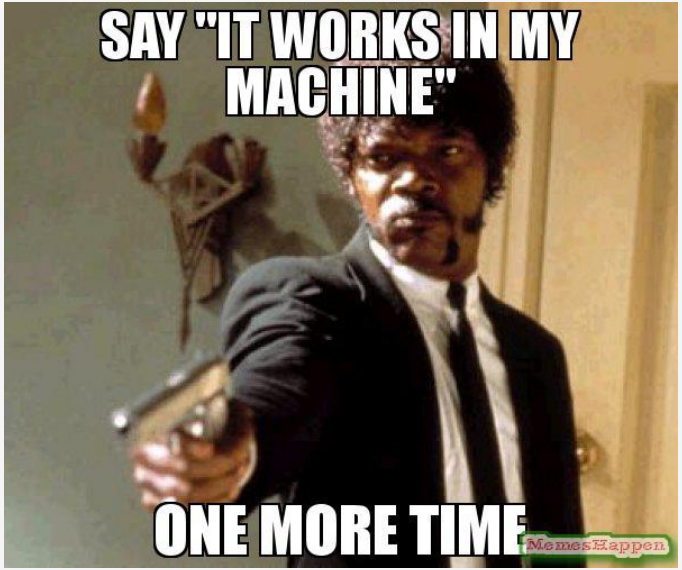
Optimist: The glass is ½ full.
Pessimist: The glass is ½ empty.
Excel: The glass is January 2nd.

Reproducible Research:
Moving From Excel to Scripting

source



Reproducibility in Research

source

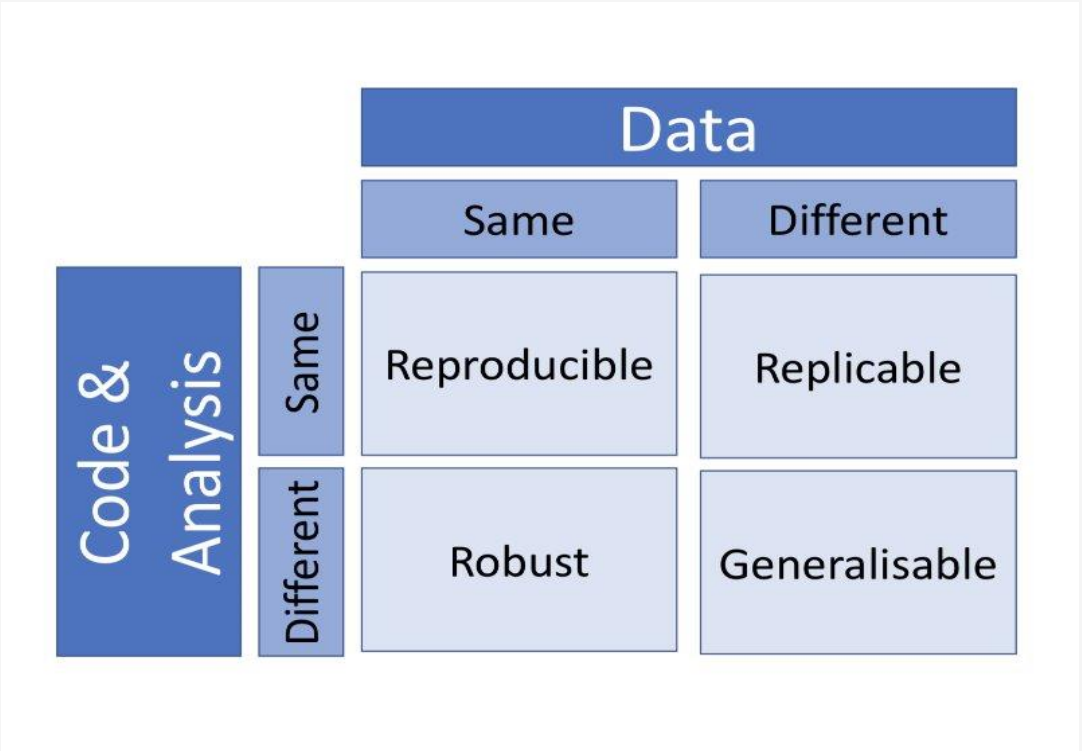
What is Reproducibility?

