

DATA WRANGLING AND REPRODUCIBILITY IN R

DR. COLLEEN FORTIER

BIOINFORMATICS OFFICE HOUR - JULY 24, 2025

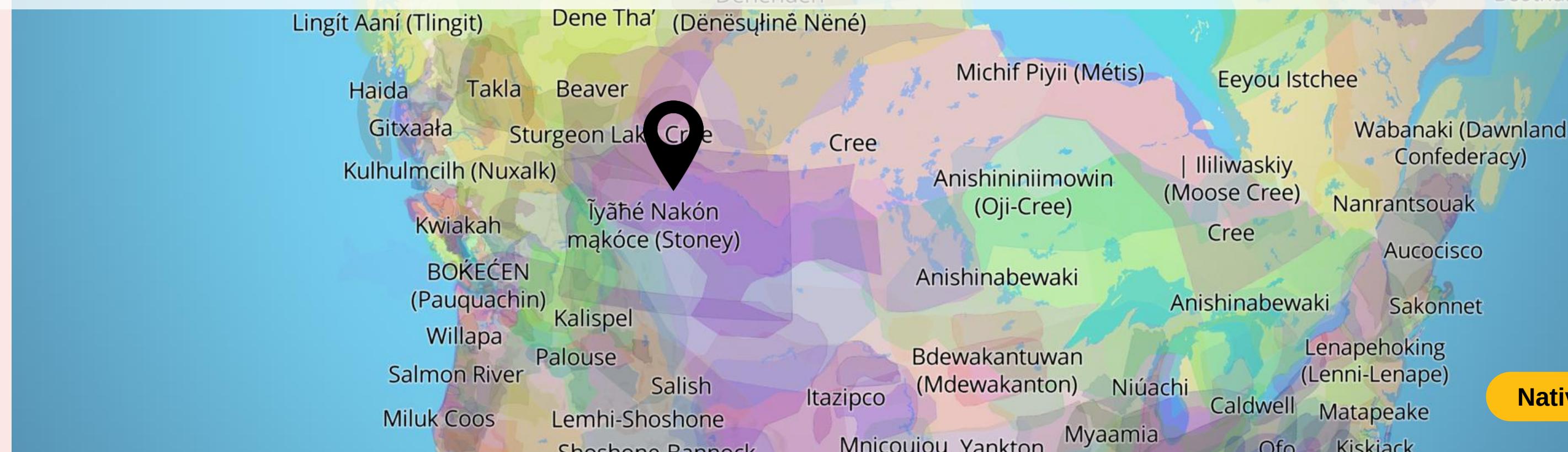
LAND ACKNOWLEDGEMENT



LAND ACKNOWLEDGEMENT

The University of Alberta, its buildings, labs and research stations are primarily located on the territory of the Néhiyaw (Cree), Niitsitapi (Blackfoot), Métis, Nakoda (Stoney), Dene, Haudenosaunee (Iroquois) and Anishinaabe (Ojibway/Saulteaux), lands that are now known as part of Treaties 6, 7 and 8 and homeland of the Métis.

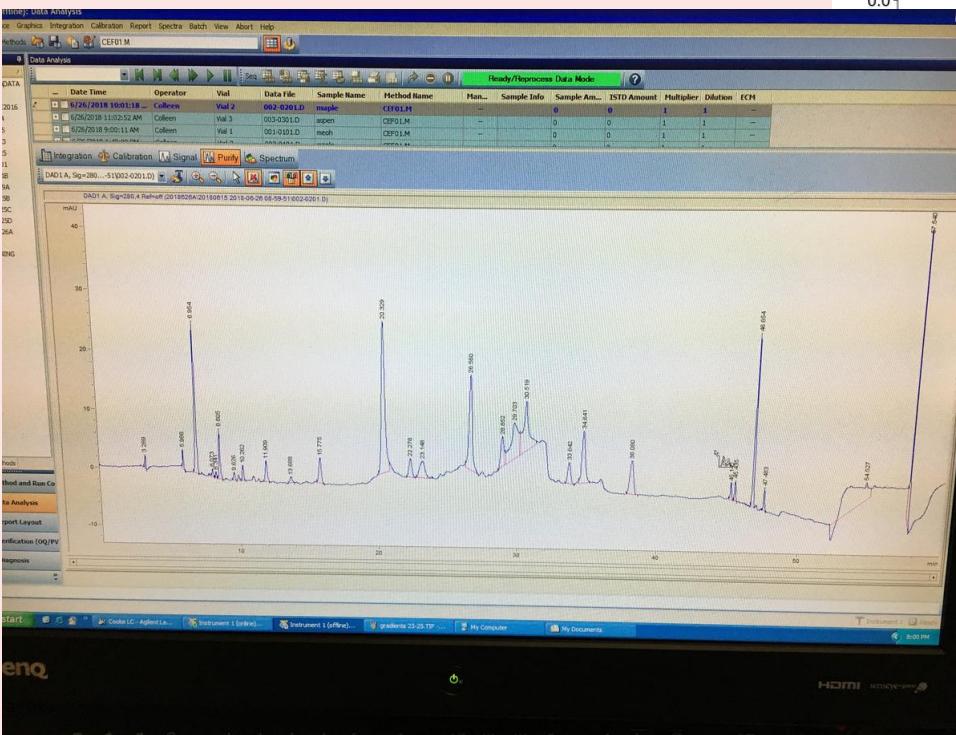
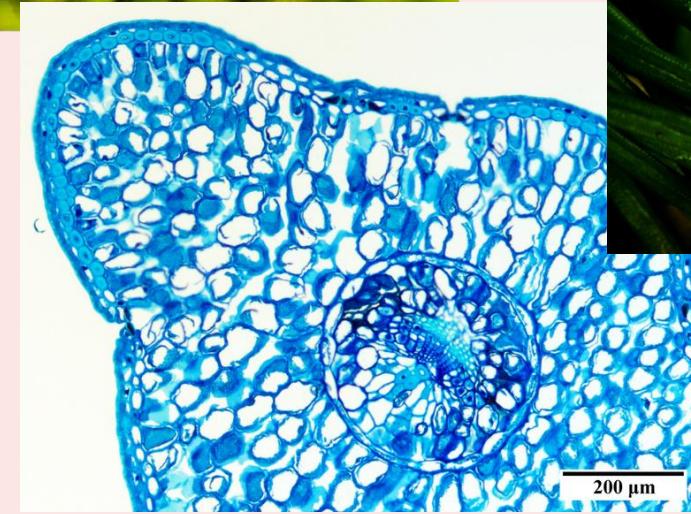
The University of Alberta respects the sovereignty, lands, histories, languages, knowledge systems and cultures of all First Nations, Métis and Inuit nations.



