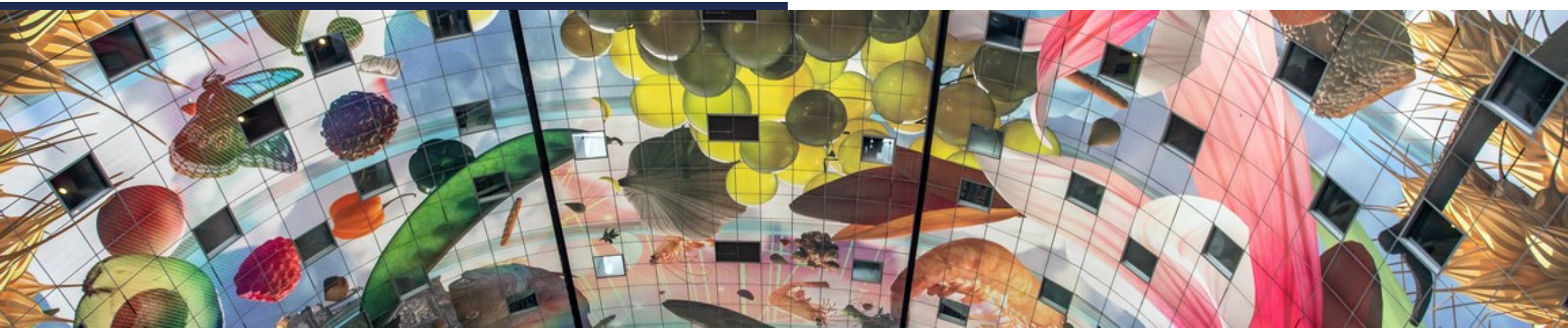


# Open Science Workflow: Open Data, Code, Materials

Ana Martinovici

Aurélie Lemmens



# Why Adopting an Open Science Workflow?



Information Asymmetry = Lack of Trust

# Why Adopting an Open Science Workflow?

*Author*

=

*Seller*



*Reviewer*

=

*Buyer*

# Transparency = Remedy

# Transparency Opens New Ways to Review

## Closed Science

Information  
Asymmetry

Publication Count  
Significance levels

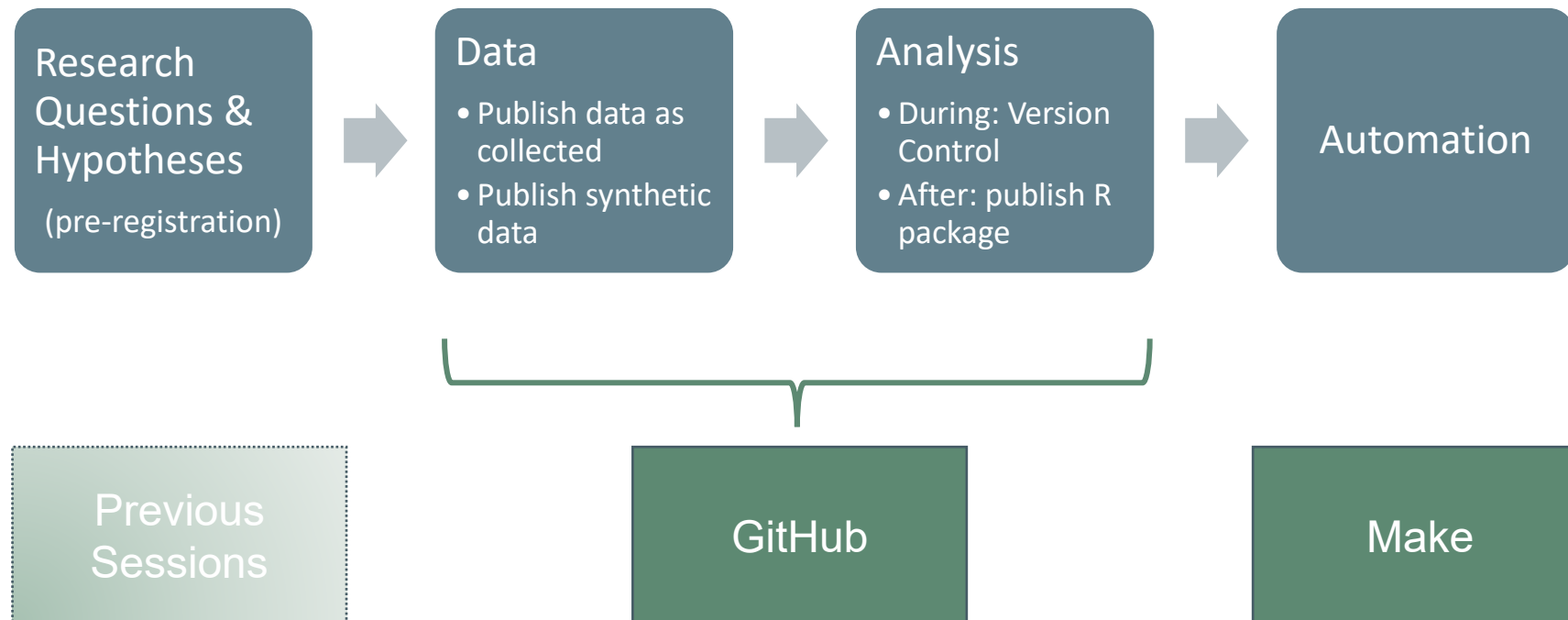


## Open Science

Information  
symmetry

Strength of scientific  
evidence

# Our Agenda



# Data: Publish Synthetic Data

- (How) Can I Share My Data if
  - I work under a NDA?
  - I want to make sure I can use them again in a next paper?
  - They belong to co-authors?



Yes, you can generate new data sets that – on average – have similar properties as the original data set but that are different from it:

$$\frac{\sum_i^m \hat{\beta}_i}{m} = \hat{\beta}$$

$\hat{\beta}$  the construct of interest estimated on the original data set