What is Reproducibility?

Reproducible

- The **existing data** can be reanalyzed using the **same research methods**, and will yield the same results

What is Reproducibility?



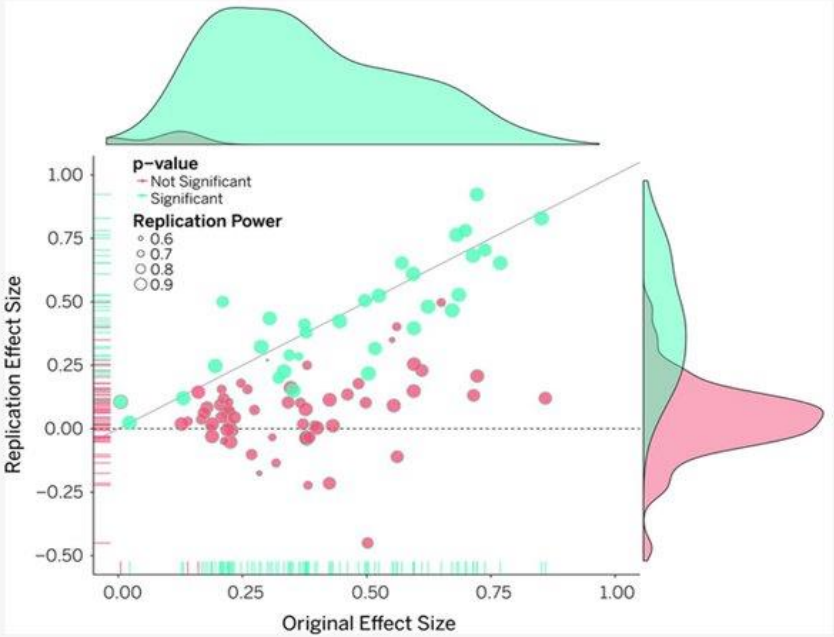
source

The Reproducibility Crisis

source

The Reproducibility Crisis

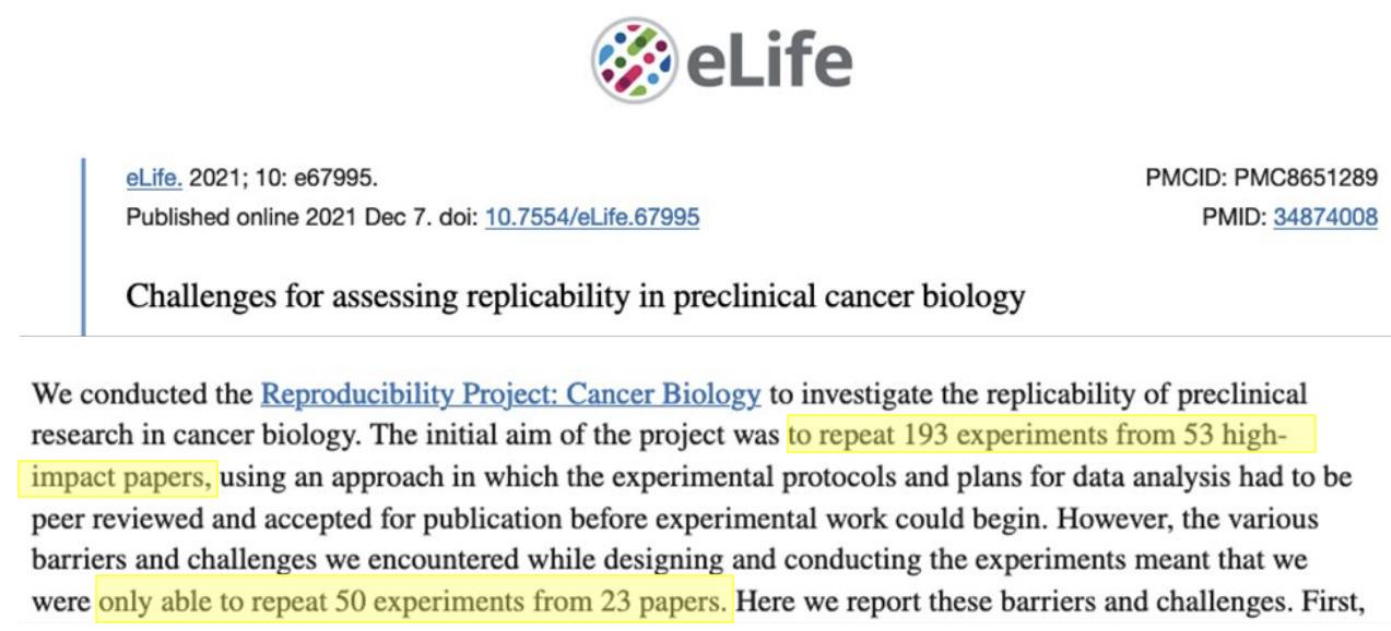
- 100 studies published in 3 psychology journals
- 36% of studies were able to be reproduced



