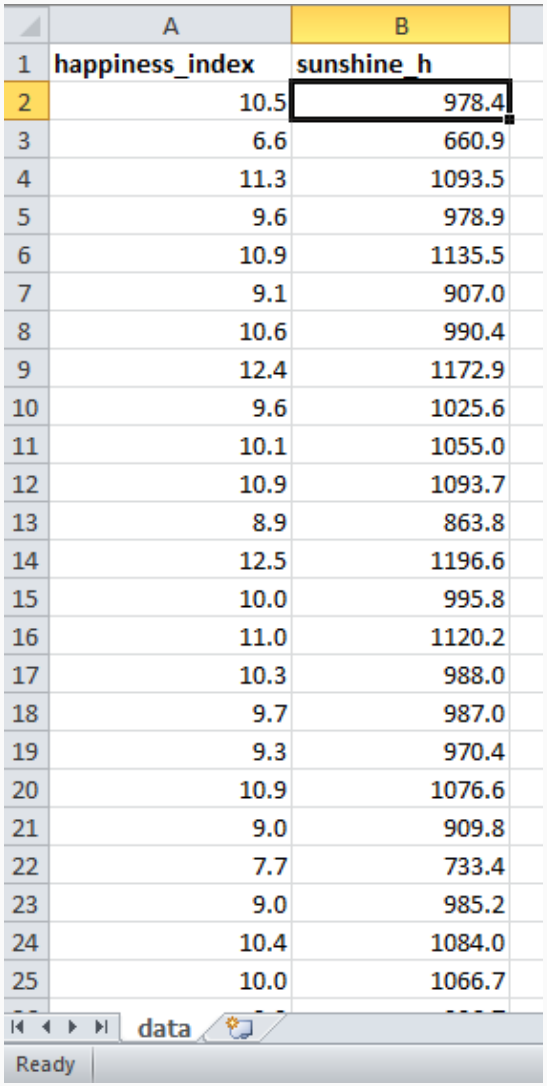
Doing reproducible science: from your hard-won data to a publishable manuscript without going mad

Francisco Rodriguez-Sanchez

@frod_san

A typical research workflow

1. Prepare data (**EXCEL**)



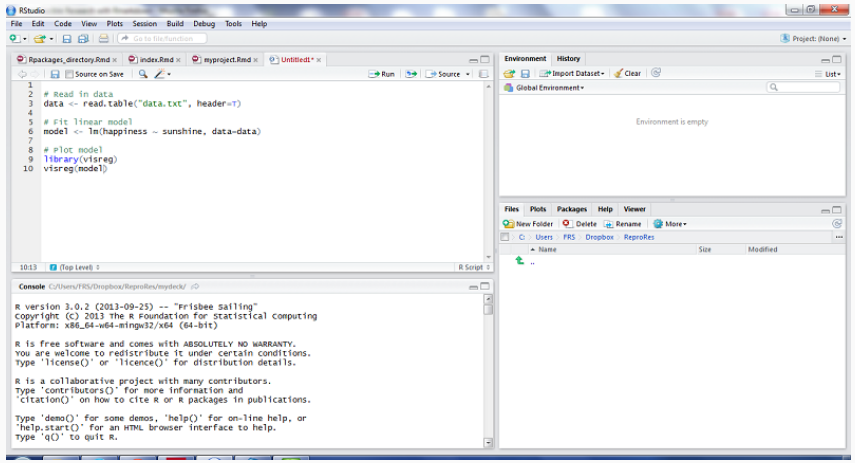
	A	B
1	happiness_index	sunshine_h
2	10.5	978.4
3	6.6	660.9
4	11.3	1093.5
5	9.6	978.9
6	10.9	1135.5
7	9.1	907.0
8	10.6	990.4
9	12.4	1172.9
10	9.6	1025.6
11	10.1	1055.0
12	10.9	1093.7
13	8.9	863.8
14	12.5	1196.6
15	10.0	995.8
16	11.0	1120.2
17	10.3	988.0
18	9.7	987.0
19	9.3	970.4
20	10.9	1076.6
21	9.0	909.8
22	7.7	733.4
23	9.0	985.2
24	10.4	1084.0
25	10.0	1066.7

