[1] "/home/jason/school/Laurier teaching/GG606 Winter 2023"
```

```
> getwd()
[1] "/home/jason/github/GG606AW24"
```


Microsoft Excel
This workbook contains links to one or more external sources that could be unsafe.
If you trust the links, update them to get the latest data. Otherwise, you can keep working with the data you have.
Update
Don't Update
Help


Microsoft Excel
We can't update some of the links in your workbook right now.
You can continue without updating their values, or edit the links you think are wrong.
Continue
Edit Links...

Project-Oriented Workflow

- Commands
- R Scripts (series of commands)
- R Project (series of related things)
- `here` package
  (project-oriented workflow & portability)

Why are Jenny Bryan & Timothée Poisot So Worked Up?

If the first line of your R script is

```
setwd("C:\Users\jenny\path\that\only\I\have")
```

I* will come into your office and
SET YOUR COMPUTER ON FIRE 🔥.

* or maybe Timothée Poisot will

If the first line of your R script is

```
rm(list = ls())
```

I will come into your office and
SET YOUR COMPUTER ON FIRE 🔥.

Project-Oriented Workflow

- https://www.tidyverse.org/blog/2017/12/workflow-vs-script/

Two specific slides generated much discussion and consternation in #rstats Twitter:

> *If the first line of your R script is*
>
> `setwd("C:\Users\jenny\path\that\only\I\have")`
>
> *I will come into your office and SET YOUR COMPUTER ON FIRE* 🔥.

> *If the first line of your R script is*
>
> `rm(list = ls())`
>
> *I will come into your office and SET YOUR COMPUTER ON FIRE* 🔥.

I stand by these strong opinions, but on their own, threats to commit arson aren't terribly helpful! Here I explain *why* these habits can be harmful and may be indicative of an awkward workflow. Feel free to discuss more on community.rstudio.com.

Caveat: only you can decide how much you care about this. The importance of these practices has a lot to do with whether your code will be run by other people, on other machines, and in the future. If your current practices serve your purposes, then go forth and be happy.

github.com/allisonhorst


"~/mm/nope.csv"
non-portable pa
setwd("/haha/good/luck/")
"/NoT/HERE"
Setwd("C:\userlohlno")
setwd("c:\iamyournightmare\!")

here

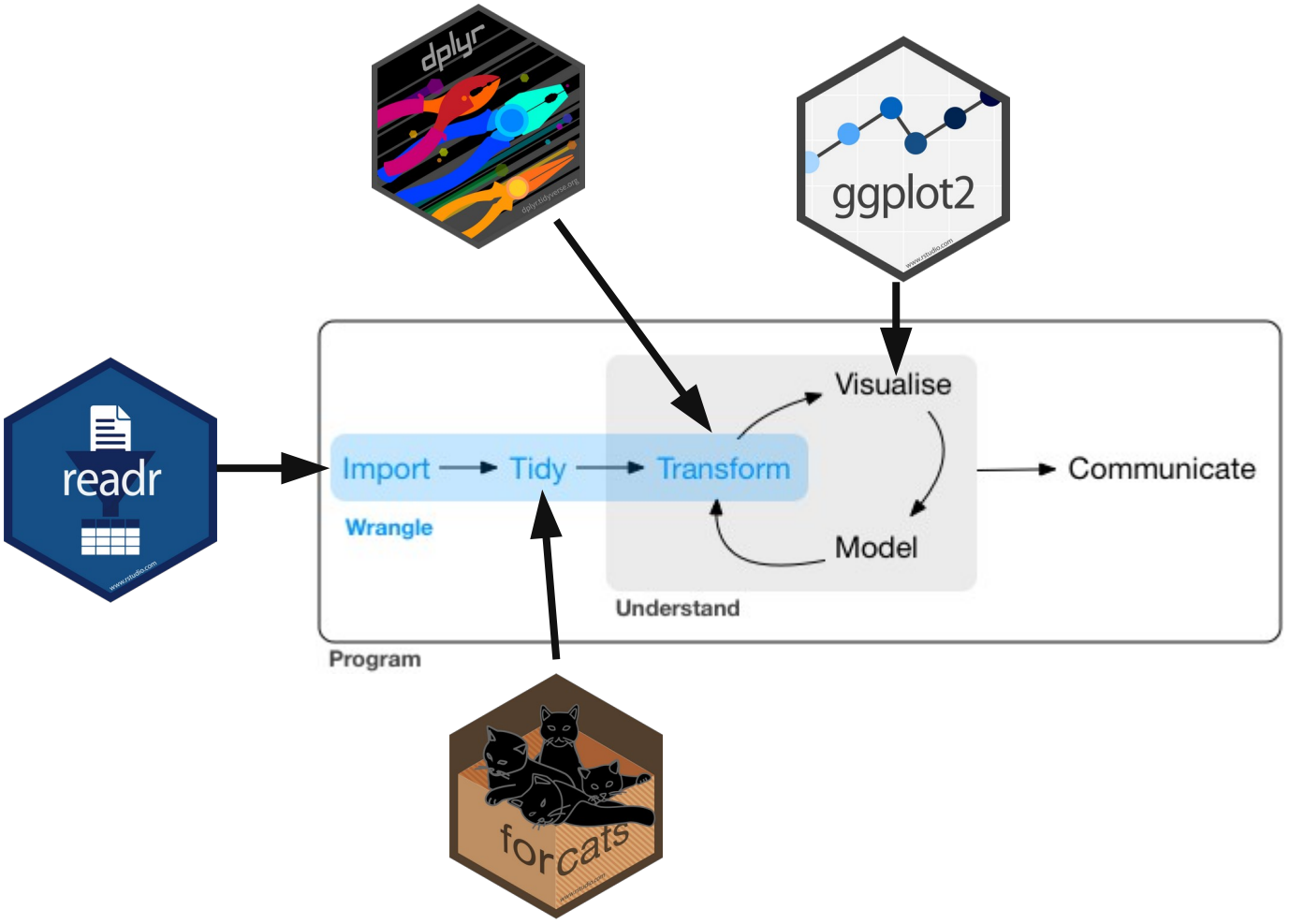