Original study effect size versus replication effect size (correlation coefficients).
Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

source

The Reproducibility Crisis



eLife

eLife. 2021; 10: e67995.
Published online 2021 Dec 7. doi: 10.7554/eLife.67995
PMCID: PMC8651289
PMID: 34874008

Challenges for assessing replicability in preclinical cancer biology

We conducted the Reproducibility Project: Cancer Biology to investigate the replicability of preclinical research in cancer biology. The initial aim of the project was to repeat 193 experiments from 53 high-impact papers, using an approach in which the experimental protocols and plans for data analysis had to be peer reviewed and accepted for publication before experimental work could begin. However, the various barriers and challenges we encountered while designing and conducting the experiments meant that we were only able to repeat 50 experiments from 23 papers. Here we report these barriers and challenges. First,

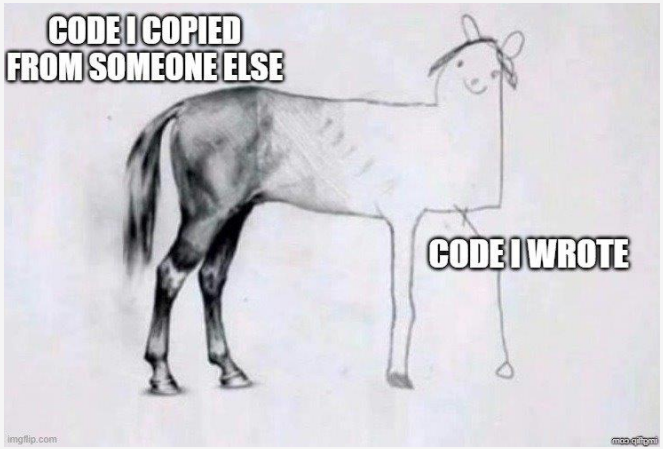
source

What Can We Do?

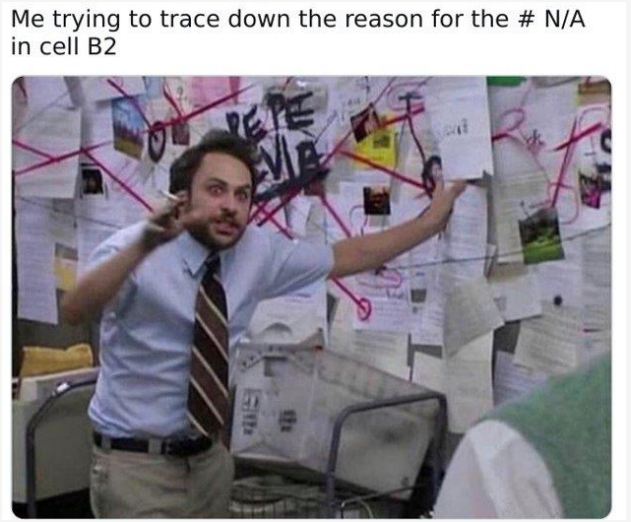
- Libraries provide a lot of support around academic/scholarly integrity.
- This largely consists of citation support and avoiding plagiarism, with the focus being **students as consumers of research.**
- However, as graduate students step into their theses and dissertations, we should be thinking of academic integrity through the lens of **students as producers of research.**
- **Data integrity, open and reproducible research!**