data

Ready

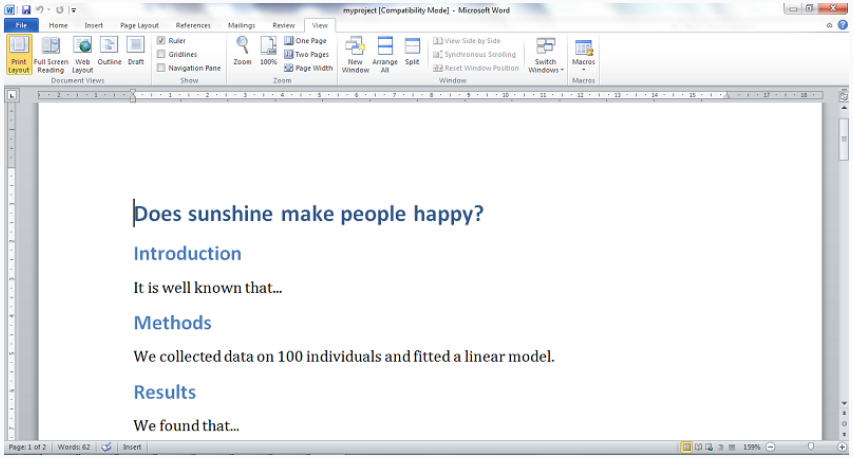
A typical research workflow

1. Prepare data (**EXCEL**)
2. Analyse data (**R**)



A typical research workflow

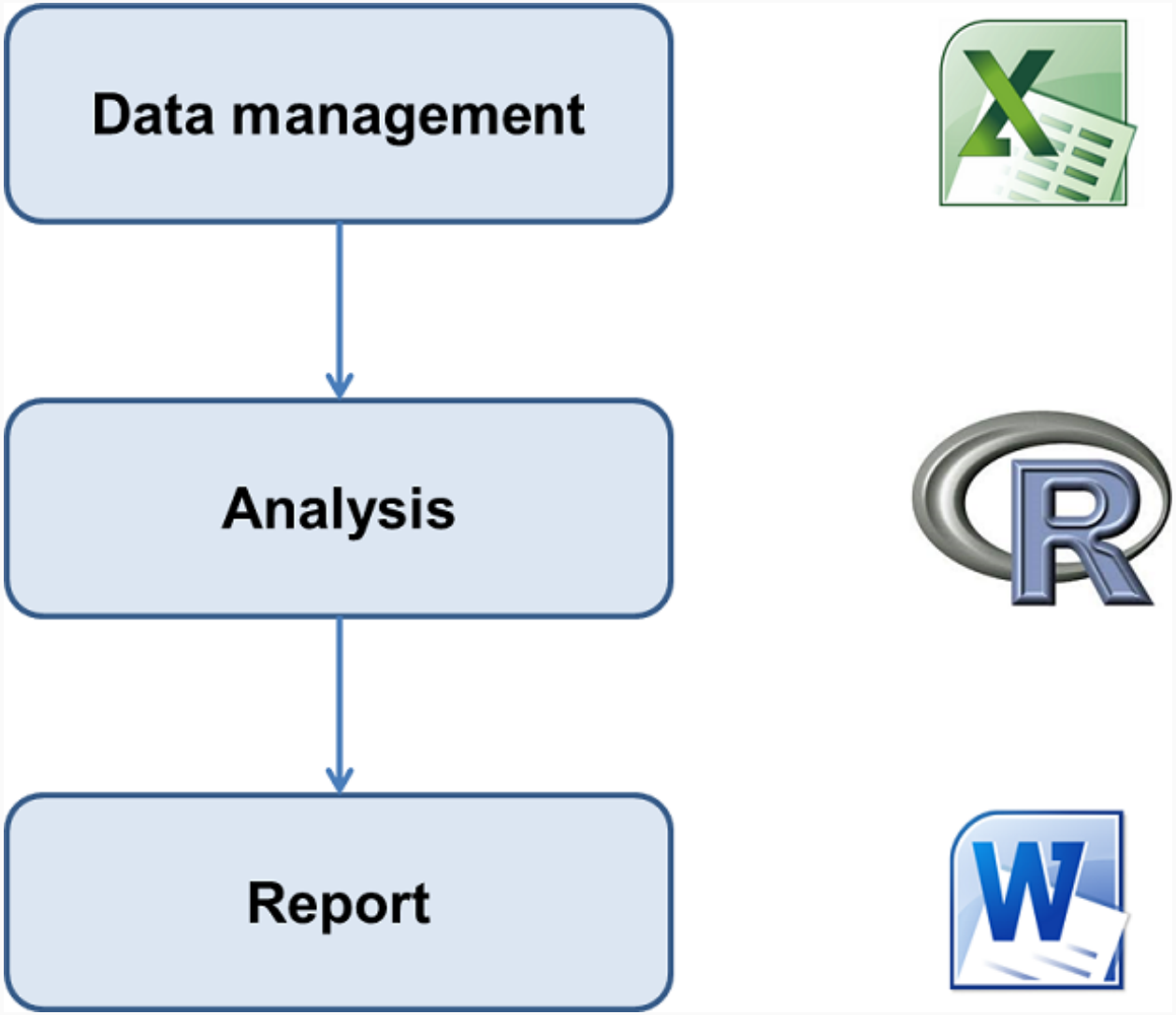
1. Prepare data (**EXCEL**)
2. Analyse data (**R**)
3. Write report/paper (**WORD**)



