

Analysis Reproducibility

Why it matters and how to do it?

8 July 2021

Learning objectives

About today:

- Understand what you can gain from analysis reproducibility.
- Know what the main technical requirements are to set up for their analysis to be reproducible.
- Have a demonstration of a practical way to make a cake using household survey data: crunching, analysis & interpretation & data stories!

Not today:

- Induction Training on R language! For this head to [UNHCR Learn & Connect- R training](#)

A Vision for data analysis

"Multi-functional teams, with strengthened data literacy, regularly conduct meaningful and documented joint data interpretation sessions to define their strategic directions based on statistical evidences"

A Theory of Change for Data analysis

Proper user of data for advocacy & programmatic decision making

- ↪ Corporate **Standards** exist to define how to encode & process household surveys dataset
- ↪ Field data experts are trained based on precise recipes and predefined tools at each step of the **data life cycle**
- ↪ Data are presented, discussed and linked to expert knowledge during data **interpretation** sessions with a multi-functional team
- ↪ All potential valid interpretations, including diverging views, are systematically **recorded**
- ↪ **Persuasive** "Data Stories" and Policy papers are generated

Data Science is like cooking

When a chef is starting out with a new dish...

- Hypothesis Tasting -- Setting the right questions
- Ingredients = source the Data
- Wash your food = clean your data
- Flavor engineering = create calculated & derived variables
- Taste and explore = reshape & visualize the data
- Tune your oven = statistical modeling
- Art of plating = use styled brand
- Document your recipe = add technical comments

Data



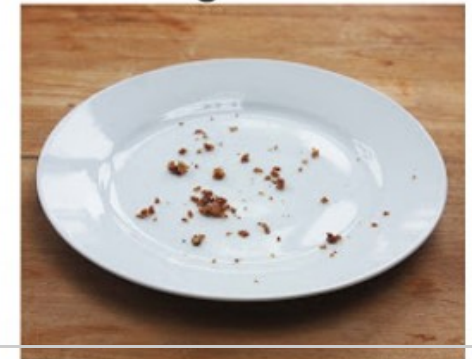
Information



Presentation



Knowledge



Information Anxiety & Analysis paralysis

When people do not want to eat the cake...

Potential source of reluctance...

- I do not know how to eat it: I see all those elements on it without being able to understand why they were added there and how this works...
- I do not trust this cake: How was it created? Did you follow correctly the recipe? Were the ingredient fresh? Can I trust how you sourced the ingredient?
- This is not the cake I need! It looks too heavy & too big: I will not be able to digest it...
- I am not hungry and do not even know what cake I want...



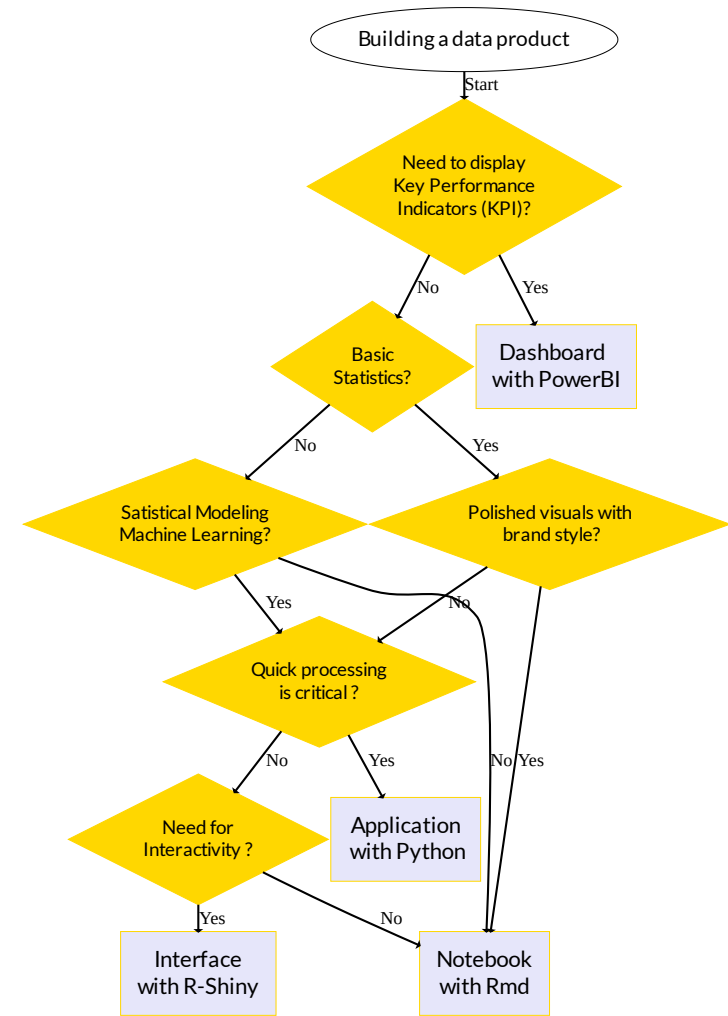
Data Products: When What?