- It figures out *real* path to project

  - `"C:\Users\jenny\path\that\only\I\have"`
  - `"/Users/jenny/cuddly_broccoli/verbose_funicular/foofy/data"`

here

- It figures out *real* path to project

  - `here("folder", "file")`
  - `ggsave(here("figures", "beak_size_by_species.pdf"))`


dplyr
ggplot2
readr
Import
Tidy
Transform
Visualise
Model
Communicate
Wrangle
Understand
Program
forcats

Template

- What folders do you want in each project?
- Inputs & Outputs
- Code
- Documents

Build a basic template


readr
www.rstudio.com

Reproducible Scientific Workflows

- lack of reproducibility in scientific studies is problem – especially in environmental science
- consistent and reproducible reading and writing of data used for scientific work is critical

Wrap Up

- Make folders that make sense
- Put a data file in there
- `library(readr)`
- `read_csv()` for your found data and `penguins_raw.csv`
- Update your code to use `here`, load data, save figures
- Think about metadata

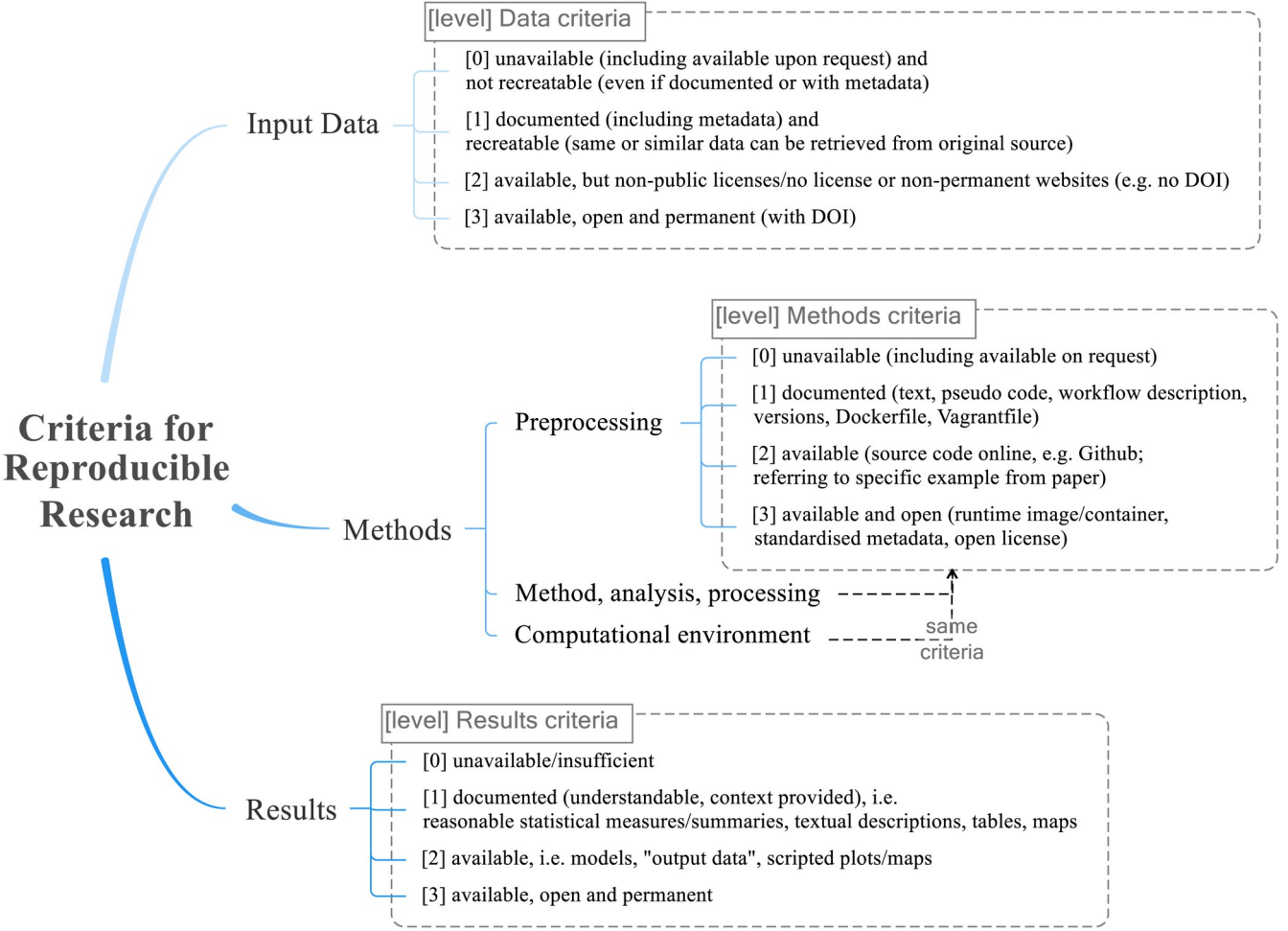
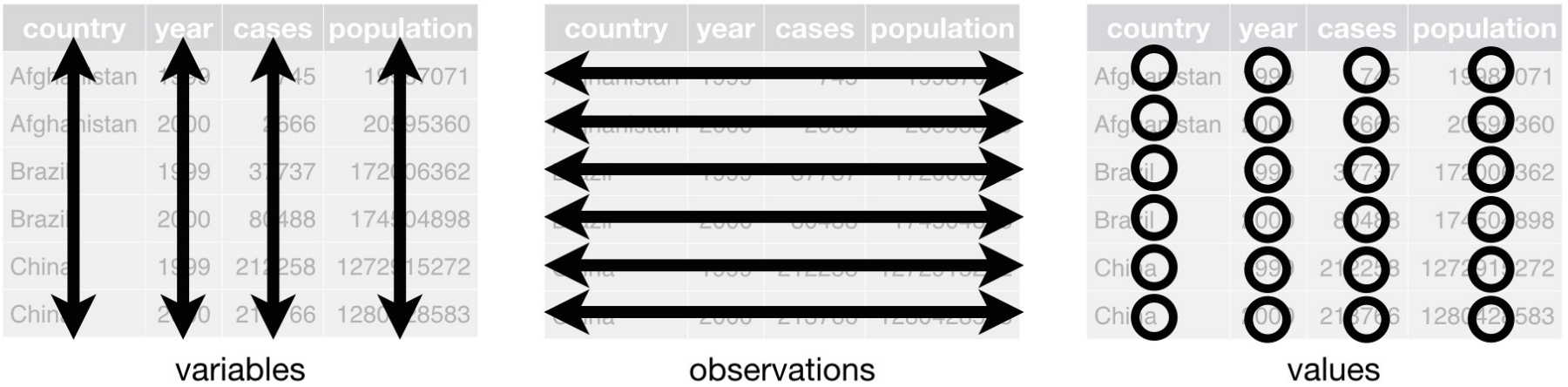


Figure 2: The final reproducible research criteria used for the evaluation.