A typical research workflow

1. Prepare data (**EXCEL**)
2. Analyse data (**R**)
3. Write report/paper (**WORD**)
4. Start the email attachments **nightmare**...

This workflow is broken



Problems of a broken workflow

- How did you do this? What analysis is behind this figure? Did you account for . . . ?

Problems of a broken workflow

- How did you do this? What analysis is behind this figure? Did you account for . . . ?
- What dataset was used? Which individuals were left out? Where is the clean dataset?

Problems of a broken workflow

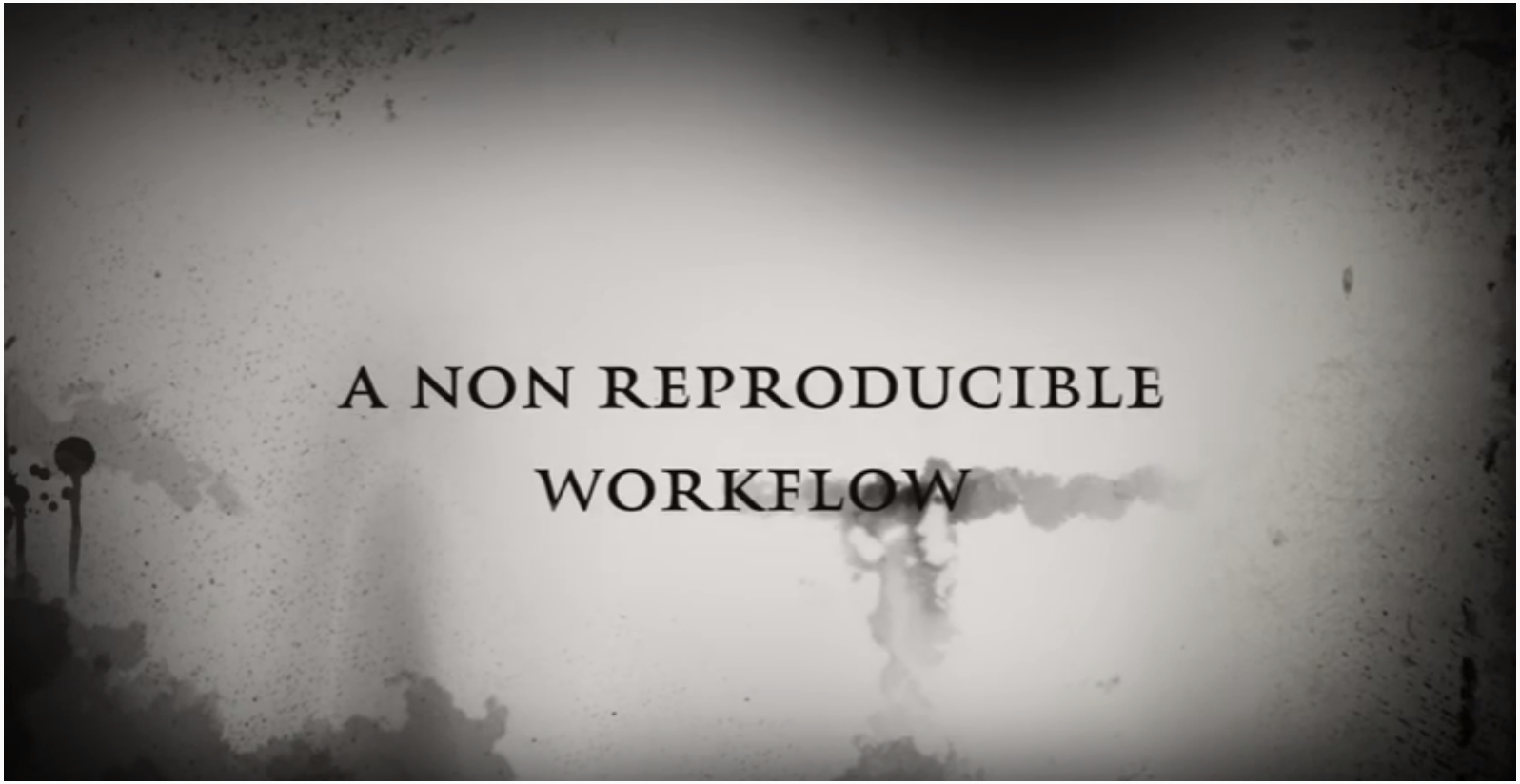
- How did you do this? What analysis is behind this figure? Did you account for . . . ?
- What dataset was used? Which individuals were left out? Where is the clean dataset?
- Oops, there is an error in the data. Can you repeat the analysis? And update figures/tables in Word!

