Data science in Unix and R

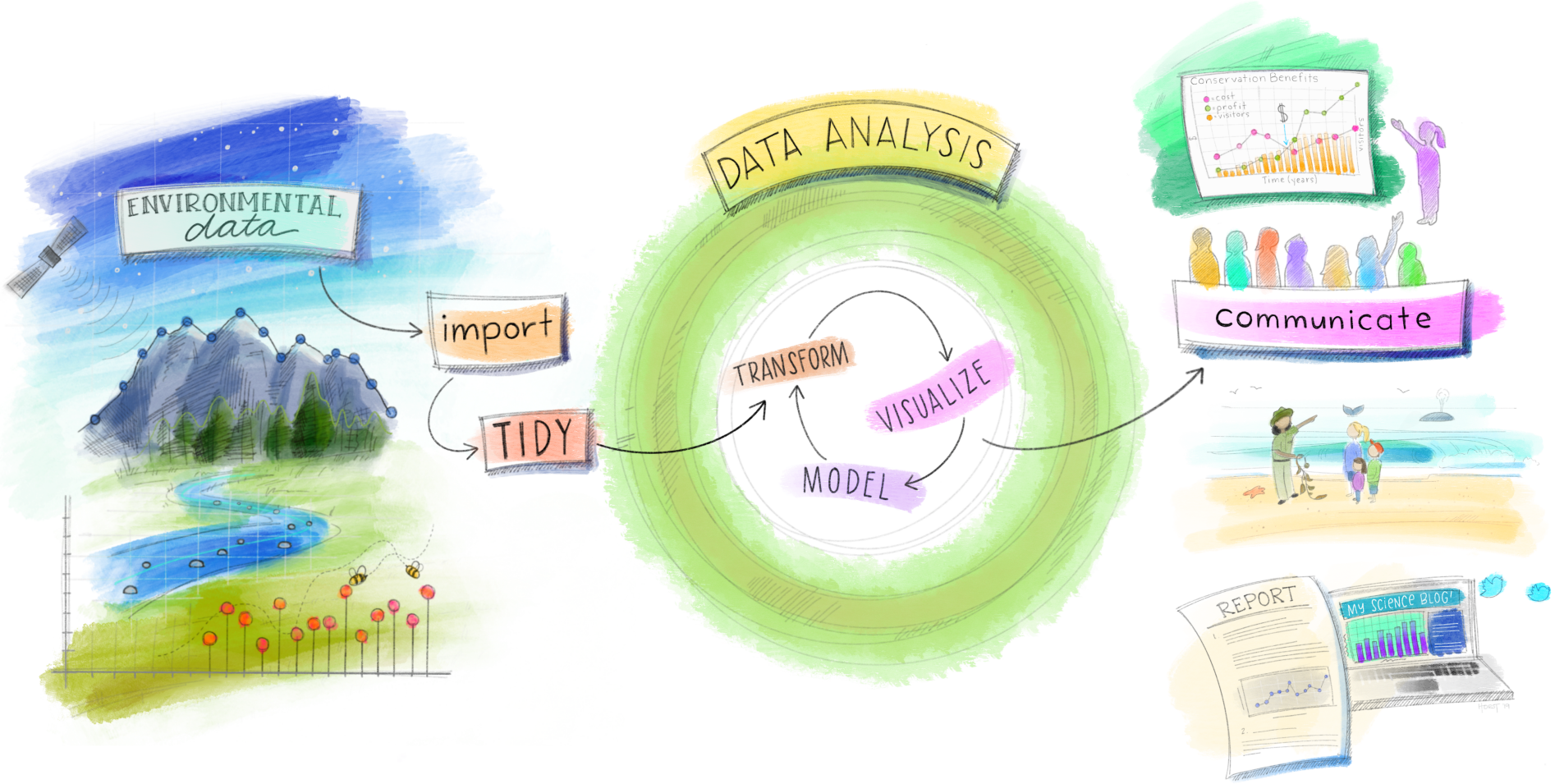
Reproducible Research

Marco Chiapello

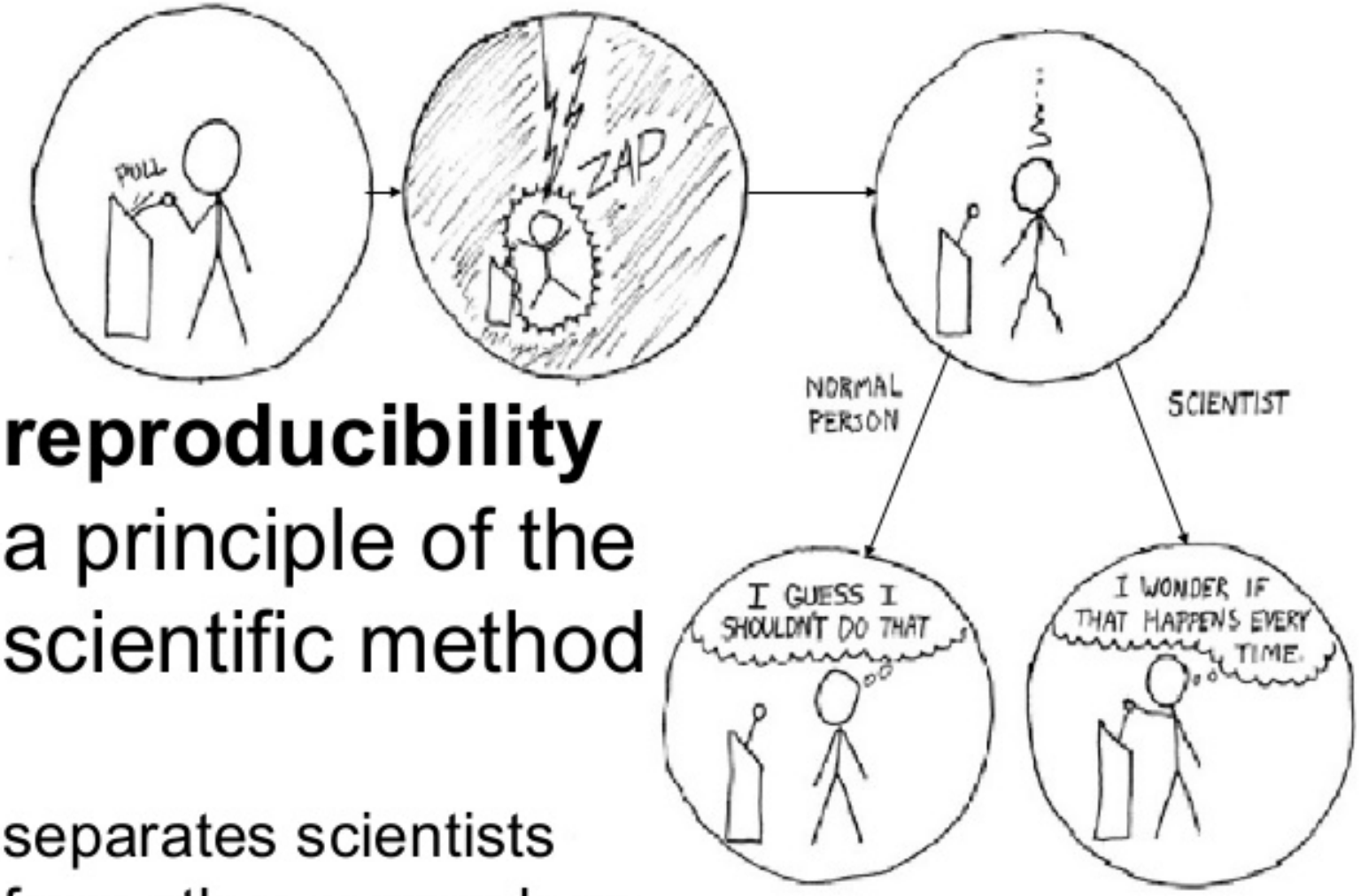
2020-12-15

Slack - marco.chiapello@unito.it

PhDToolbox Course



What reproducible research is

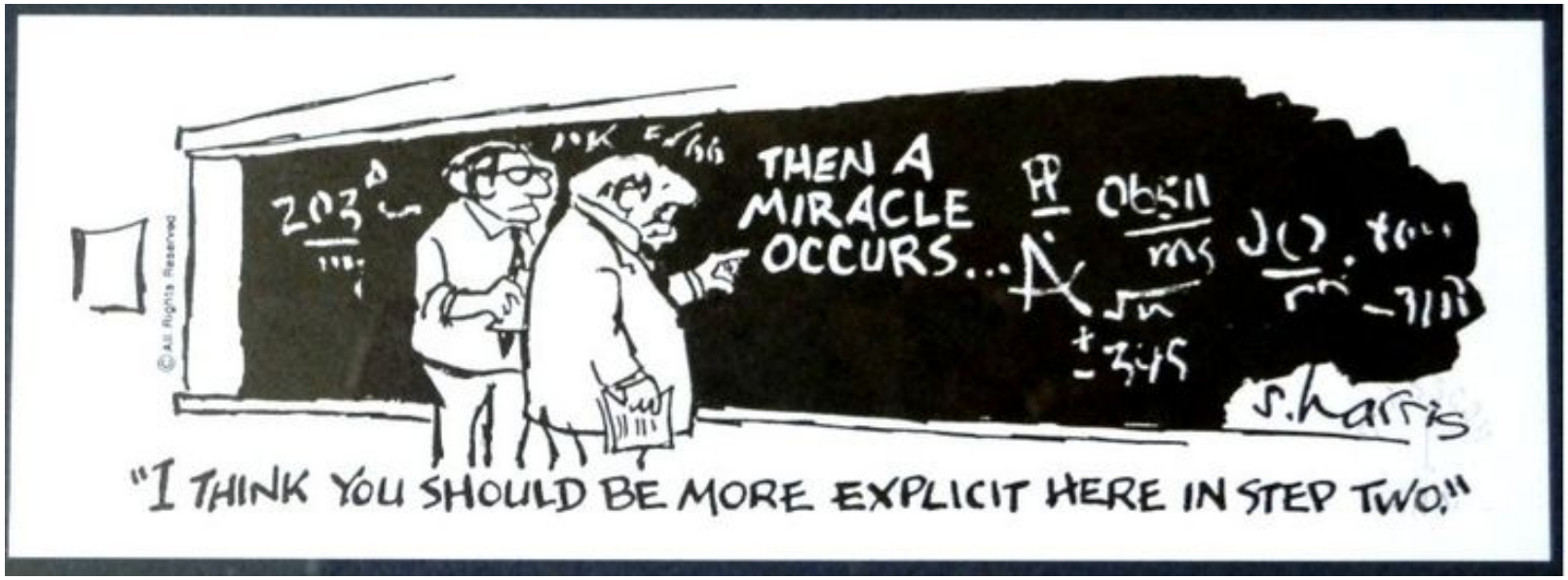


reproducibility
a principle of the scientific method

separates scientists from other researchers and normal people

http://xkcd.com/242/

What reproducible research is



This is exactly how it seems when you try to figure out how authors got from a large and complex data set to a dense paper with lots of busy figures. Without access to the data and the analysis code, a miracle occurred.

What reproducible research is