The categories Data, Methods (sub-categories: preprocessing, method/analysis/processing, and computational environment), and Results each have four levels ranging from 0 = not reproducible to 3 = fully reproducible.

Nüst, Daniel, Carlos Granell, Barbara Hofer, Markus Konkol, Frank O. Ostermann, Rusne Sileryte, and Valentina Cerutti. 2018. "Reproducible Research and GIScience: An Evaluation Using AGILE Conference Papers." PeerJ 6 (July): e5072. https://doi.org/10.7717/peerj.5072

Three Rules of Tidy Data



1. Each variable must have its own column.
2. Each observation must have its own row.
3. Each value must have its own cell.

Messy Data are Everywhere

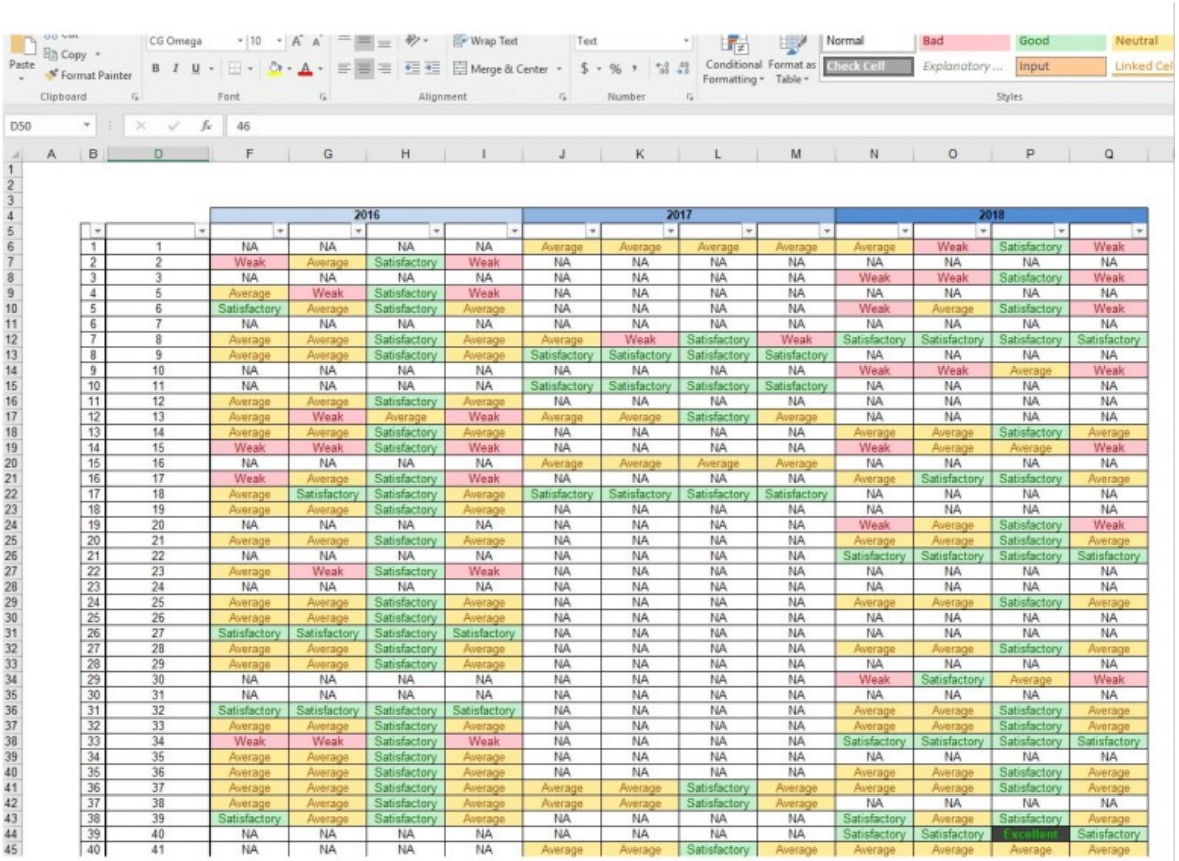
Wide data format

Time	A	B	C
0	1.1	4.2	5.6
1	1.0	4.5	5.8

Tidy data format

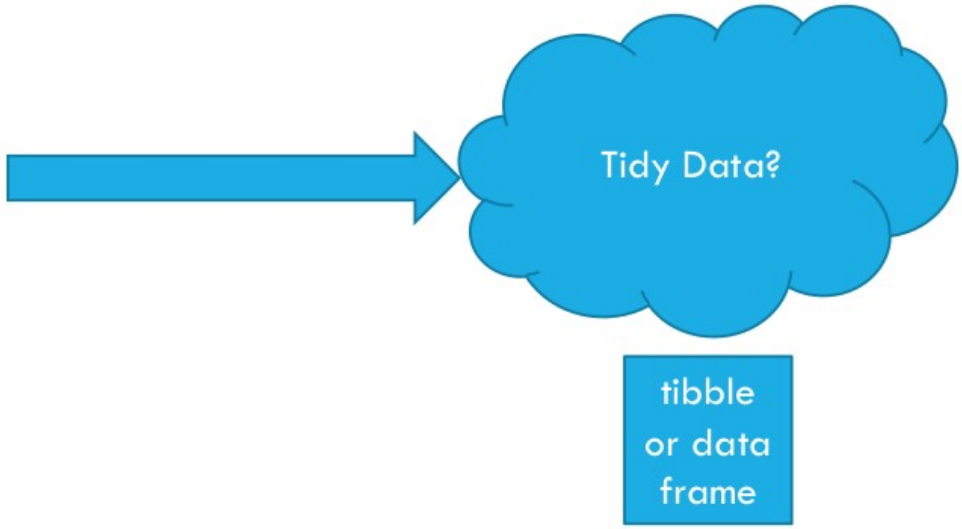
Time	Sample	Value	id
0	A	1.1	1
1	A	1.0	1
0	B	4.2	1
1	B	4.5	1
0	C	5.6	1
1	C	5.8	1

Joachim Goedhart



Messy Data are Everywhere

	Population Estimates			
District	MOH area	2009	2010	2011
Colombo	Dehiwala	233664	236809	240018
	Piliyandala	168958	171232	173553
	Homagama	204699	207454	210266
	Kaduwela	233612	236757	239966
	Kolonnawa	178675	181080	183534
	Kotte	69302	70234	71186
	Maharagama	157089	159203	161361
	MC-Colombo	715249	724877	734702
	Moratuwa	197357	200013	202724
	Nugegoda	103230	104619	106037
	Padukka	60589	61407	62234
	Boralesgamuwa	61944	62777	63628
	Hanwella	104218	105621	107053
Gampaha	Attanagalla	180397	183907	187521
	Biyagama	188435	192102	195877
	Divulapitiya	149474	152382	155377
	Gampaha	199015	202888	206875
	Ja-Ela	148509	151399	154374



Invertebrate Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Date		2016_11_15				2016_11_15				2016_11_15				2016_11_16				2016_11_16			
2	Time (EST)		8:30:00				11:30:00				14:30:00				8:30:00				11:30:00			