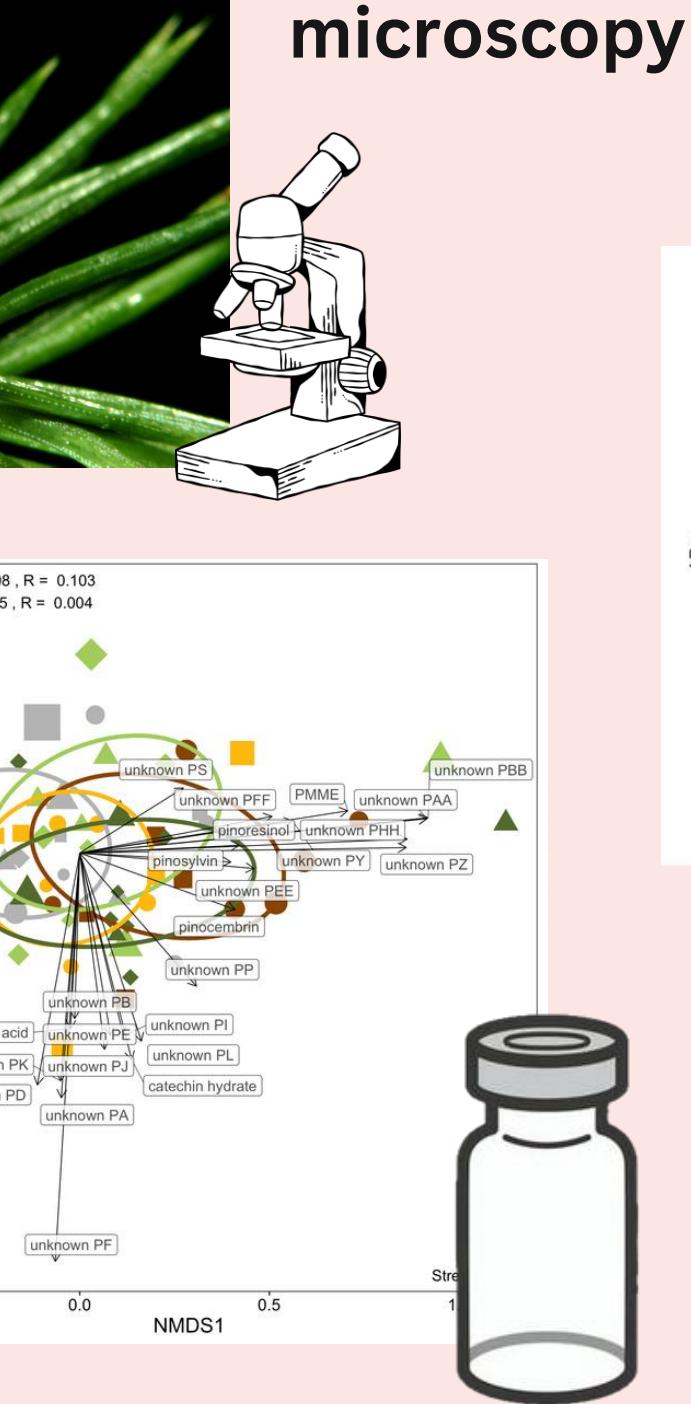
A BIT ABOUT ME



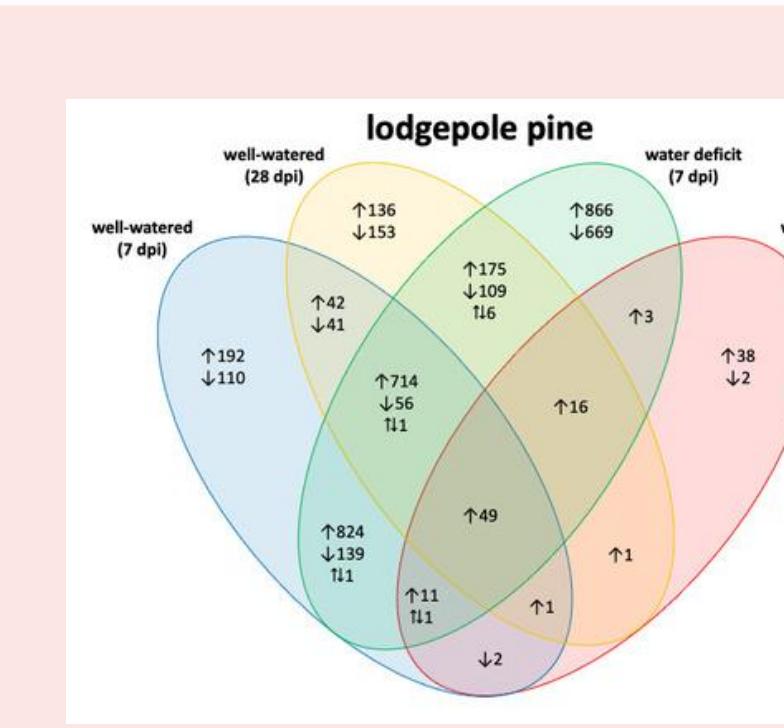
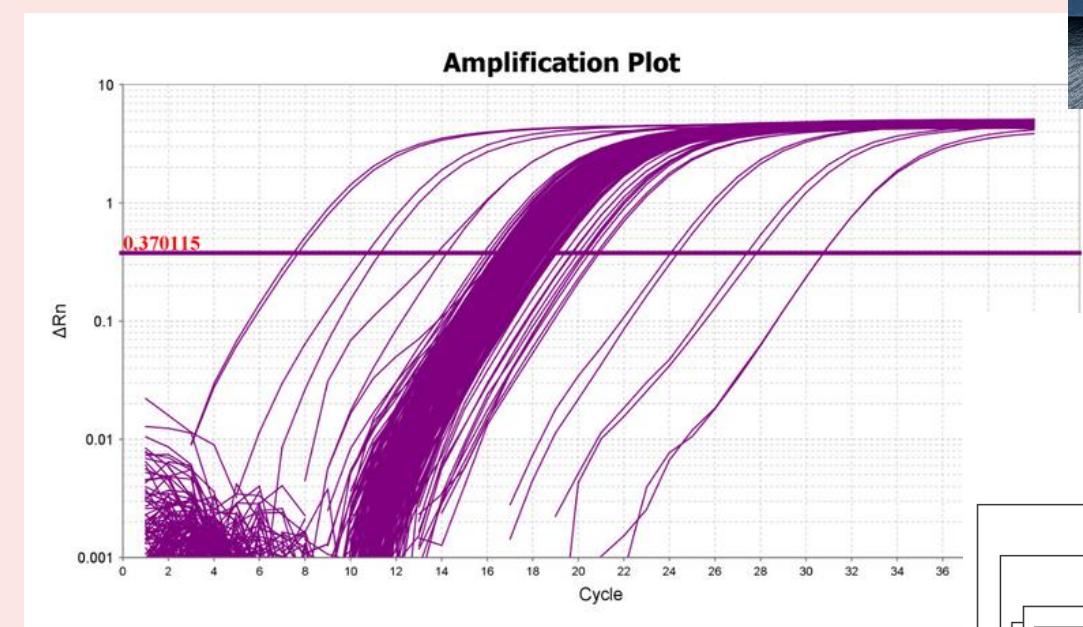
A BIT ABOUT ME



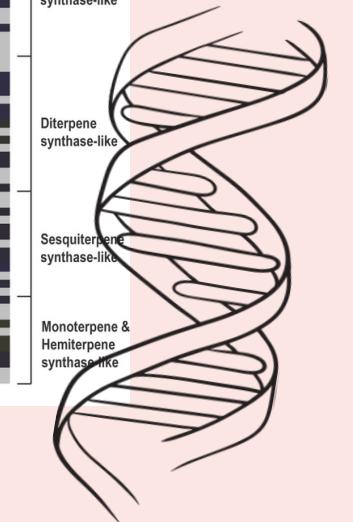
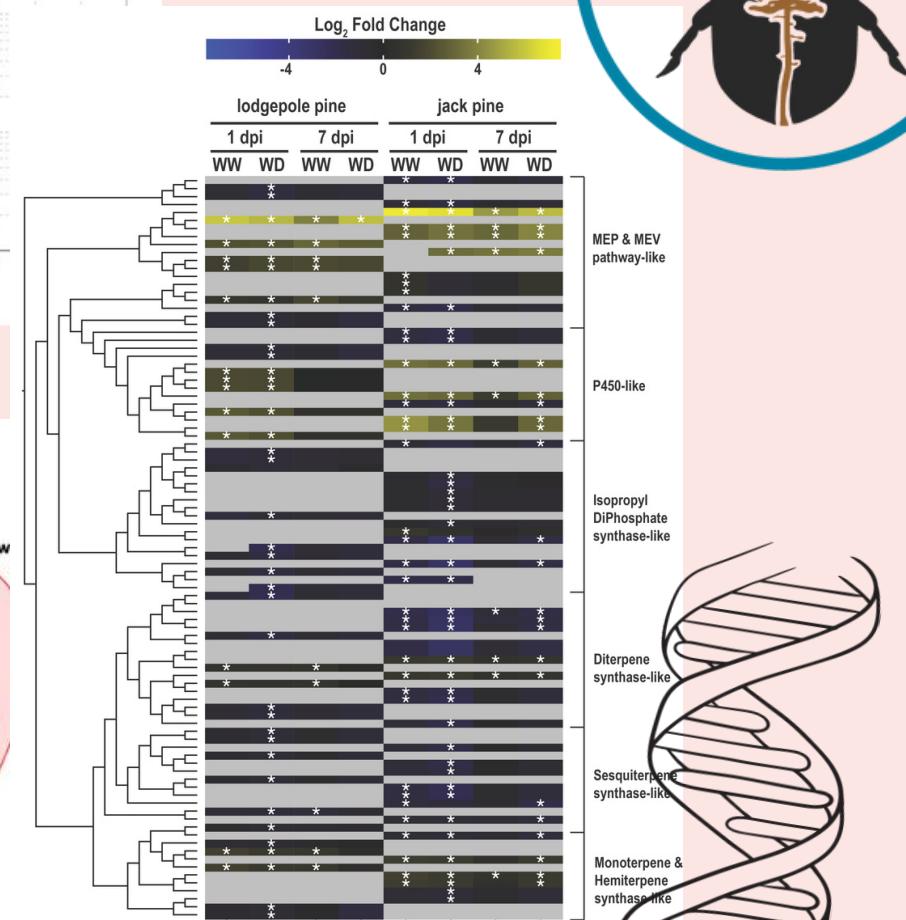
metabolomics