Trevor A. Branch
@TrevorABranch

Follow

My rule of thumb: every analysis you do on a dataset will have to be redone 10–15 times before publication. Plan accordingly. #Rstats

Our everyday scary movie



https://youtu.be/s3JldKoA0zw

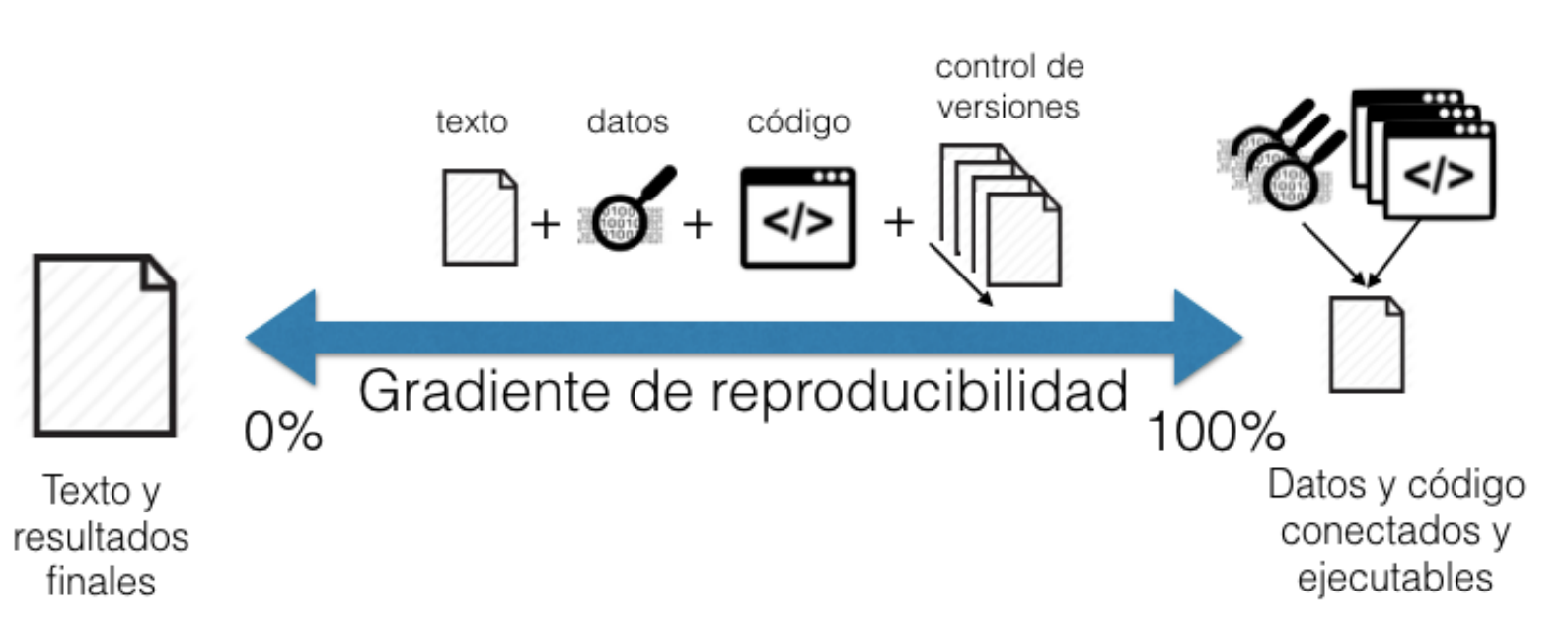
WHAT is Reproducible Science?