Dashboard are relevant for displaying KPIs! (*like when you drive your car...*)

Key Performance Indicators (KPIs) are indicators specifically designed to show progress toward an intended result, i.e a predefined **target**

Create an analytical basis for **decision making**, aka Business Intelligence

Help focus attention of Snr Management on what matters most - a good dashbaord needs to be **concise**



Why we need to work in a reproducible way?

Ethics, Productivity, Learning

Ethics: Science is '*show me*' - not '*trust me*'

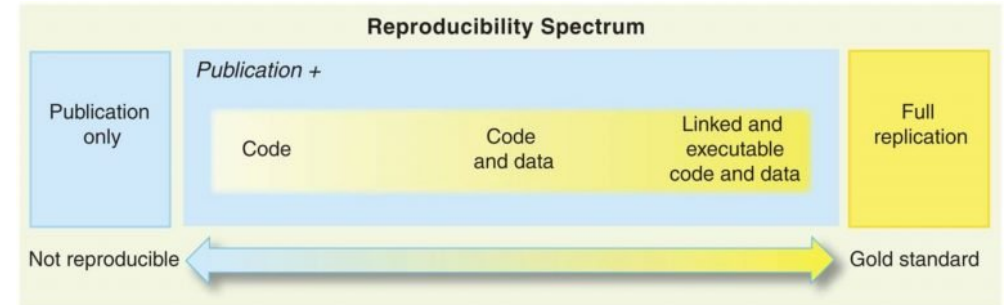
Reproducibility allows for **peer review**

Peer Review allows for **transparency**

Transparency allows for **scrutiny**

Scrutiny allows for **accountability**

It's okay to make mistakes, as long as one can detect them and that we can learn from them...

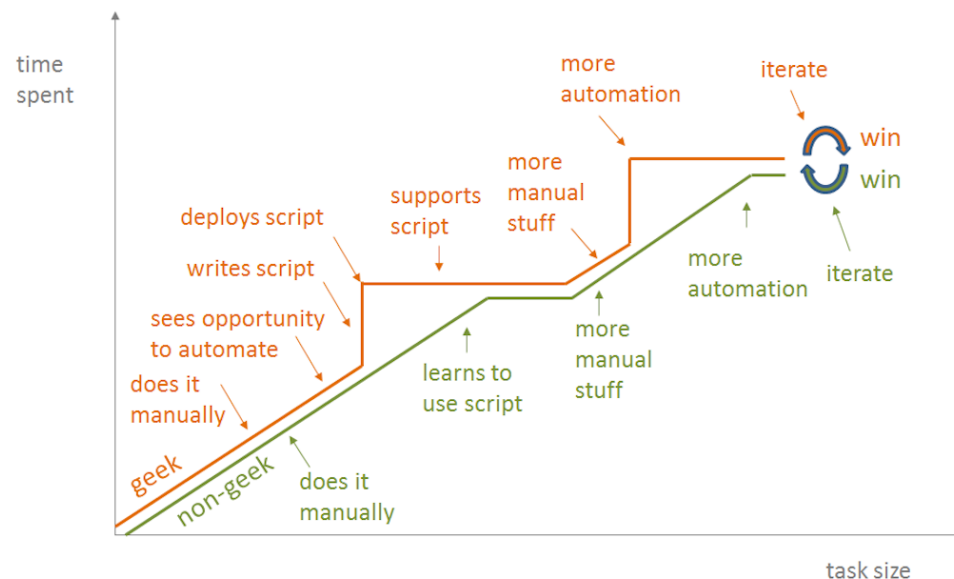


Productivity: getting things done quickly and safely!