microscopy



transcriptomics



BEST DATA PRACTICES

 **F**indable

 **A**ccessible

 **I**nteroperable

 **R**Reusable

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and **persistent identifier** 
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a **standardized communications protocol**
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and **broadly applicable language** for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are **richly described** with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with **detailed provenance**
 - R1.3. (meta)data meet domain-relevant community standards

**R is great
for this!**

HAVING A PLAN

Tri-Agency Research Data Management Policy

4. Implementation dates

The agencies plan to implement the policy incrementally, as determined through ongoing engagement with the research community and other stakeholders, and in step with the continuing development of research data practices and capacity in Canada and internationally.

- **Institutional strategies:** By March 1, 2023, research institutions subject to this requirement must post their RDM strategies and notify the agencies when they have been completed.
- **Data management plans:** By spring 2022, the agencies will identify the initial set of funding opportunities subject to the DMP requirement. The agencies will pilot the DMP requirement in targeted funding opportunities before this date.
- **Data deposit:** After reviewing the institutional strategies and in line with the readiness of the Canadian research community, the agencies will phase in the deposit requirement.
 - **CIHR-funded researchers:** Since January 1, 2008, recipients of CIHR funding have had to comply with the limited data deposit requirements included in the Tri-Agency Open Access Policy on Publications. They must continue to comply with these requirements, which are specific to bioinformatics, atomic, and molecular coordinate data.

The policy will be updated as the implementation dates for requirements are further specified.



NSERC
CRSNG



Institutional/Organizational Strategy

Project Data Management Plan

Data Collection & Analysis

Data Deposit

Publication

MAKING DATA FINDABLE



borealis

UAlberta Research Data Collection
(University of Alberta)

Borealis > University of Alberta Borealis-Dataverse > UAlberta Research Data Collection >

Data for: Evidence that Ophiostomatoid fungal symbionts of mountain pine beetle do not play a role in overcoming lodgepole pine defenses during mass attack

Version 1.1

Fortier, Colleen, 2024, "Data for: Evidence that Ophiostomatoid fungal symbionts of mountain pine beetle do not play a role in overcoming lodgepole pine defenses during mass attack", <https://doi.org/10.5683/SP3/QNUVEZ>, Borealis, V1, UNF:6:Vn8QbqhubcsGhcv9daQU8g== [fileUNF]

Cite Dataset ▾ Learn about [Data Citation Standards](#).

Description ▾ Data for the publication "Evidence that Ophiostomatoid fungal symbionts of mountain pine beetle do not play a role in overcoming lodgepole pine defenses during mass attack" (doi: 10.1094/MPMI-06-23-0077-R). All statistical analyses were conducted with R v4.2.3 (RCoreTeam 2023) and RStudio v2021.9.0.351 (RStudio Team 2021), and all plots were generated using ggplot2 v3.4.2 (Wickham 2016) from the tidyverse package v2.0.0 (Wickham et al. 2019), cowplot v1.1.1 (Wilke 2020), and ggpubr v0.6.0 (Kassambara 2023) packages for R. For additional details about analyses and software used, please consult the publication. (2024-01-16)

Access Dataset ▾ Contact Owner Share

Dataset Metrics ⓘ 39 Downloads ⓘ



FRDR

Federated **Research**
Data Repository



Government
of Canada

Gouvernement
du Canada

[Canada.ca](#) > [Science and innovation](#) > [Federal Science Libraries Network](#)

Federal Open Science Repository of Canada

From: [Federal Science Libraries Network](#)



National Center for
Biotechnology Information



MAKING DATA ACCESSIBLE

METADATA!



README file(s): description of the dataset, how was it generated, sharing and access info, acknowledgements



Data Dictionary: defines data terms, column names, codes, abbreviations, acronyms, etc.



File Naming: short and relevant, consider versioning, be consistent

UBC Library Research Commons
“How to Create a README File”

Harvard Medical School
“File Naming Conventions”

York U “Naming for Electronic
Files and Folders”



MAKING DATA ACCESSIBLE

METADATA!



Directory Structure

Think about this as soon as you collect your data

- keep raw data separate, work with copies
- having designated locations for things will prevent data loss!

Name	Size	Kind
> Data	--	Folder
> Graphs	--	Folder
M050_Hormone_RAnalysis.Rproj	205 bytes	R Project
> R_scripts	--	Folder
README_M050_Hormone_RAnalysis.txt	739 bytes	text
> Tables	--	Folder

MAKING DATA ACCESSIBLE AND *INTEROPERABLE*



Directory Structure

Consider using R Projects!
(require RStudio)



Name	Size	Kind
> Data	--	Folder
> Graphs	--	Folder
> M050_Hormone_RAnalysis.Rproj	205 bytes	R Project
> R_scripts	--	Folder
README_M050_Hormone_RAnalysis.txt	739 bytes	text
> Tables	--	Folder

- work with git for version control
- easily shareable with collaborators

MAKING DATA *INTEROPERABLE*



Annotations

R version

package versions

what functions do

```
1 #R.version·4.4.1·last·run·by·Colleen·Apr·15, ·2025·
2 ·
3 #install.packages("tidyverse")·#only·run·once, ·don't·typically·need·to·include·in·script·
4 #repeat·for·any·other·packages·you·don't·have·installed·already·
5 ·
6 library(tidyverse)·#v2.0.0·
7 library(janitor)·#v2.2.1, ·for·clean_names()·
8 library(scales)·#v1.3.0, ·for·scale_x_date()·on·graph·
9 ·
10 #read·in·the·untidy·dataset·
11 untidy_seminar_data<-·read_csv("data/untidy_seminar_data.csv")·|>·
12 #tidy·up·the·column·names·
13 ..janitor::clean_names()·|>·
14 #replace·the·"omit"s·to·NAs·
15 ..mutate(across(where(is.character), ·~na_if(.,·"omit")))|>·
16 #now·that·the·omits·are·gone, ·change·the·weight·columns·to·all·be·numeric·type·
17 ..mutate(across(bud_1_wax_mg:bud_3_dry_weight_mg, ·~as.numeric(x)))|>·
18 #change·date·type·so·that·R·recognizes·format·(helpful·for·graph·later)·
19 ..mutate(date_collected=·as.Date(date_collected, "%b%d-%Y"))·
20 ¶
```

MAKING DATA *INTEROPERABLE*



Consider using RMarkdown (in RStudio)

- Can run multiple languages (R, python, bash, SQL) as “chunks”

```
Fig1_Hormone_RAnalysis - main - RStudio
M050-2P_hormones_stats_summary.Rmd
Go to file/function Addins Environment History Conne
Source Visual
1 ---  
2 title:: "M050-2P.Hormones.Stats.Summary--ACC"  
3 author:: "Report.generated.by.Colleen.Fortier"  
4 date:: `r Sys.Date()`  
5 output:--  
6 ..pdf_document:--  
7 ....keep_tex::FALSE  
8 ---  
9  
10 ``{r,setup,.include=FALSE}--  
11 knitr::opts_chunk$set(echo=.FALSE)--  
12 ``--  
13  
14 ``{r.load.packages,.message=.FALSE,.echo=.TRUE}--  
15 #R.version.4.2.3--  
16 library(tidyverse).#v2.0.0.includes.ggplot2--  
17 library(hablar).#v0.3.2.convert.function.to.specify.data.as.numerical--  
18 library(lme4).#v1.1-32--
```

The RStudio interface shows the following components:

- Top Bar:** Shows the project name "Fig1_Hormone_RAnalysis - main - RStudio".
- Toolbar:** Includes standard icons for file operations like Open, Save, Print, and a "Knit on Save" button.
- Code Editor:** Displays the RMarkdown code with syntax highlighting for different languages.
- Language Selector:** A dropdown menu next to the "Knit" button lists supported languages: R, Python, GraphViz, Mermaid, Bash, Julia, Rcpp, SQL, and Stan.
- Environment Tab:** Shows the global environment is empty.
- Files Tab:** Shows the project structure: Hormone_RAnalysis > R_scripts.