$\hat{\beta}_i$  the estimate on synthetic data set  $i = 1, \dots, m$

# Synthesize Data for Sharing



*Journal of Statistical Software*

October 2016, Volume 74, Issue 11.

doi: 10.18637/jss.v074.i11

## synthpop: Bespoke Creation of Synthetic Data in R

Beata Nowok  
University of Edinburgh

Gillian M. Raab  
University of Edinburgh

Chris Dibben  
University of Edinburgh

### Abstract

In many contexts, confidentiality constraints severely restrict access to unique and valuable microdata. Synthetic data which mimic the original observed data and preserve the relationships between variables but do not contain any disclosive records are one possible solution to this problem. The **synthpop** package for R, introduced in this paper, provides routines to generate synthetic versions of original data sets. We describe the methodology and its consequences for the data characteristics. We illustrate the package features using a survey data example.

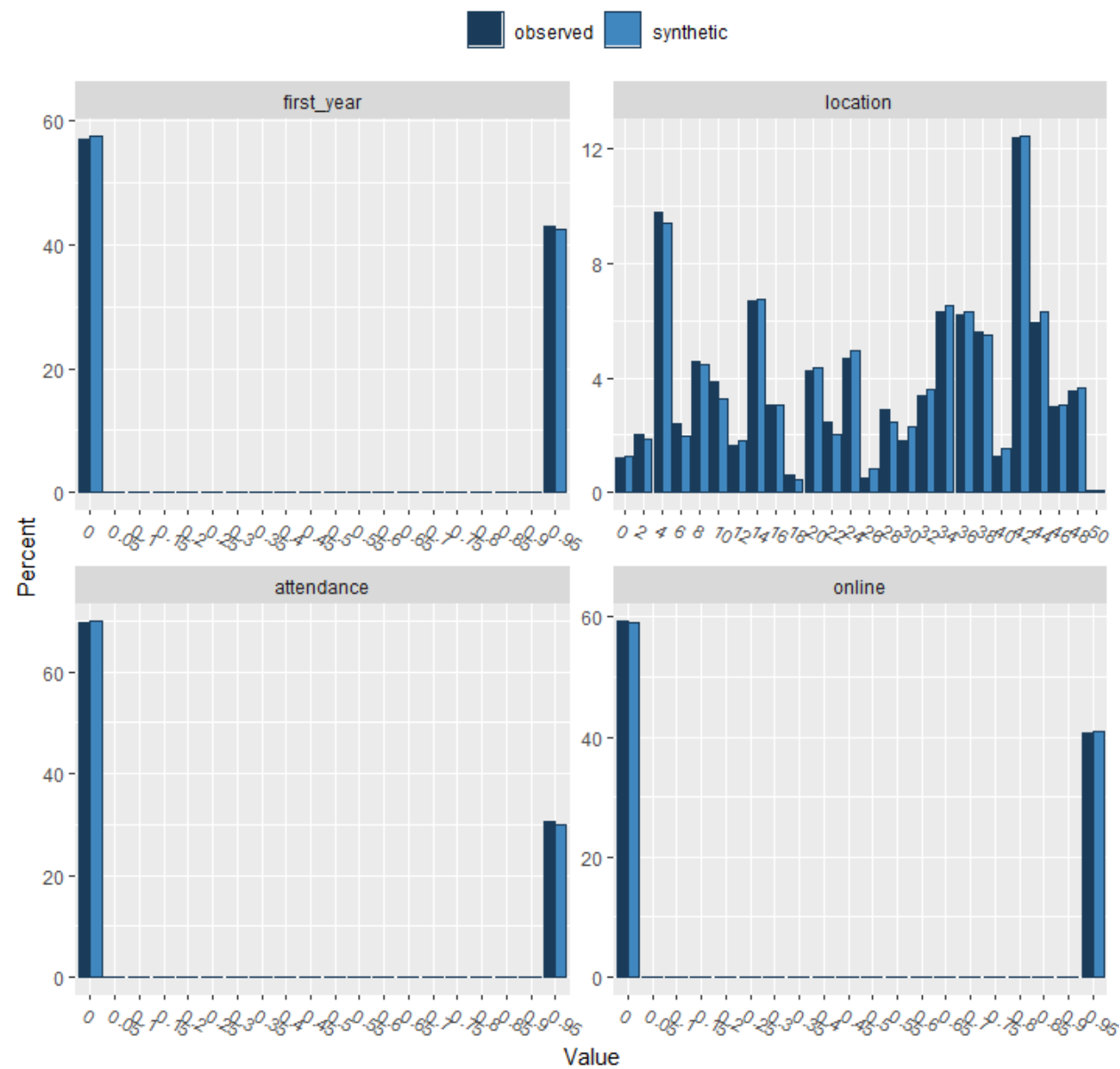
*Keywords:* synthetic data, disclosure control, CART, R, UK longitudinal studies.