Nuances of Reproducibility

Computational vs. manual processes



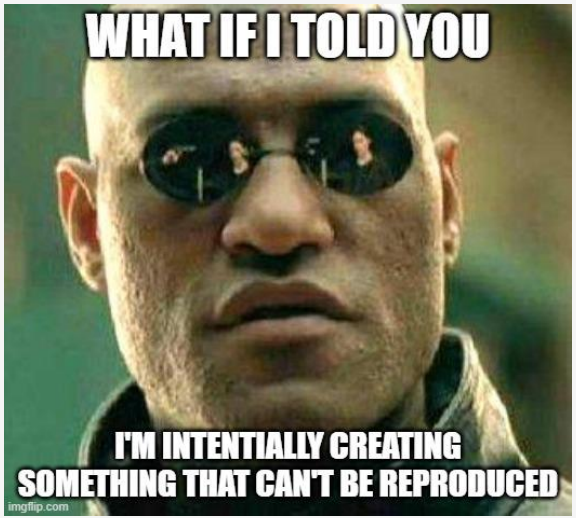
source



source

Nuances of Reproducibility

Arts, humanities, and creative outputs



source

Overarching Best Practices

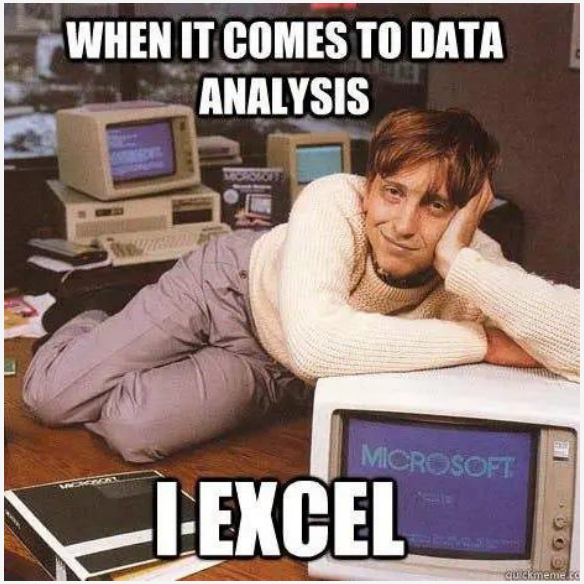
Your practices vs. the practices they told you not to worry about

source

Overarching Best Practices

FAIR Principles

- Sharing/preserving research data in a data repository (**Findable**)
- Ensuring materials are in open file formats (**Accessible, Interoperable**)
- Supplementing data with documentation that facilitates interpretability and reuse (**Reusable**)



Let's Talk About Excel!

source

Have you ever used Excel?

Have you ever used Excel?

- What did you use it for?
- Did it do a good job? Why/why not?