MAKING DATA INTEROPERABLE



Consider using RMarkdown (in RStudio)

- Can “knit” into html or pdf report, showing all code and output

All tools Edit Convert E-Sign Find text or tools Share Ask AI Assistant

M050-2P Hormones Stats Summary - ACC

Report generated by Colleen Fortier

2023-04-04

```
#R version 4.2.3
library(tidyverse) #v2.0.0 includes ggplot2
library(hablar) #v0.3.2 convert function to specify data as numerical
library(lme4) #v1.1-32
library(emmeans) #v1.8.5 #estimated marginal means (least square means) for contrasts
library(multcomp) #v1.4-23 #needed for cld results from emmeans
library(car) #v3.1-2 #Anova function for glms
```

INOC GLM

Fit your data to a glmm - may need to change family to better fit the model to the data/meet model assumptions

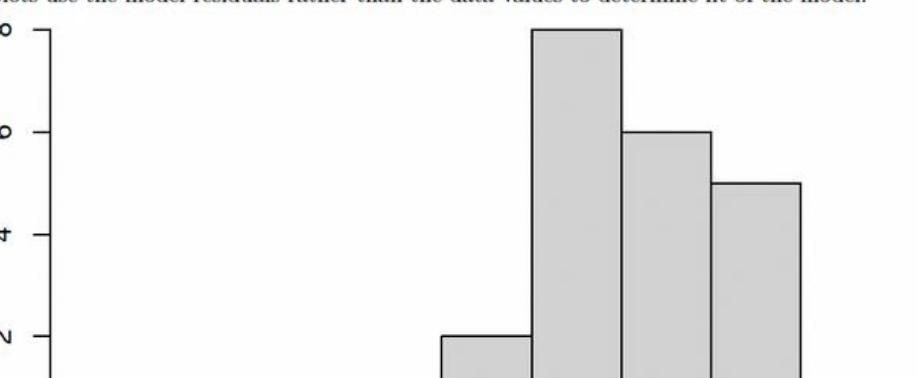
```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
## Gauss-Hermite Quadrature, nAGQ = 0) [glmerMod]
## Family: Gamma ( log )
## Formula: QTY ~ treatment * timepoint + (1 | tree_number)
## Data: inoc_data
##      AIC      BIC   logLik deviance df.resid
##  285.7404 292.8087 -136.8702  273.7404     18
## Random effects:
## Groups      Name        Std.Dev.
## tree_number (Intercept) 0.3531
## Residual            0.3614
## Number of obs: 24, groups: tree_number, 16
## Fixed Effects:
```

CHECK THAT YOUR DATA MEETS ASSUMPTION OF YOUR MODEL

```
##
## Shapiro-Wilk normality test
##
## data: inoc_model_residuals
## W = 0.89898, p-value = 0.02048
##
## Bartlett test of homogeneity of variances
##
## data: inoc_model_residuals and inoc_data$timepoint
## Bartlett's K-squared = 5.1333, df = 1, p-value = 0.02347
##
## Bartlett test of homogeneity of variances
##
## data: inoc_model_residuals and inoc_data$treatment
## Bartlett's K-squared = 0.87047, df = 1, p-value = 0.3508
```

Plots examining distribution of model residuals

These plots use the model residuals rather than the data values to determine fit of the model.