Automation through functions & scripts can help skipping **repetitive tasks**

Tasks that involve recurrent **data manipulation** are undertaken by teams.... but not everyone in the team needs to be a **geek/coder!**

When enough investment can be made, **Graphical User Interface (GUI)** can be developed for specific functions to ease the learning curve of new users while they are still in the process of building up their personal R skills.



An R-Community geared towards learning

Which approach is the most appealing exercise among the 2 proposed aside?!!

Start from an end-product
and **reverse engineer** it!

Eat the cake first! (then play with and change ingredients...)

(a)

- Declare the following variables
- Then, determine the class of each variable

```
# Declare variables
x ← 8
y ← "monkey"
z ← FALSE

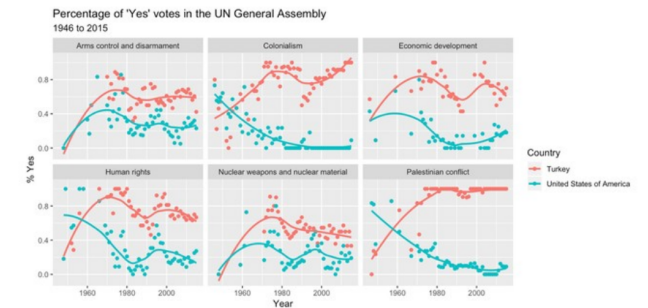
# Check class of x
class(x)
#> [1] "numeric"

# Check class of y
class(y)
#> [1] "character"

# Check class of z
class(z)
#> [1] "logical"
```

(b)

- Open today's demo project
- Knit the document and discuss the results with your neighbor
- Then, change **Turkey** to a different country, and plot again



Conditions for reproducibility.

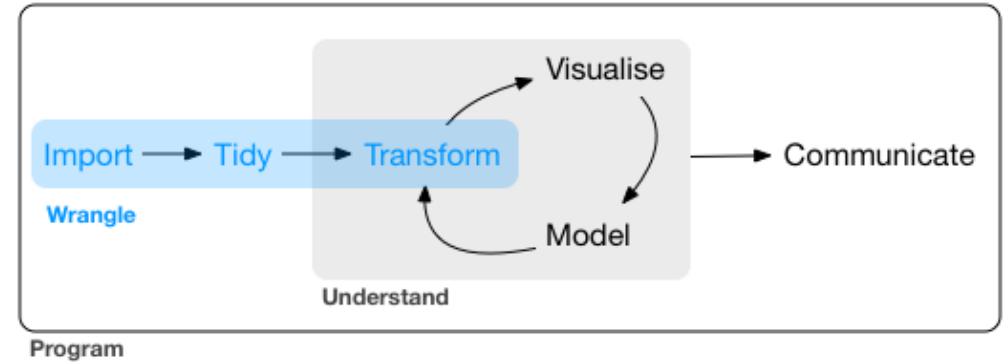
Sourcing data, documenting analysis, & packaging output

Preparing data

Data Wrangling takes usually more than 80% of any data project time...

Imagine if you need to rewind your analysis...

Correct at any steps in the process and re-run all..



Documenting analysis

DO

Cutting corners to meet arbitrary management deadlines



Essential

Copying and Pasting
from Stack Overflow

How to actually learn any new programming concept



Essential

Changing Stuff and
Seeing What Happens

The internet will make those bad words go away



Essential

Googling the
Error Message

Software can be chaotic, but we make it work



Expert

Trying Stuff
Until it Works

Where's the fun in just knowing what the code is supposed to do?