DATA + ANALYSIS → RESULTS

BUT not like this

What reproducible research is

MOTIVATED REASONING

PEOPLE TEND TO EVALUATE EVIDENCE IN WAYS CONSISTENT WITH THEIR PREFERENCES

WISHFUL THINKING

PEOPLE FORM BELIEFS BASED ON WHAT THEY'D LIKE RATHER THAN LOOKING AT DATA

CONFIRMATION BIAS

PEOPLE SEEK CONFIRMING EVIDENCE TO THEIR HYPOTHESIS & PUT MORE WEIGHT ON IT THAN DISCONFIRMING EVIDENCE

INFORMATION PURSUIT BIAS

PERSUING INFO LEADS TO US PUTTING MORE WEIGHT ON IT

SUNK COST FALLACY

THE HIGHER THE SUNK COSTS, THE MORE LIKELY PEOPLE STAY THE COURSE

What reproducible research is

DATA + ANALYSIS → RESULTS

Common practice of writing statistical reports:

- We import a dataset into Excel
- Run a procedure to get all results
- Copy and paste selected pieces into a typesetting program (usually Word)
- Add a few descriptions
- Finish a report

What reproducible research is

There are obvious dangers and disadvantages in this process:

1. It is **error-prone** due to too much manual work;
2. It requires lots of human effort to do **tedious jobs**;
3. The workflow is barely recordable, therefore it is **difficult to reproduce**;
4. A **tiny change** of the data source in the future will require the author(s) to go through the same procedure again;
5. The analysis and writing are separate, so close attention has to be paid to the **synchronization of the two parts**.