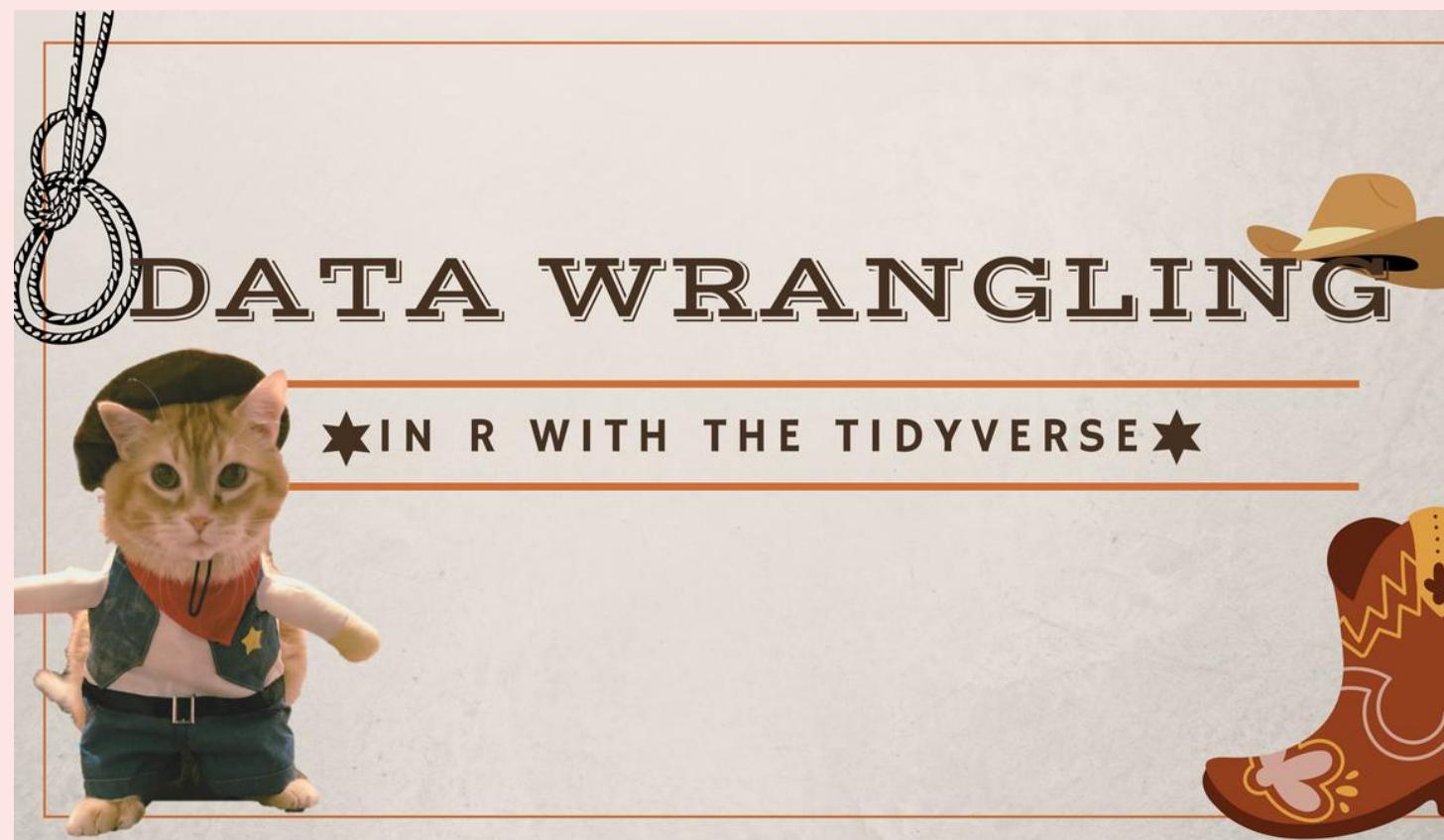


MAKING DATA RESUABLE

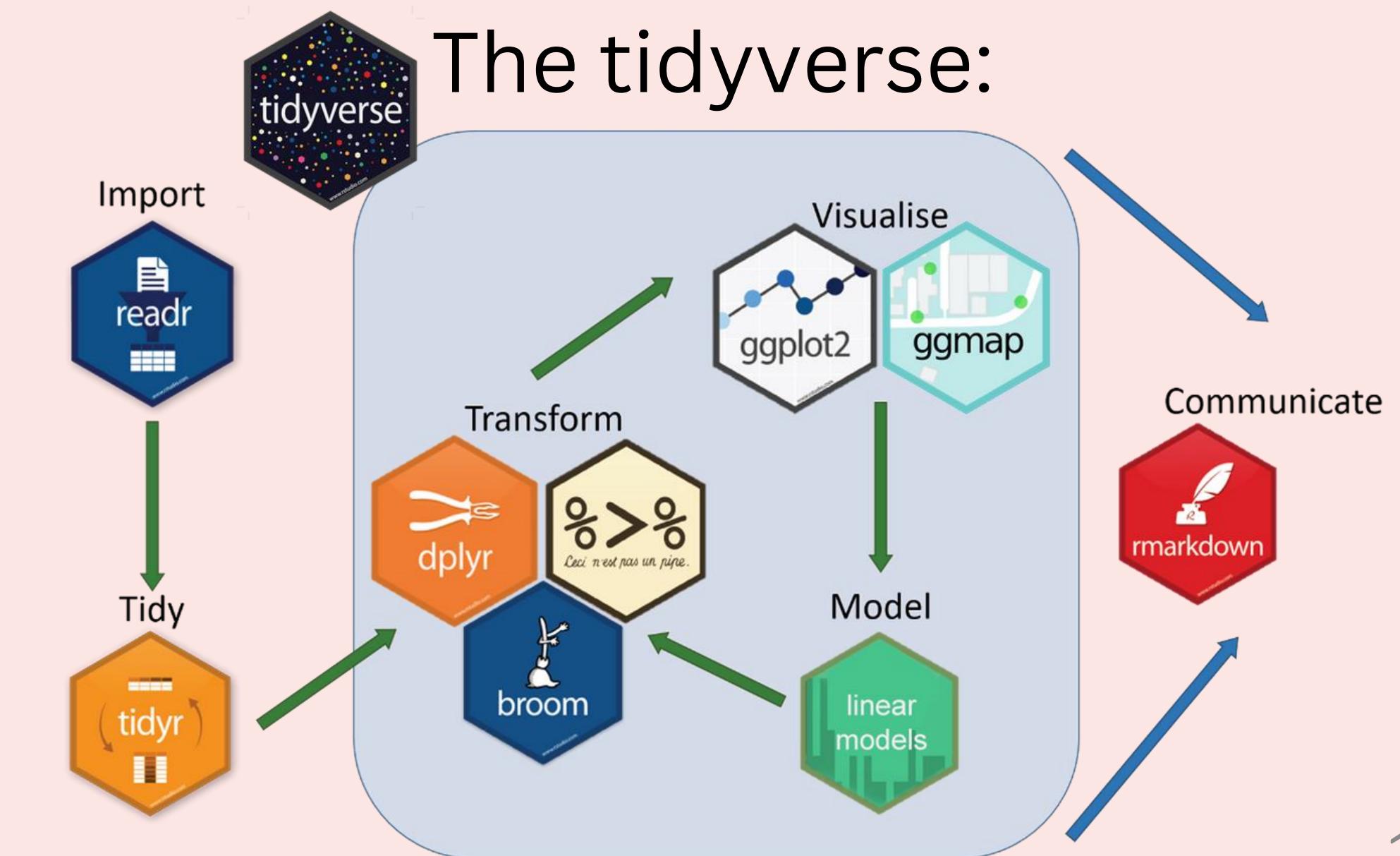


Use Smarter Code and TIDY YOUR DATA!

- why use many lines when few line do trick? (Kevin, *The Office*)



Slides and RProject with data and scripts to try out tidyverse! ↪

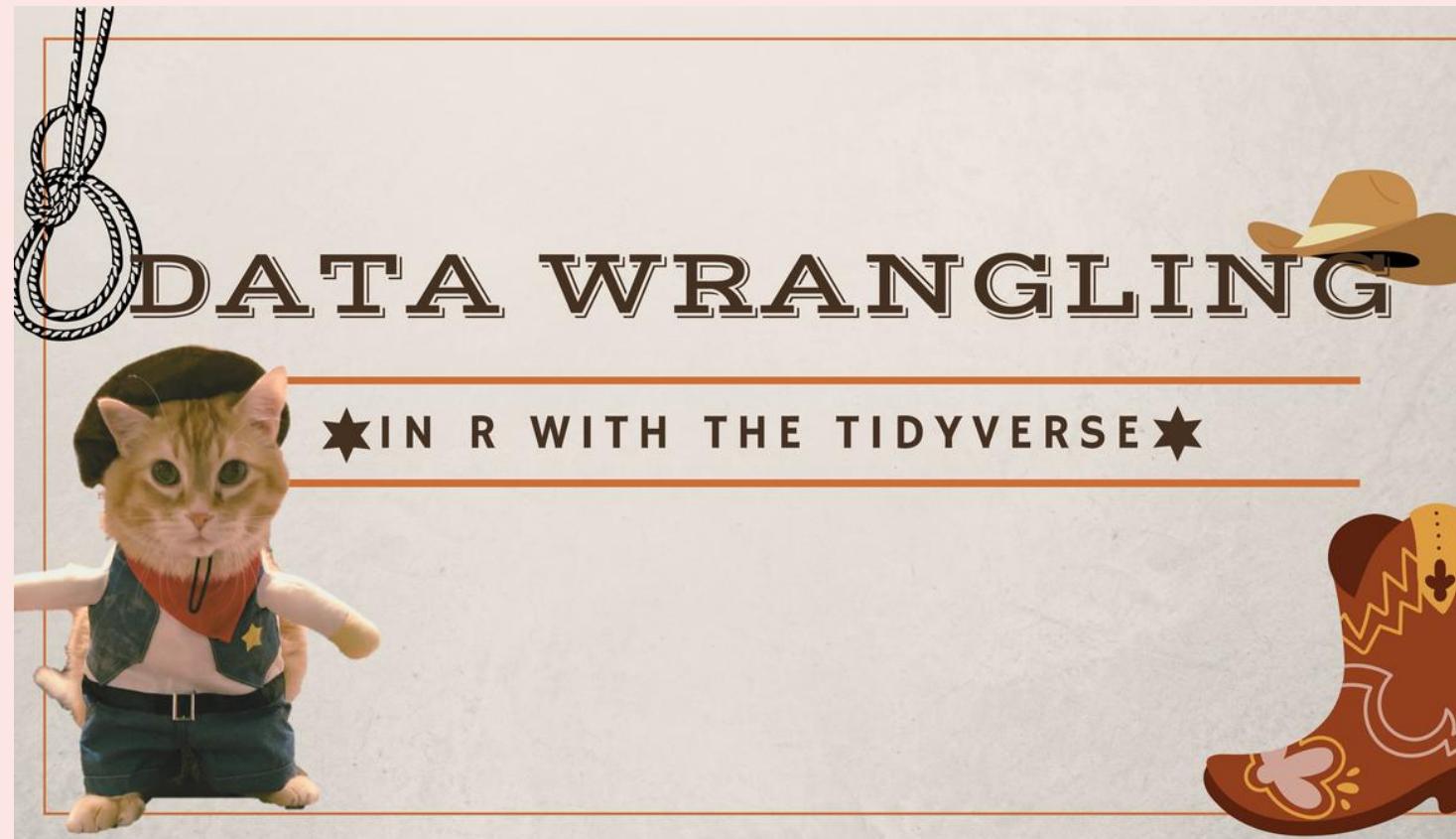


MAKING DATA RESUABLE



Use Smarter Code and TIDY YOUR DATA!

- why use many lines when few line do trick? (Kevin, *The Office*)



#Base R:

```
>names(df)[names(df) == "old_name"] <- "new_name"
```

#tidyr:

```
>rename(df, new_name = old_name)
```

#if your old name has any spaces, make sure to include " "

Slides and RProject with data and
scripts to try out tidyverse! 