Diving Into Excel: What it does well

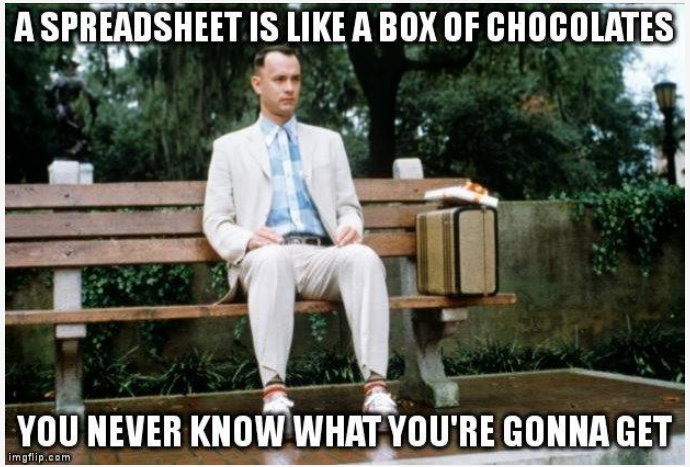
- Intuitive point-and-click graphical user interface (GUI)
- Easily group and order datasets based on specific values
- Easily calculate things like sum, mean, and other basic statistical methods
- Create graphs and charts with a few clicks
- **TLDR: it's easy and convenient**

Diving Into Excel: What it doesn't do so well

- Any manual changes made to a document are subject to human errors, and tracking these changes and identifying errors can be challenging.
- Formulaic changes are often hidden in cells and are subject to "breaking", making manipulations and analyses difficult to track and reproduce.
- Manipulations require manual processes for every file.

Diving Into Excel: What it doesn't do so well

- The auto-formatting of cells can be very annoying and cause issues.
- Rows/columns can be hidden, causing issues with collaboration.
- Changing variables will overwrite original data, causing issues with reproducibility and provenance.
- **TLDR: it's not very reproducible.**



What's in a Spreasheat?:

.xlsx vs .csv files

source

.xlxs Files

- Excel files (.xlxs) are a proprietary file format, meaning that the ability to use them relies on a subscription to the Microsoft suite.

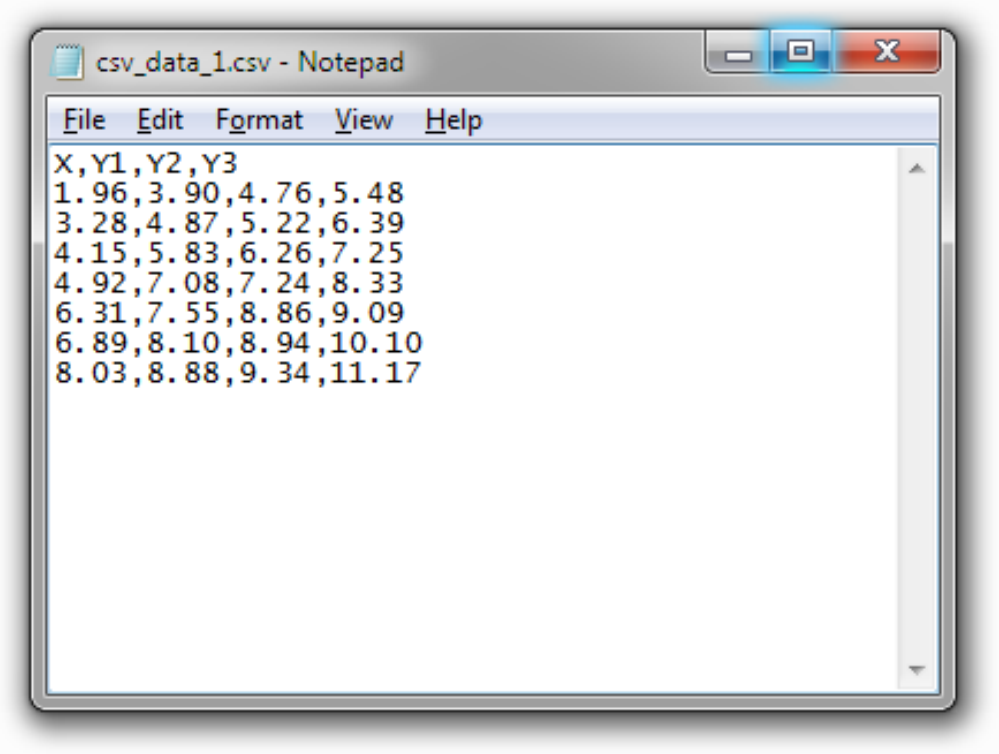
.xlxs Files

- While it's natural to think of an Excel file as a single file, in reality these files are zipped archives containing multiple underlying files:
    - Worksheet data
    - Worksheet formatting
    - Charts and graphs
    - Pivot tables
    - Embedded objects
    - External links
    - Metadata

.csv Files

- .csv files (Comma Separated Values) are tabular (spreadsheet) files, with each row representing a record and values within each row separated by commas (even though they can be presented in cells).