What is Reproducible Research?

The ability to reproduce someone else results

What do you need?

- Data

What reproducible research is

Reproducible vs Replicable

		DATA	
		Same	Different
CODE	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Ref: https://github.com/KirstieJane/ReproducibleResearch

What reproducible research is

Reproducibility/reproduce

A study is reproducible if there is a specific set of computational functions/analyses (usually specified in terms of code) that exactly reproduce all of the numbers in a published paper from raw data.

Replication/replicate

A study is only replicable if you perform the exact same experiment (at least) twice, collect data in the same way both times, perform the same data analysis, and arrive at the same conclusions.

Reproducible Research Rules

– based on (Sandve, Nekrutenko, Taylor, et al., 2013)

Rule 1

For Every Result, Keep Track of How It Was Produced

- The **full sequence** of pre- and post-processing steps are often critical in order to reach the achieved result
- **Every detail** that may influence the execution of the step **should be recorded**
- Include the name and **version of the program**, as well as the exact parameters and inputs

As a minimum, you should at least **record sufficient details** on **programs, parameters, and manual procedures** to allow yourself, in a year or so, to approximately reproduce the results

Rule 2

Avoid Manual Data Manipulation Steps

- Manual procedures are not only **inefficient and error-prone**, they are also **difficult to reproduce**
- Manual operations like the use of **copy and paste** between documents **should also be avoided**
- Manual modification of files can usually be replaced by the use of standard **UNIX commands** or scripts
- Manual tweaking of data files to attain format compatibility should be replaced by **format converters** that can be reenacted and **included into executable workflows**

If manual operations cannot be avoided, you should as a minimum

Rule 3

Archive the Exact Versions of All External Programs Used

- In order to exactly reproduce a given result, it may be necessary to use programs in the **exact versions used originally**
- It is **not always trivial to get hold of a program** in anything but the current version

As a minimum, you should note the **exact names** and **versions** of the main programs you use

Rule 4

Version Control All Custom Scripts

- Only that **exact state of the script** may be able to produce that **exact output**, even **given the same input data and parameters**
- The standard solution to **track evolution of code** is to use a version control system
  - A version control system is a **repository of files** with monitored access.
  - **Every change** made to the source **is tracked**, along with who made the change, why they made it

As a minimum, you should **archive copies of your scripts** from

Rule 5

Record All Intermediate Results, When Possible in Standardized Formats

- In principle, as long as the **full process** used to produce a given result **is tracked**, all **intermediate data can also be regenerated**
- In practice, having easily accessible intermediate results may be of great value (eg, to **spot errors**)
- When the full process is not readily executable, it allows parts of the process to be rerun
- **It allows critical examination** of the full process behind a result

Rule 6

For Analyses That Include Randomness, Note Underlying Random Seeds

- Many analyses and predictions include some **element of randomness**, meaning the same program will typically give **slightly different results every time it is executed**
- **Given the same initial seed**, all random numbers used in an analysis will be equal, thus giving identical results every time it is run

As a minimum, you should **note which analysis steps involve**

Rule 7

Always Store Raw Data

Always store in a safe place the raw data

Never touch or mofidy the raw data

Rule 8

Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected (test units)

- The **final results** that make it to an article, be it plots or tables, often represent **highly summarized data**
- In order to validate and fully understand the main result, it is often useful to **inspect the detailed values** underlying the summaries

When working with summarized results, you should as a minimum at least once

Rule 9

Connect Textual Statements to Underlying Results

- The results of analyses and their corresponding textual interpretations are clearly interconnected but **often lie in different places**
- Results usually live on a personal computer, while interpretations live in text documents
- To allow efficient retrieval of details behind textual statements, we suggest that **statements are connected to underlying results already from the time the statements are initially formulated**

Integrate reproducible analyses directly into textual documents

RMarkdown

Rule 10

Provide Public Access to Scripts, Runs, and Results

- All input data, scripts, versions, parameters, and inter-mediate results should be made **publicly and easily accessible**
- Making reproducibility of your work by peers a realistic possibility sends a **strong signal of quality, trustworthiness, and transparency**

Reproducible research Tools

Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to humans what we want the computer to do.