MAKING DATA RESUABLE



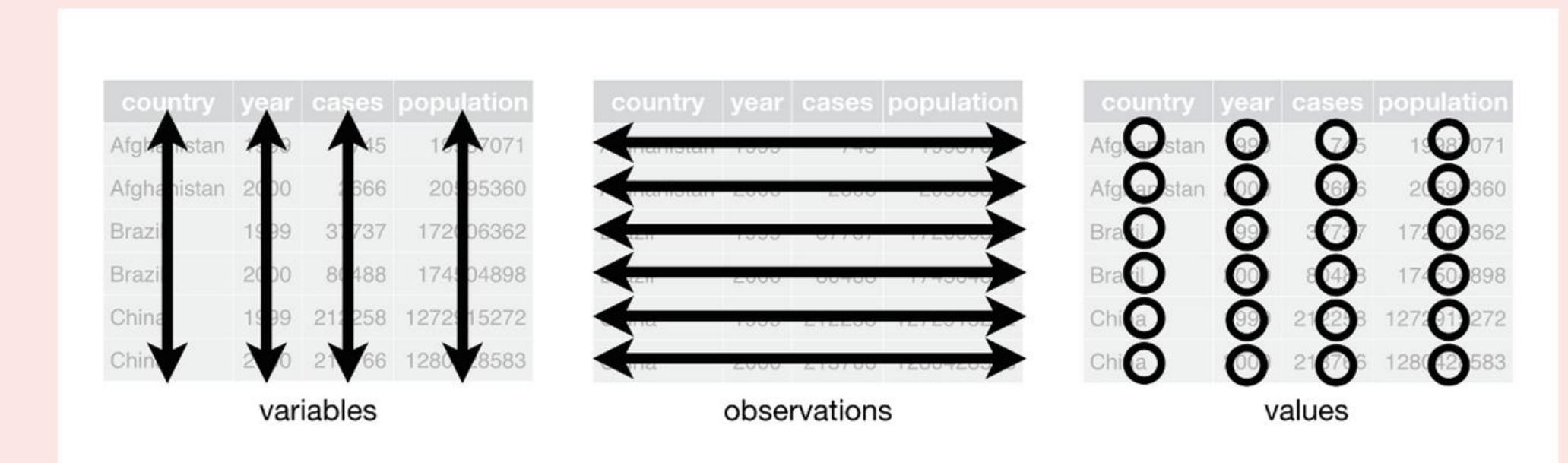
Use Smarter Code and TIDY YOUR DATA!

- why use many lines when few line do trick? (Kevin, *The Office*)



Slides and RProject with data and scripts to try out tidyverse!

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



MAKING DATA RESUABLE



Use Headers and Annotations to break up script

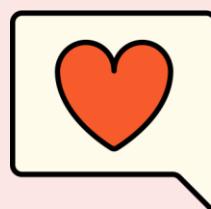
- RStudio recognizes 4 dashes (----) as cues for headers

The screenshot shows an RStudio interface with a script file open. The code uses four dashes (----) to create section headers:

```
65 .....percent_water_treatment_avg..=mean(avg_percent_water_per_tree),-- DATA ORGANIZATION
66 .....wax_mg_treatment_avg..=mean(avg_wax_per_tree_mg)-- SUMMARY TABLE
67 -
68 #export.datasets.as.csv--use.this.new.file.for.additional.analyses.so.don't
69 write_csv(tidier_data,"data/tidier_seminar_data.csv")-
70 -
71 -
72 #####----SUMMARY TABLE---- library(gt)-
73 -
74 -
75 #data.is.piped.directly.into.table-
76 summarized_data.|>-
77 #replace._.in.column.names.with.spaces.so.headers.look.nicer-
78 ..rename_with(~gsub("_",".",.,x,.fixed=TRUE),.ends_with("avg")).|>-
79 ..gt()-
80 |
```

A red arrow points from the word "SUMMARY" in line 72 to the word "SUMMARY" in the "SUMMARY TABLE" header. A blue arrow points from the "SUMMARY TABLE" header to the "Global Environment" tab in the right sidebar, which lists the datasets: sample..., summar..., tidier..., and untidy... with their respective sizes and variable counts.

MAKING DATA RESUABLE



Use Headers and Annotations to break up script

- continuous code lines will still run even if broken up by spaces or annotations - just don't put a # in front of the code you want to run

The screenshot shows an RStudio interface with a dark theme. The left pane displays an R script named "Tidy_seminar_script_Apr17-2025.R". The code contains numerous pipes (|>) used for data manipulation. A large teal curly brace groups the entire script under the heading "all one code chunk". A teal arrow points from the word "pipes!" to one of the pipes in the code. The right pane shows the "Global Environment" tab, which lists a dataset named "untidy_seminar_data" with 97 observations and 7 variables.

```
10 #read.in.the.untidy.dataset:-
11 untidy_seminar_data <- read_csv("data/untidy_seminar_data.csv") |>...
12 #tidy.up.the.column.names-
13 ..janitor::clean_names() |>...
14 #replace.the."omit"s.to.NAs-
15 ..mutate(across(where(is.character), ~na_if(., "omit"))) |>...
16 #now.that.the.omits.are.gone, .change.the.weight.columns.to.all.be.numeric.type-
17 ..mutate(across(bud_1_wax_mg:bud_3_dry_weight_mg, ~as.numeric(.x))) |>...
18 #change.date.type.so.that.R.recognizes.format.(helpful.for.graph.later)-
19 ..mutate(date_collected = as.Date(date_collected, "%b%d-%Y")) |>...
20 #pivot.the.weight.columns.longer,.so.only.one.column.for.each.type.of.weight.measurement-
21 ..#will.need.to.add.a.separate.column.for.bud.number-
22 ..pivot_longer(cols=c(bud_1_wax_mg:bud_3_dry_weight_mg), ...
23 .....names_to = c("bud", ".value"), names_pattern = "bud_(.)_(.*)" ) |>...
24 #filter.out.rows.with.NA.values.for.weight.(check.that.this.applies.to.all.measurements.first!)-.
25 ..dplyr::filter(!is.na(wax_mg))|
```

pipes!