.csv Files


csv_data_1.csv - Notepad
File Edit Format View Help
X,Y1,Y2,Y3
1.96,3.90,4.76,5.48
3.28,4.87,5.22,6.39
4.15,5.83,6.26,7.25
4.92,7.08,7.24,8.33
6.31,7.55,8.86,9.09
6.89,8.10,8.94,10.10
8.03,8.88,9.34,11.17

.csv Files

- They are **non-proprietary plain text files,** meaning that they don't have any underlying structures like Excel files, and can be opened using any simple free text editor like Notepad, TextEdit, etc.

- Because of their simplicity, they are an efficient way to store and transfer tabular data because they are smaller in size, but with this they have limited formatting options.

- While it is possible to handle Excel files in coding languages like R and Python, .csv files are generally the preferred format due to their open and simple nature.

Example: Genetic Data

In a survey looking at over 11,000 papers with Excel gene lists published between 2014 and 2020, more than 30% contained at least one gene name error caused by Excel's autoformatting.

This was due to gene names being converted to standard dates, internal data numbers (5 digits), and floating point numbers.

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008984

Converting Excel Files to Other Formats

Original value/ format	What can happen	Example of change	Why it happens
01234	Leading zero lost	1234	Treated as a whole number
Résumé	Characters corrupted	Rsum / RÃ©sum Ã©	Encoding issues
31/12/2024	Date misread or changed	12/31/2024	Different regional settings for dates
=SUM(A1:A10)	Formula lost	55 (only value)	Plain text only keeps values
Hello	Formatting lost	Hello	Plain text doesn't store formatting
123456789012345	Rounded or scientific notation	1.23E+19 or rounded value	Precision limits

Moving to Scripting

source

Moving to Scripting: What is it?

- Writing lines of code, or "scripts", is creating a set of instructions that a coding language can perform.
- For the purpose of this program, these instructions are performed on a dataset (noting there can be broader usage).

Literate Programming

- Scripting was just referenced as instructions for a coding language/computer to perform, which is true, but a big focus of this series is making our work reproducible, which needs to include a second element of **instructions for humans to interpret.**

Literate Programming

- Literate programming is a framework that provides a human-language explanation of how a script works in combination with the script itself, so people can accurately interpret and reuse the script.

- We've already discussed how to document our files and data, but as we move through the program we're also going to talk about implementing documentation in our code and literate coding principles.

Moving to Scripting: Why you wouldn't use it

- Learning a coding language is like learning a human language, and can be a painful process.
- Because the things we do with tabular data can seem like isolated and time-sensitive activities ("I need this chart now!"), the effort needed to learn how to perform a new task can take much longer than just doing it in Excel.

Moving to Scripting: Why you wouldn't use it

- Even if you are familiar with a coding language, it can still be a struggle to figure out the exact way to do something.
- **TLDR: it can be a pain in the a**.**

Moving to Scripting: Why you would use it

- Other than the learning curve, scripting overcomes all of the downfalls of Excel:
    - There are no "manual" changes, and everything can be tracked and checked.
    - Scripts are kept in a document/documents and provide a clear and ordered list of everything that was done.
    - A single script can be performed across large amounts of similar files, and chunks can be taken from one script and altered slightly to perform the same task on different datasets.
    - No hidden rows or columns.
    - Can create new variables that are derived from existing variables, so the original data isn't overwritten.