- Assumption: observed data  $y_{obs}$  come from a population that our synthesizer can estimate,  $f(y_{obs}|x_{obs}, \theta)$
  - Different models can be used (tree, ensemble, regression, etc.).
  - Joint distribution of the data = series of conditional distributions:
    - *Synthpop* generates one column at a time conditional on all others (including the new ones)
- ➔  $m$  data sets are generated. When estimating a quantity of interest for these  $m$  data sets,  $\hat{\beta}_1, \dots, \hat{\beta}_m$  will have mean equal to the  $\hat{\beta}$  estimated on the original data set

# Example

In R:

```
syn(mydata, seed = my.seed)  
compare(syn(mydata), mydata)
```





# Analysis: Create a R Package

```
#### Step 1: Install and Load Packages
install.packages("devtools")
install.packages("roxygen2")
library("devtools")
library(roxygen2)

#### Step 2: Create Directory
setwd("C://Users/ndbae/Dropbox (Personal)/Open_Science/tutorial")
create("myfirstpackage")

#### Step 3: Create a function, document it properly, and save as "my_first_function.r" file in the R folder

#' My First Open Science Function
#'
#' This function allows you to practice with Open Science
#' @param open_science Do you want to practice Open Science? Default is TRUE.
#' @keywords
#' @export
#' @examples
#' my_first_function()

my_first_function <- function(open_science){
  if(open_science == TRUE){
    print("it worked :)")
  }
  else {
    print("great job!")
  }
}

#### Step 4: Attach the documentation
setwd("myfirstpackage")
devtools::document()

#### Step 5: when ready, you can install and load your new package
setwd("..")
install("myfirstpackage")
library(myfirstpackage)
```

```
> myfirstpackage::my_first_function(TRUE)
[1] "it worked :)"
```



Files Plots Packages Help Viewer

R: My First Open Science Function Find in Topic

my\_first\_function {myfirstpackage} R Documentation

## My First Open Science Function

### Description

This function allows you to practice with Open Science

### Usage

```
my_first_function(open_science)
```

### Arguments

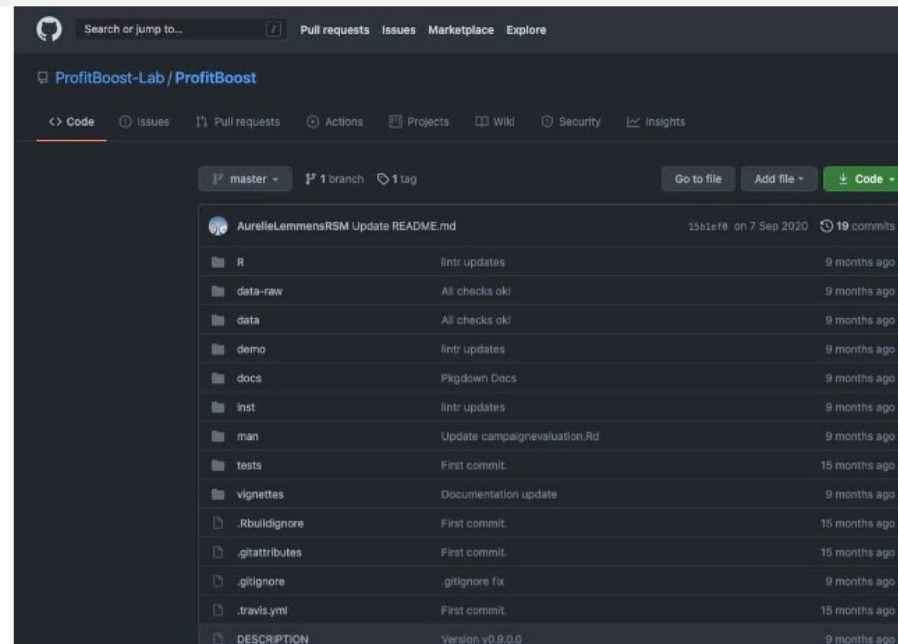
open\_science Do you want to practice Open Science? Default is TRUE.