Reproducible Science: WHAT

A scientific article is **reproducible** if there is computer **code** that can **regenerate** all results and figures from the original data.

- Transparent
- Traceable
- Comprehensive
- Useful

Most science is not reproducible

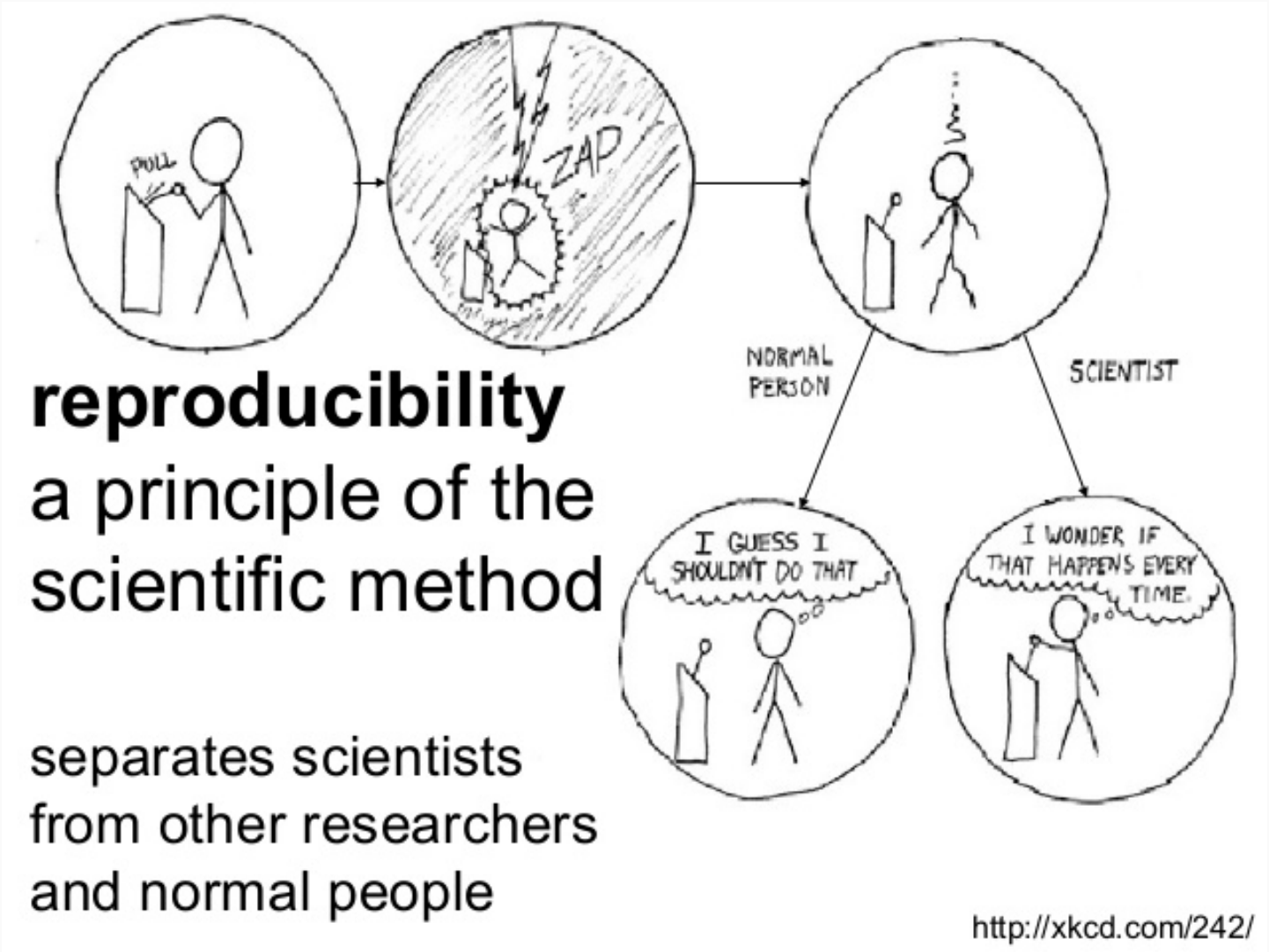


Even **you** will struggle to reproduce **your own results** from a few weeks/months ago.

You can't reproduce if you don't understand where a number came from. You can't reproduce what you don't remember. And trust me: you won't. You can't reproduce what you've lost. What if you need access to a file as it existed 1, 10, 100, or 1000 days ago?

Ben Bond-Lamberty

WHY Reproducible Science?



Carole Goble http://www.slideshare.net/carolegoble/ismb2013-keynotecleangoble

Noam Ross
@noamross

Follow

Gelman: "Reproducible research is even better when you're wrong" #stancon2017

Reproducible Science: WHY

- Fundamental pillar of **scientific method**

Reproducible Science: WHY

- Fundamental pillar of **scientific method**
- Much less prone to **errors**

Reproducible Science: WHY

- Fundamental pillar of **scientific method**
- Much less prone to **errors**
- Regenerate results **automatically**

Reproducible Science: WHY

- Fundamental pillar of **scientific method**
- Much less prone to **errors**
- Regenerate results **automatically**
- **Code reuse** & sharing accelerates scientific progress

Reproducible Science: WHY

- Fundamental pillar of **scientific method**
- Much less prone to **errors**
- Regenerate results **automatically**
- **Code reuse** & sharing accelerates scientific progress
- Increasingly required by **journals**