Essential

Excuses for Not
Writing Documentation

Does it run? Just leave it alone.



Writing Code that
Nobody Else Can Read

git commit -m "changes"



Writing

Useless Git
Commit Messages

Who are you kidding?



"Temporary"
Workarounds

DON'T

Packaging functions

Gradual automation

- level 1: write a command
- level 2: organize multiple command together in reusable function
- Level 3: organize multiple functions together in a package
- Level 4: includes test data & Documentation
- Level 5: **Unit testing, aka code review**
- Level 6: **Graphical User Interface (GUI)**

Package

- DESCRIPTION
- R/
- tests/
- man/
- vignettes/
- data/
- NAMESPACE

SETUP

WRITE CODE

TEST

DOCUMENT

TEACH

ADD DATA

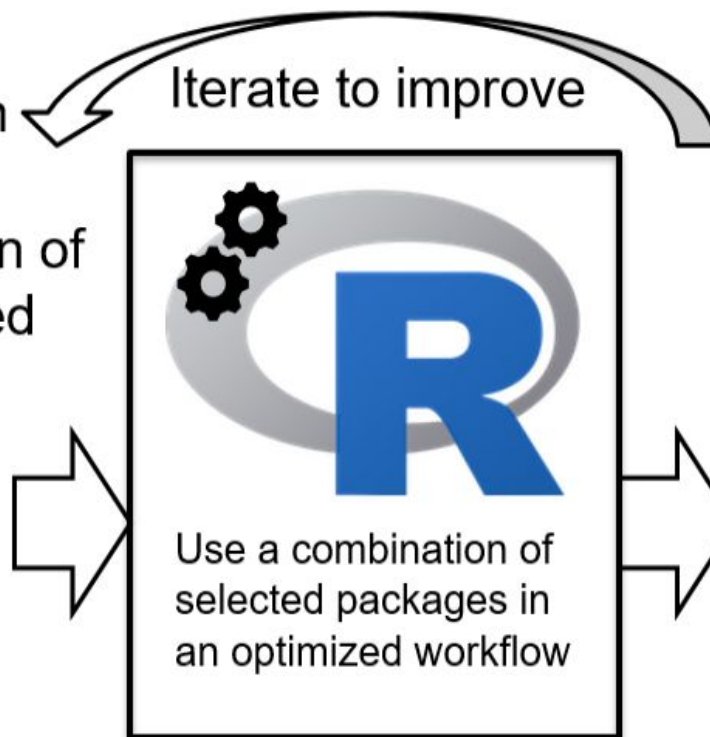
ORGANIZE

Hands-on practice: a practical run-through based on Household survey dataset

Crunching, Interpretation & Dissemination

Step 1- Notebook for Automatic Data exploration, aka "crunching"

Configure the analysis plan in Excel within an extended version of the XlsForm used for the survey



Generate an **Rmd Notebook** and get standard report generated in multiple formats



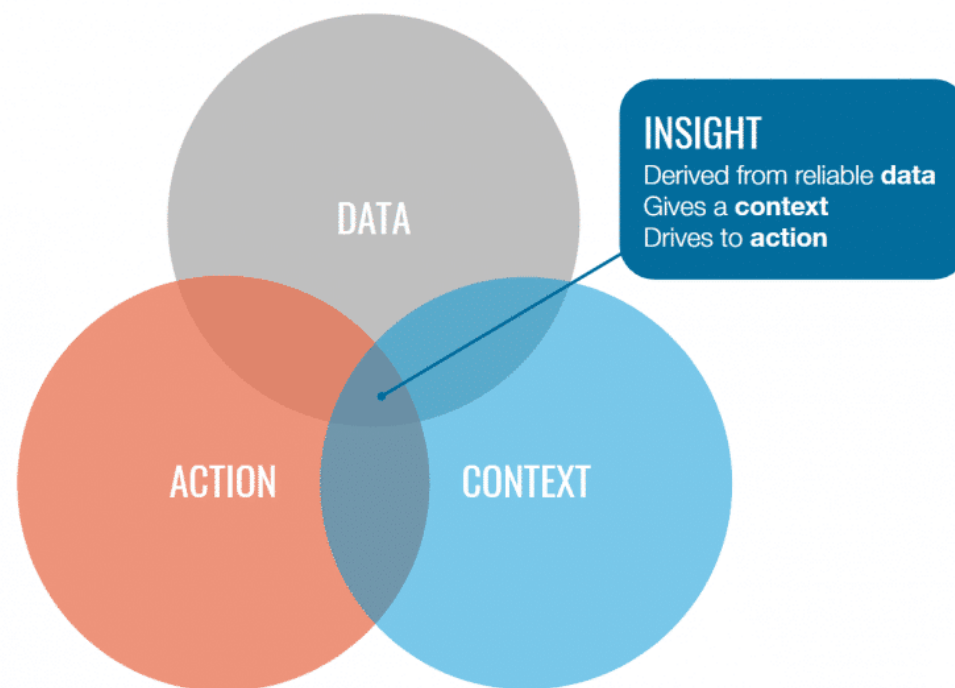
Step 2- Notebook for Data Insights documentation: Analysis Repo