Moving to Scripting: Why you would use it

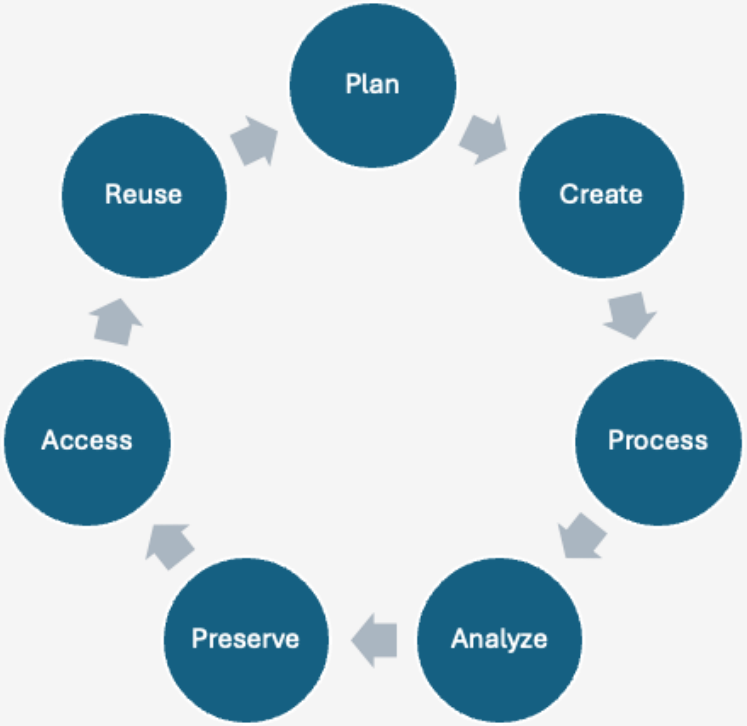
- The actions you can perform with scripts can also be far more precise and powerful than what Excel is capable of.
- **TLDR: it's powerful and reproducible.**

A Relevant (?) Example

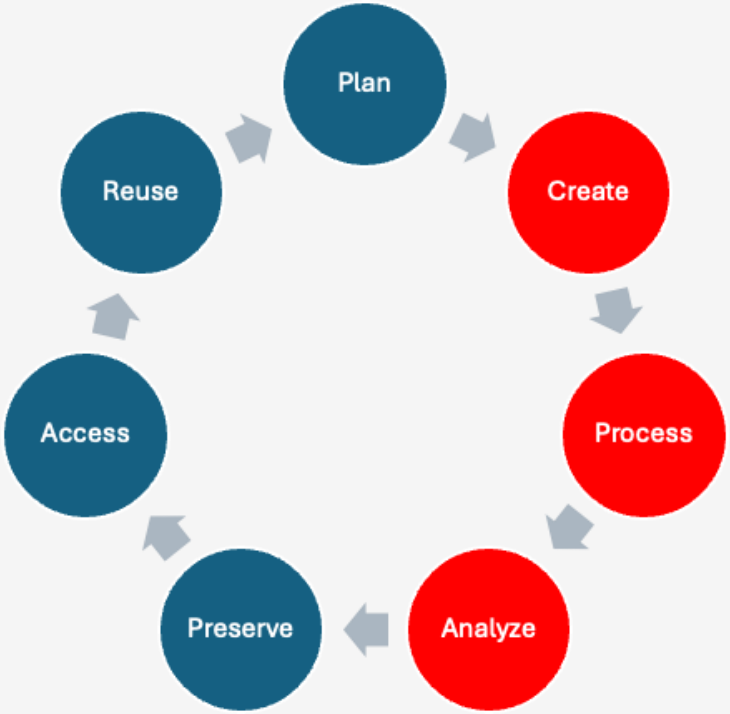
The Data Science Workflow

source

The Research Data Lifecycle



The Research Data Lifecycle



The Data Science Workflow

- A framework to support transparent and open research processes throughout the active phases of the research data lifecycle.
- While the research data lifecycle is focused on making research outputs FAIR, the data science workflow focses on asking questions from data in an open and reproducible way.