3	Scientific_name	Common_name	Site_36	Site_42	Site_43	Site_44	Site_6	Site_25	Site_33	Site_34	Site_1	Site_25	Site_29	Site_30	Site_36	Site_42	Site_43	Site_44	Site_6	Site_25	Site_33	Site_34
4	Amphipoda	Scuds	0	0	0	0	0	0	0	2	1	0	1	0	0	0	0	0	1	0	0	0
5	Anisoptera	Dragonflies	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
6	Bivalvia	Clams_Mussels	0	0	6	0	2	0	0	7	0	0	0	0	0	0	8	0	2	0	0	0
7	Ceratopogondia	No-See-Ums	0	0	0	0	5	0	0	0	0	0	1	7	0	0	0	1	1	0	0	0
8	Chironomidae	Midges	0	2	0	17	8	11	8	3	0	1	4	5	7	16	2	4	5	3	11	2
9	Coelenterata	Hydras	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Coleoptera	Beetles	1	1	2	0	0	0	0	2	0	1	0	0	2	0	8	0	0	1	1	2
11	Culicidae	Mosquitos	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	Decapoda	Crayfish	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	Ephemeroptera	Mayflies	10	3	0	0	1	3	2	2	0	9	2	0	1	1	0	0	0	2	0	1
14	Gastropoda	Snails	0	0	0	0	0	3	0	2	0	0	0	0	0	0	0	0	0	0	0	1
15	Hemiptera	True_bugs	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
16	Hirudinea	Leeches	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
17	Hydrachnida	Mites	0	0	0	0	0	0	0	0	5	2	0	1	0	0	1	0	4	0	1	0
18	Isopoda	Sow_bugs	2	0	0	2	0	5	2	1	0	0	0	0	3	0	3	2	2	5	2	0
19	Lepidoptera	Aquatic_moths	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	Megaloptera	Fishflies_and_Alderflies	0	0	4	0	0	0	0	0	0	0	0	0	0	0	15	0	0	0	0	0
21	Misc._Diptera	True_flies	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
22	Nematoda	Roundworms	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Oligochaeta	Aquatic_earthworms	22	8	3	2	64	5	36	13	18	1	8	8	0	1	26	3	23	10	3	20
24	Plecoptera	Stoneflies	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	Simuliidae	Black_flies	5	2	0	0	0	0	0	1	2	0	1	0	6	0	0	0	1		0	1
26	Tabanidae	Horse_flies_Deer_flies	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	Tipulidae	Crane_flies	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	1