Insight: The capacity to gain an accurate and deep understanding of someone or something

Not all charts will emulate need for interpretation - the data analyst need to generate the one that can create **debates**.

Charts need to be **crafted** - for instance use chart title framed as "opening question"...

Insights arise when a multifunctional team is able to explain **unexpected patterns**, to challenge or revise **existing assumptions**, or to identify evidence to support **Call to action**.



Step 3- Notebook to communicate with data: Microsite

From **assumptions** to **evidence based** statement

Data is to support Narrative - not the other way around!

Leverage Art Data Storytelling to:

- Explain,
- Enlighten,
- Engage



Conclusion

R in Humanitarian Context

You are not alone

More than 450 users from multiple organisation in the [humanitarian-useR-group](#)

Around already ≈20 R champions within UNHCR vs more than 420 PowerBI Pro users

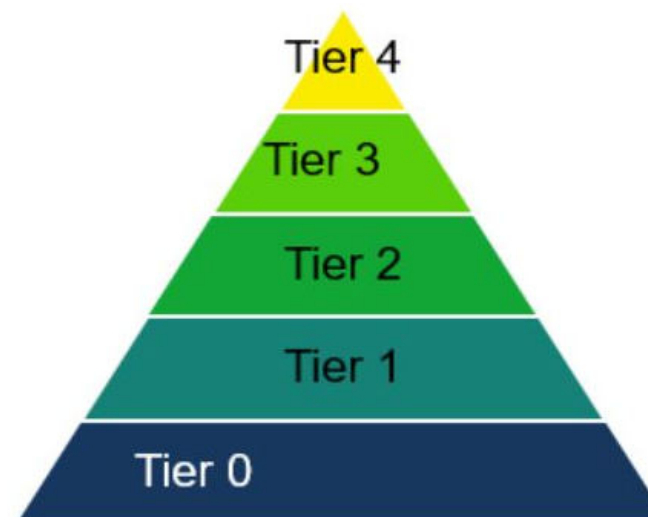
Try to start by using existing UNHCR packages and start from a project you can reproduce



A call for Institutionalisation

Using Standard Multi Tier IT Standard Support model to enhance reproducible analysis...

- Tier 4: Code Review & Quality Insurance / Contracted Company with global frame agreement
- Tier 3: Internal package development / Internal R champions team (cost: one yearly Rdev meeting to incentivize contributing staff)
- Tier 2: User induction & Advanced User Support / Global Data Service/DIMA (Data Science Team)
- Tier 1: Basic User Troubleshooting / Global Service Desk (WIPRO according to Documented Scenario)
- Tier 0: Self-support / Package documentation (maintained and improved on continuous basis)



Your Opinion Count

Please [fill this survey](#) to share your opinion and thoughts on the topic presented here

Slides are available in [English](#) & [French](#) - slides note can be displayed by pressing keyboard shortcut p - and in PDF ([English](#) & [French](#)) within this [github repo](#)