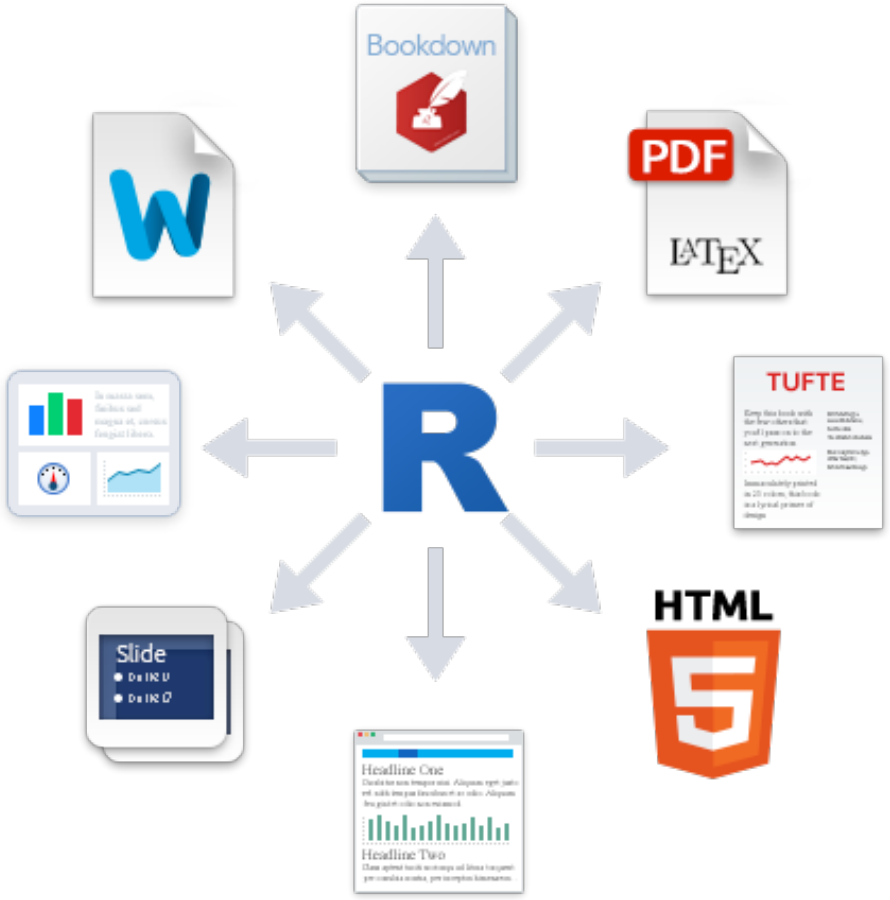
Tools

Literate programming is a methodology that combines a programming language with a documentation language

- Write program code
- Write narratives to explain what is being done by the program

Tools

RMarkdown



DEMO Rmarkdown

Tools

Project organization

I strongly advise to split your project in folder and not dumpo everything on the desktop.

1. **README file**: explain the purpose of the project and describe the folder/files in it
2. **RAWDATA folder**: contains the rawdata. We advise to have it with "read only permission"
3. **SCRIPT folder**: contains the script used for your analysis
4. **ANALYSIS folder**: contains the results of your analysis

Tools

Version Control

Version control software keeps track of every modification to the code in a special kind of database. If a mistake is made, developers can **turn back the clock** and compare earlier versions of the code to help fix the mistake while minimizing disruption to all team members.

Conclusion

Conclusion

- Learning to use these tools will **require commitment and a massive investment of your time and energy**
- **A priori** it is not clear why the benefits of working reproducibly outweigh its costs.
- Does reproducibility sound like **extra work**?
- It can be, particularly when one is first trying to do it, that is, to **break one's own previous nonreproducible habits**

Conclusion

My advice is:

Learn the tools of reproducibility as quickly as possible

Use them in every project.

Questions