all one code chunk

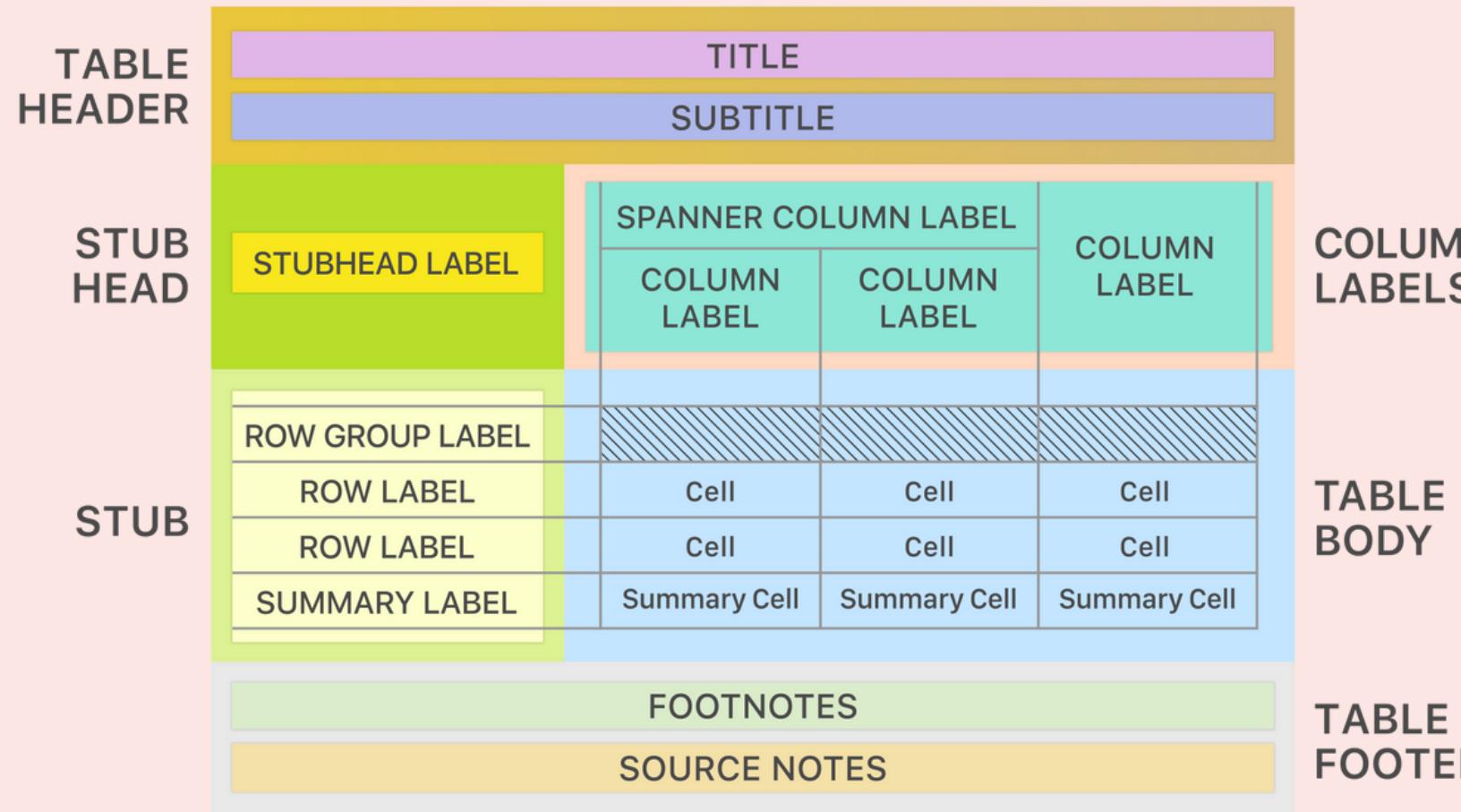
MAKING DATA RESUABLE



Try to code as much as you can

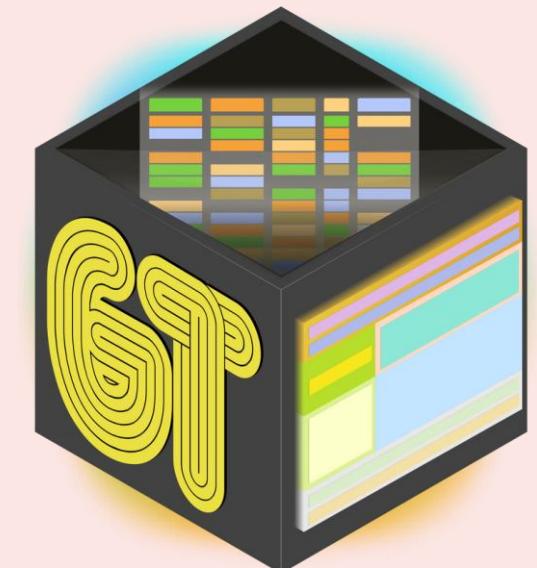
- avoid copy + paste between programs, or editing outputs
- use packages like gt to create results tables IN YOUR SCRIPT

The Parts of a gt Table



The title of the table						
The table's subtitle						
	num	char	fctr	date	time	datetime
grp_a						
row_1	1.111e-01	apricot	one	2015-01-15 13:35	2018-01-01 02:22	49.950
row_2	2.222e+00	banana	two	2015-02-15 14:40	2018-02-02 14:33	17.950
row_3	3.333e+01	coconut	three	2015-03-15 15:45	2018-03-03 03:44	1.390
row_4	4.444e+02	durian	four	2015-04-15 16:50	2018-04-04 15:55	65100.000
min	0.11	—	—	—	—	1.39
max	444.40	—	—	—	—	65,100.00
grp_b						
row_5	5.550e+03	NA	five	2015-05-15 17:55	2018-05-05 04:00	1325.810
row_6	NA	fig	six	2015-06-15 NA	2018-06-06 16:11	13.255
row_7	7.770e+05	grapefruit	seven	NA	19:10	2018-07-07 05:22
row_8	8.880e+06	honeydew	eight	2015-08-15 20:20	NA	0.440
total	—	—	—	—	—	66,508.79

¹ This is a footnote.
This is a source note.