### Examples

```
my_first_function()
```

[Package myfirstpackage version 0.0.0.9000 [Index](#)]

# Analysis: Create a R Package



## ProfitBoost R Package

The R package ProfitBoost contains the algorithm developed in Lemmens and Gupta (2020), Managing Churn to Maximize Profits, Marketing Science. The package is available free of charge and can be downloaded on ...

[Read More](#)



Let's Make GitHub Fun!

A brief intro to  
Git, GitHub, and Make

**WHAT DO YOU MEAN**

**THERE'S NO VERSION CONTROL?**





**neil\_ferguson** ✓  
@neil\_ferguson

I'm conscious that lots of people would like to see and run the pandemic simulation code we are using to model control measures against COVID-19. To explain the background - I wrote the code (thousands of lines of undocumented C) 13+ years ago to model flu pandemics...

10:13 PM · Mar 22, 2020 · [Twitter for iPhone](#)

**1.5K** Retweets **5K** Likes

<https://en.wikipedia.org/wiki/CovidSim>



**neil\_ferguson** ✓ @neil\_ferguson · Mar 22

Replying to @neil\_ferguson

I am happy to say that [@Microsoft](#) and [@GitHub](#) are working with [@Imperial\\_JIDEA](#) and [@MRC\\_Outbreak](#) to document, refactor and extend the code to allow others to use without the multiple days training it would currently require (and which we don't have time to give)...

48

246

1.3K



**neil\_ferguson** ✓ @neil\_ferguson · Mar 22

They are also working with us to develop a web-based front end to allow public health policy makers from around the world to make use of the model in planning. We hope to make v1 releases of both the source and front end in the next 7-10 days...

50

139

1.1K



**neil\_ferguson** ✓ @neil\_ferguson · Mar 22

That timescale reflects the balancing of those priorities with the multitude of other urgent policy-relevant COVID-19 questions we are addressing....

15

52

501



**neil\_ferguson** ✓ @neil\_ferguson · Mar 22

As well as the partners listed above, I would also like to thank all the other individuals and orgs who have offered to help. It is appreciated, but we only have limited bandwidth to manage such partnerships right now. We hope to bring more partners on board over time.

36

52

625



WHAT

# What is git?

- Formal version control system
- Developed by Linus Torvalds
  - Used to manage the source code for Linux
- Tracks content:
  - Source code
  - Data analysis projects
  - Websites
  - Presentations
  - Manuscripts

# What is GitHub?

- A home for git repositories
- Interface for exploring public git repositories
- A 'safe place' to keep code and data
- Additional benefits: issues, projects, wiki, insights



**WHY**

# Why use GitHub?


- Facilitates
  - Exploring code
  - Tracking issues
  - Learning from others
- Lowers the barriers to collaboration
  - Email “there’s a typo in your code in file X, line 30” vs
  - Pull request “here’s a correction to your code”
- Free for researchers and students


**I SHOULD USE GITHUB**



HOW

<https://osf.io/zmu3k/>

 OSFHOME ▼

 Workshop—GitFun: Introduction to git ... Files Wiki Analytics Registrations


Society for the Improvement of Psychological Science  
(SIPS) 2021 Meeting /

Workshop—GitFun: Introduction to git and  
GitHub


Contributors: [Ana Martinovici](#)

Date created: 2021-05-03 03:16 AM | Last Updated: 2021-07-07 12:34 PM

Identifier: DOI 10.17605/OSF.IO/ZMU3K

Category:  Software

Description: Slides and recording of the June 25, 2021 workshop.

License: Other 

<https://www.eur.nl/werken-bij/trainingen-eur/share-your-knowledge>

Donderdag 7 april

9.30 - 11.00

11.15 - 12.45

GitFun: Introduction to git and GitHub - Ana Martinaovici