Reproducible Science: WHY

- Fundamental pillar of **scientific method**
- Much less prone to **errors**
- Regenerate results **automatically**
- **Code reuse** & sharing accelerates scientific progress
- Increasingly required by **journals**
- Higher publication **impact** (citations, future collaborations, etc)

HOW TO DO Reproducible Science?

Reproducible Science: HOW

1. File **organisation**.
2. **Data management**. Spreadsheet good practices.
3. **Code-based** data analysis. **Rmarkdown**
4. Software **dependencies**.
5. **Version control** & collaborative writing.

File organisation

- All files in **same directory** (Rstudio project).

File organisation

- All files in **same directory** (Rstudio project).
- **Raw data untouched** in independent folder.

File organisation

- All files in **same directory** (Rstudio project).
- **Raw data untouched** in independent folder.
- Derived, **clean data** in another folder.

File organisation

- All files in **same directory** (Rstudio project).
- **Raw data untouched** in independent folder.
- Derived, **clean data** in another folder.
- Figures, code, etc also have their own folder.

File organisation example

```
|- README          # general info about the project

|- analysis.R      # master script that executes everything

|- data-raw/       # original raw data

|- data/           # clean data (produced w/ script)

|- R/              # functions definitions

|- doc/            # manuscript files

|- figs/           # final figures

|- output/         # other code output
```

Data management

Data management

See https://www.dataone.org/best-practices

.