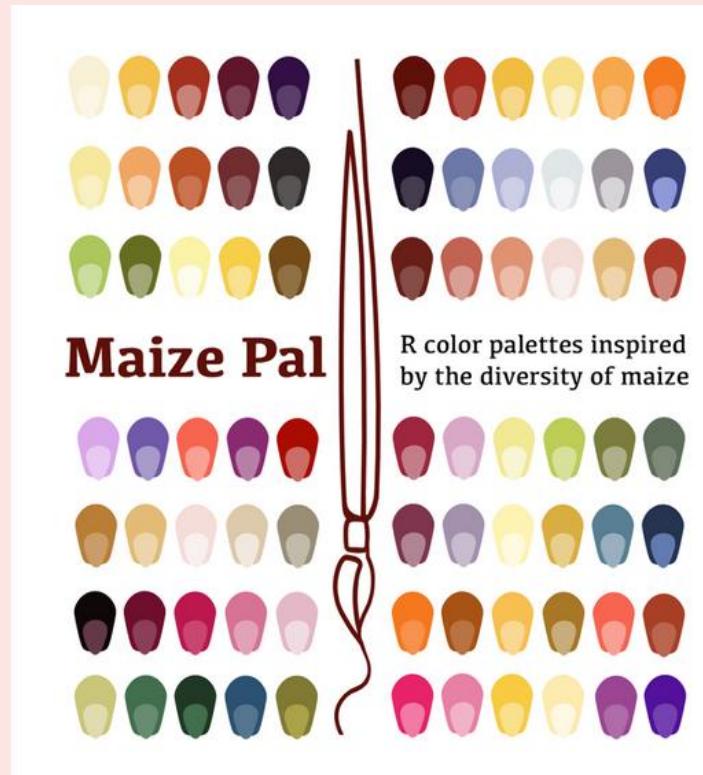
{gt} 0.7.0

MAKING DATA RESUABLE

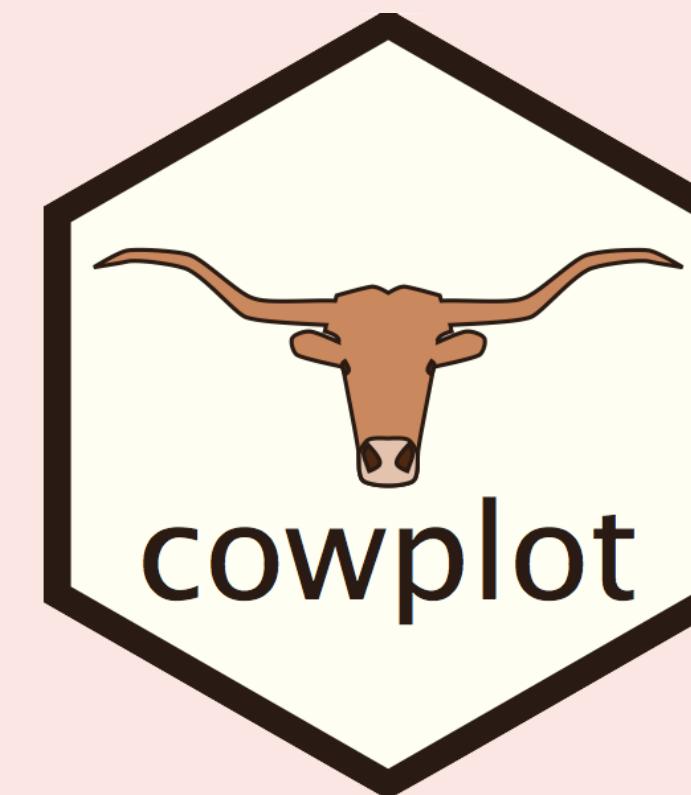


Try to code as much as you can

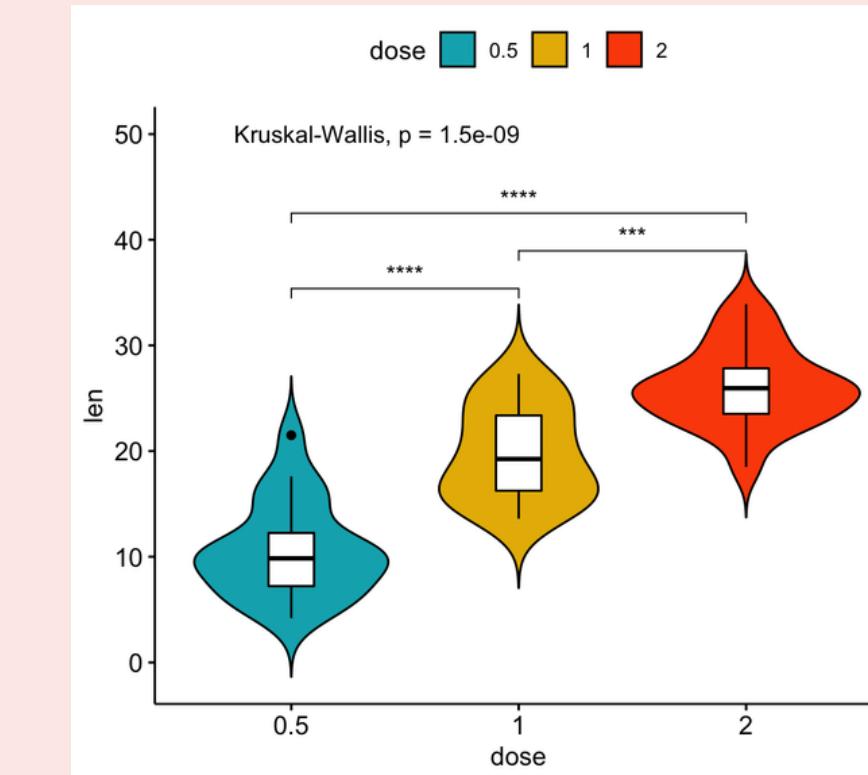
- avoid copy + paste between programs, or editing outputs
- explore packages that build on ggplot to do even more



Maize Pal
color palettes



combine plots, add axes
and labels with **cowplot**



add statistics and regression
lines with **ggpubr**

more
ideas!

MAKING DATA RESUABLE

more on citing
packages



Don't forget to cite your packages (and versions)!

The screenshot shows the RStudio interface with the following details:

- Code Editor:** The script file `Tidy_seminar_script_Apr17-2025.R` contains R code for generating a citation file. A green arrow points from the text "knitcitations package" to the line `library(knitcitations)`.
- Project Explorer:** The project name is `Tidy_seminar_Rproject`. It includes files for data, figures, r_scripts, and a main R project file.
- Environment:** The Global Environment pane shows data frames like `sample_m...`, `summariz...`, `tidier_d...`, and `untidy_s...`.
- Console:** The status bar at the bottom left shows the time as 110:45 and the number of citations as # CITATIONS.

knitcitations package

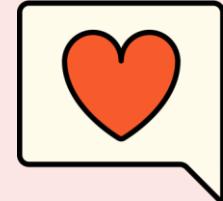
**create a .bib file with all current citations
(can import into citation managers)**

```
102 #export::the::ggplot,::specify::file::type::uri::uri::-->::URL::Save::the::most::recent::plot
103 ggsave("figures/tidier_data_plot.tiff",.dpi=600,.width=.5,.height=.5)
104 -
105 -
106 ####---CITATIONS---
107 #Generate.bib.files.for.each.of.your.current.packages...can.be.imported.to.ref.manager-
108 library(knitcitations).#v1.0.12..#for.write.bibtex.function-
109 -
110 #first.make.a.list.of.the.packages.you.used:-
111 pkgs<-c("tidyverse","janitor","gt","MaizePal","knitcitations")-
112 #can.also.use.the.packages().function.but.it.will.list.all.dependencies.as.well-
113 -
114 write.bibtex(do.call('c',lapply(pkgs,citation)),.file="tidy_seminar_pkg_citations.bib")-
115 #this.will.generate.a.bib.file.in.the.main.directory-
116 -
117 #best.to.use.the."import".option.in.zotero.(and.select..bib.file),.I.find.that.it.doesn't.drag.&.drop.well-
118 #includes.the.package.version.as.a.note.in.the.citation.in.zotero-
119
```

LAST THOUGHTS



Make a plan - get everyone involved



REVISIT THE PLAN REGULARLY



Share your code with others - they are the best at finding errors/where more annotation is needed



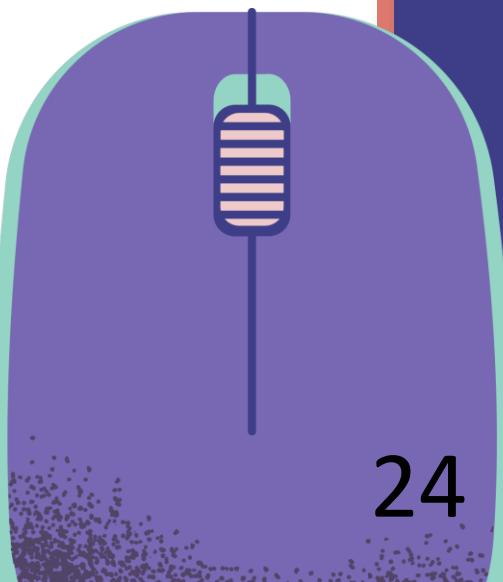
Schedule time periodically to curate your data

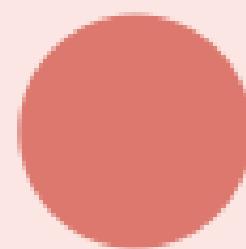


ADDITIONAL RESOURCES



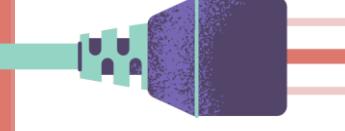
- ≡ My Github, where I post presentations + materials
- ≡ UBC Library Research Commons FREE Workshops
these have great resources
including a whole section on Data Management
- ≡ Digital Research Alliance YouTube
and newsletter signup for workshops



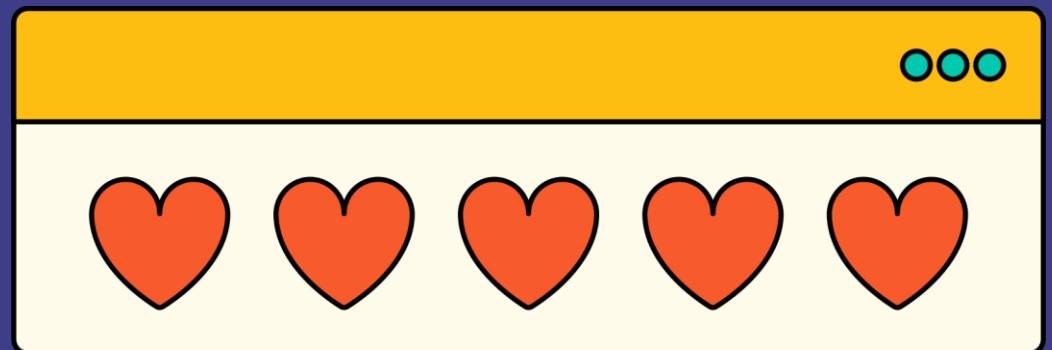


ADDITIONAL RESOURCES



-  [UBC Advanced Research Computing Workshops](#)
-  [UAlberta Research Bootcamp Workshop Recordings](#)
-  [Compute Ontario Courses](#)
-  [SHARCNET Training](#)





THANKS FOR LISTENING!

ANY QUESTIONS?



You Have A
New Messages!

OK

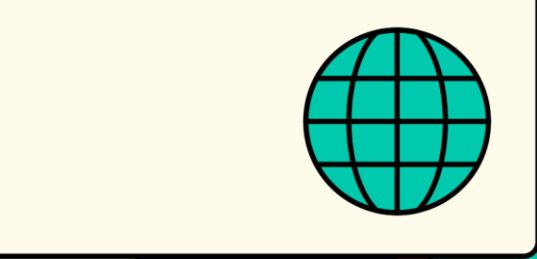


Lorem ipsum dolor sit amet,
consectetuer adipiscing elit, sed
diam nonummy nibh euismod

OK



LOADING...



...
OK

OK