The R Coding Language

- R is an open-source programming language designed for data manipulation, statistical analysis, and visualization in research.
- Enables all of these activities to be easily reproducible, allowing for efficient work across data (many) data files, as well as sharing with others to reproduce and build off your work.
- Promotes the transparency of research and academic integrity.

RStudio

- While R is the programming language, RStudio is an Integrated Development Environment (IDE), which is a fancy term for a software application that makes working in R much more user-friendly.

RStudio

A (bad?) analogy could be:

- You are writing a document in Microsoft Word. In this case, the plain human language that you are typing would be R, and Microsoft Word, which is the software that allows you to view, format, add visuals, and save your work, is RStudio.

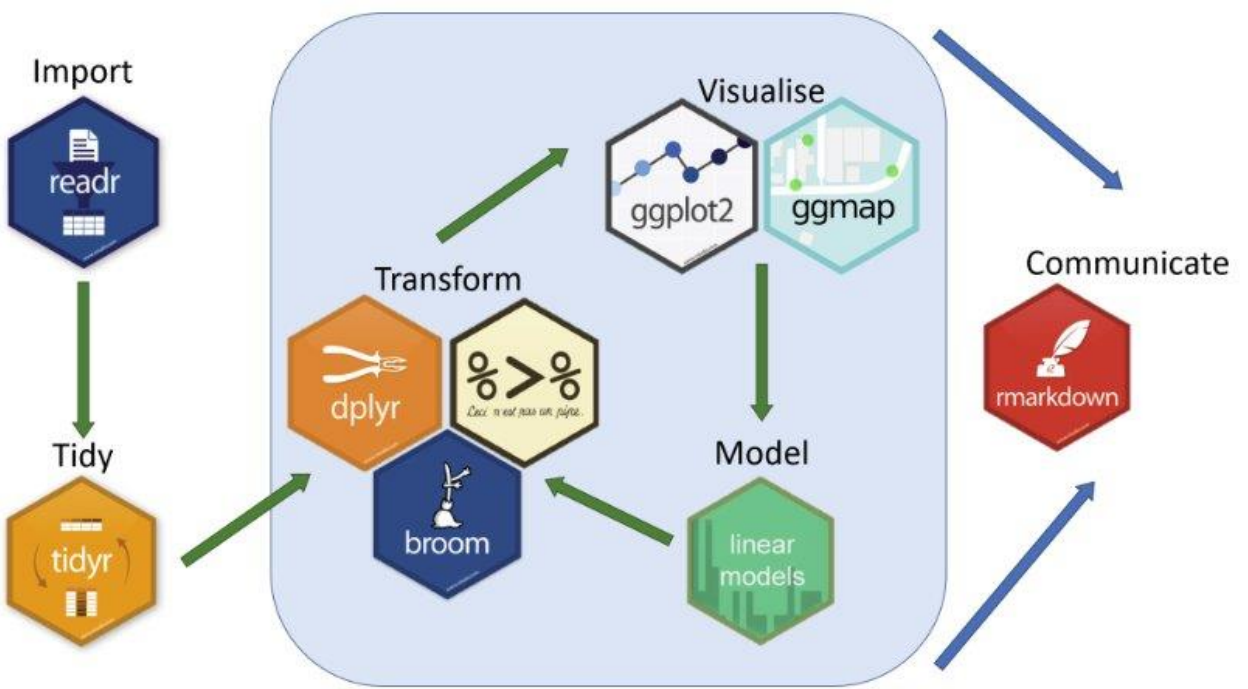
Across the Tidyverse

source

The Tidyverse

- A group of R "packages" for data manipulation, exploration, and visualization.
- Focused on the connections between the activities that make reproducible workflows possible.
- Provides a good reference point for those new to coding to engage with computational processes.

The Data Science Workflow with Tidyverse



source

Questions?