1. Planification (e.g. DMPTool)
2. Collection
3. Metadata description (EML, Morpho, Data Packages, DataPackageR)
4. Quality control (e.g. assertr, validate, pointblank)
5. Storage

Storage

Editorial expression of concern

IN THE 3 June issue, *Science* published the Report "Environmentally relevant concentrations of microplastic particles influence larval fish ecology" by Oona M. Lönnstedt and Peter Eklöv (*1*). The authors have notified *Science* of the theft of the computer on which the raw data for the paper were stored. These data were not backed up on any other device nor deposited in an appropriate repository. *Science* is publishing this Editorial Expression of Concern to alert our readers to the fact that no further data can be made available, beyond those already presented in the paper and its supplement, to enable readers to understand, assess, reproduce, or extend the conclusions of the paper.

Jeremy Berg

Editor in Chief

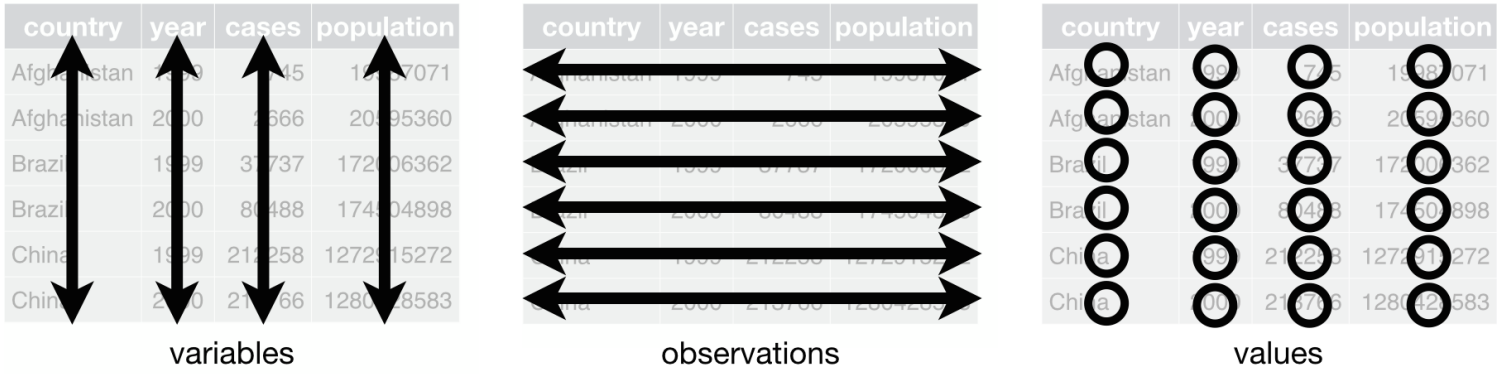
http://science.sciencemag.org/content/354/6317/1242.1

Storage

Use the **cloud**: safe, persistent, easy to share

- Open Science Framework
- GitHub
- Dropbox
- Figshare, Zenodo, etc
- See all data repositories in www.re3data.org

Tidy data



country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4

http://r4ds.had.co.nz/tidy.html

Very careful with data entry and management in Excel!

COMMENT Open Access

Gene name errors are widespread in the scientific literature

Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and.xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene sym-

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data**. Do all data manipulation through code.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data**. Do all data manipulation through code.
- Export data as plain text (txt, csv).

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data**. Do all data manipulation through code.
- Export data as plain text (txt, csv).
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data**. Do all data manipulation through code.
- Export data as plain text (txt, csv).
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/
- http://kbroman.org/dataorg/

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- **Avoid** spaces, numbers, and **special characters** in column names.
- Always **write zero values**, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as **year, month, day** in separate columns. Or **YYYY-MM-DD** as text.
- Use **Data validation** in Excel (or GForms) to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- **Don't touch raw data**. Do all data manipulation through code.
- Export data as plain text (txt, csv).
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/
- http://kbroman.org/dataorg/
- Broman & Woo: Data organization in spreadsheets