28	Trichoptera	Caddisflies	6	1	69	17	0	7	1	7	5	18	0	13	4	0	80	1	1	0	1	1
29	Turbellaria	Flatworms	0	0	6	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
30	Zygoptera	Damselflies	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	0	18	0

Lake Sediment Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Lat	Long				**Lattitude:**	**62.608333**											
2	62°36'30.00"N	114°36'19.00"W	**Predicted 210Pb annual fallout****			dpm/cm2/yr	0.334	<--See Muir et al	2009; Environ Sci Technol 43:4802-9.				**0.06**	**<-- Estimate of unsupported 210Pb inventory (±0.0126dpm/cm2) below last horizon of qu**				
3	**Lake Name**	NW20	**Total Measured 210Pb inventory****			dpm/cm2	6.284	± 1 Std Dev	0.143				**0.95**	**<-- Above as a % of total 210Pb inventory**				
4	**Coring Date:**	1-Jun-18	**Measured annual 210Pb fallout****			dpm/cm2/yr	0.196	± 1 Std Dev	0.005									
5			Mean	StDev		**Focusing Factor**	0.586	±	*0.013*		**CRS AGE MODEL**							
6	Mean supported 210Pb value:		0.6487	0.4542					above only accounts for measured error, not predicted				**USE THIS**		1 Sigma			
7	(mean of all 214Bi +214Pb values from 0-14 cm)		***sample by sample method used***				= Interpolated					CRS Date		Total	± error	Organic Matter	Inorganic Matter	**not compaction corrected
8						**Depth by**	Measured Total	Error			Raw CRS	± error	**CRS Dates**	(dry mass	(dry mass	(dry mass	(dry mass	(depth based
9	**Original Given Depth Intevals**		**Depth Interval (cm)**		**mid-**	**Cumulative**	Pb-210	1 std.dev.		Depth	Dates	2 Sigma	**with Linear**	sedimentation)	sedimentation)	sedimentation)	sedimentation)	sedimentation)
10			**Top**	**Bottom**	**depth (cm)**	**Mass (g/cm2)**	(dpm/g)	(dpm/g)		(cm)	(CE)	(Year CE)	**Extrapolation**	(g/cm2 yr)	(g/cm2 yr)	(g/cm2 yr)	(g/cm2 yr)	(cm/yr)
11			0	1	0.5	0.0265	65.4450	2.5035		1	2008.22	0.84	2008.22	0.0030	0.0001	0.0021	0.0009	0.1145
12			1	2	1.5	0.0459	49.4680	2.0452		2	2000.84	1.20	2000.84	0.0029	0.0001	0.0021	0.0009	0.1517
13			2	3	2.5	0.0689	47.3543	2.1374		3	1989.66	1.88	1989.66	0.0024	0.0001	0.0017	0.0008	0.1059
14			3	4	3.5	0.0882	33.9023	1.8079		4	1980.44	2.59	1980.44	0.0024	0.0002	0.0016	0.0008	0.1248
15			4	5	4.5	0.1169	23.1014	1.7118		5	1967.36	3.76	1967.36	0.0027	0.0002	0.0018	0.0009	0.0931
16			5	6	5.5	0.1402	15.7680	1.2120		6	1956.80	5.19	1956.80	0.0026	0.0003	0.0018	0.0008	0.1111
17			6	7	6.5	0.1731	10.2496	0.9535		7	1943.06	7.70	1943.06	0.0030	0.0004	0.0020	0.0010	0.0895
18			7	8	7.5	0.2081	6.6886	0.9172		8	1928.89	11.20	1928.89	0.0031	0.0006	0.0020	0.0010	0.0873
19			8	9	8.5	0.2428	3.8655	0.7817		9	1918.73	14.29	1918.73	0.0040	0.0012	0.0027	0.0013	0.1148
20			9	10	9.5	0.2783	2.2718	0.7794		10	1912.82	15.54	1912.82	0.0066	0.0040	0.0043	0.0022	0.1854
21			10	11	10.5	0.3037	2.6269	0.7230		11	1907.89	17.02	1907.89	0.0055	0.0035	0.0038	0.0017	0.2188
22			11	12	11.5	0.3424	2.1825	0.7436		12	1894.88	20.96	1894.88	0.0036	0.0017	0.0025	0.0012	0.0935
23			12	13	12.5	0.3888	2.2564	0.9050		13			1868.95					
24			13	14	13.5	0.4241	1.5388	0.7479		14			1854.69					
25			14	15	14.5	0.4557	1.2577	1.0666		15			1841.90					
26			15	16	15.5	0.4905	1.0132	0.7605		16			1827.88					
27			16	17	16.5	0.5324				17			1810.96					
28			17	18	17.5	0.5673				18			1796.87					
29			18	19	18.5	0.6129				19			1778.47					
30			19	20	19.5	0.6331				20			1770.29					

Homework

- folder structure for the workflow
  - we spoke about keeping raw data separate from processed data and keeping figures and/or tables together)
- use the `here` package and function
- R script to load data from pangaea.de or www.frdr-dfdr.ca
  - example, `read_csv(here("folder", "file"))`
- create and save a figure to an appropriate folder
  - hint, use the `ggsave` and `here` functions
- Put a screenshot of your success on discord