Common spreadsheet errors

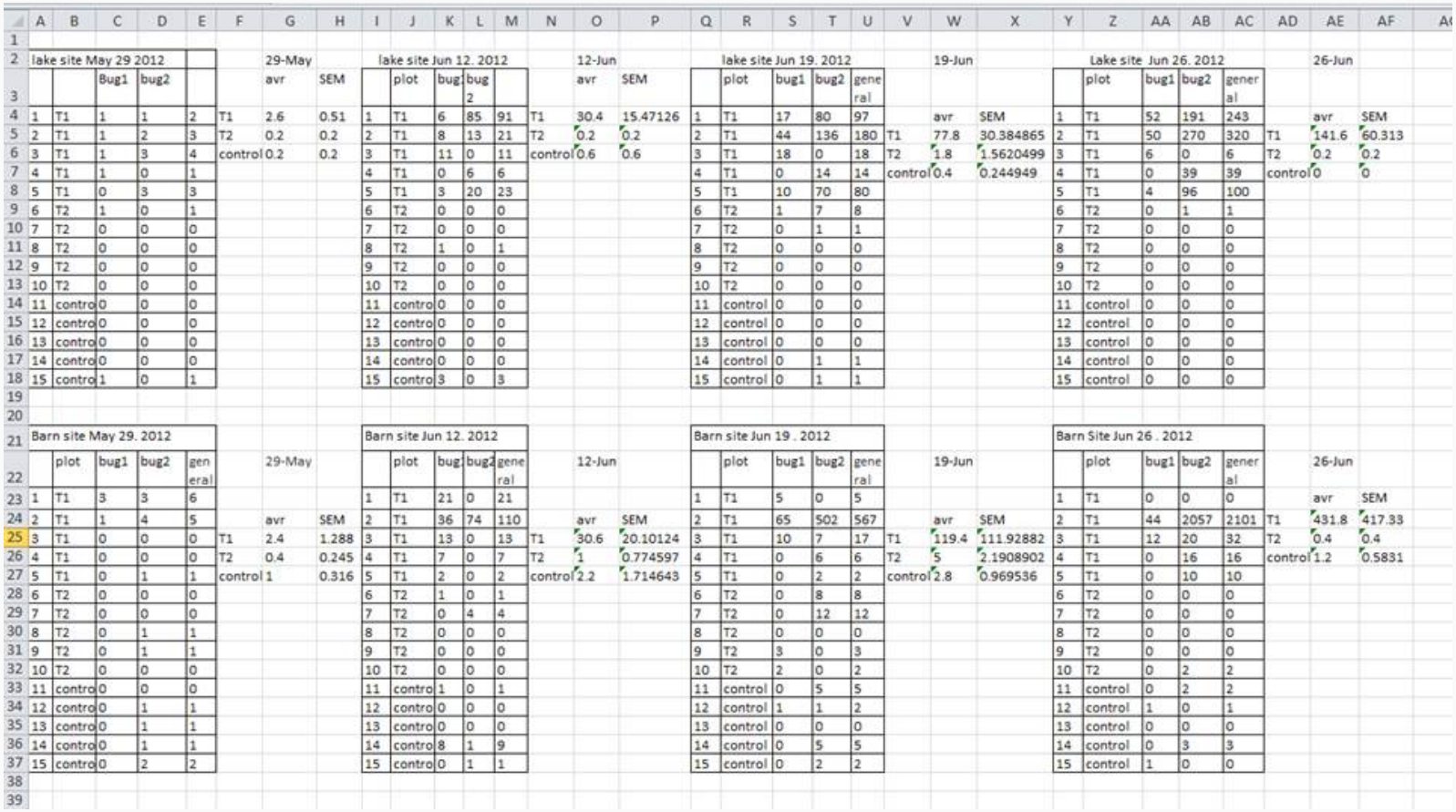
More than one variable per column

Date collected	Plot	Species-Sex	Weight
1/9/78	1	DM-M	40
1/9/78	1	DM-F	36
1/9/78	1	DS-F	135
1/20/78	1	DM-F	39
1/20/78	2	DM-M	43
1/20/78	2	DS-F	144
3/13/78	2	DM-F	51
3/13/78	2	DM-F	44
3/13/78	2	DS-F	146

Source: Data Carpentry

Date collected	Plot	Species	Sex	Weight
1/9/78	1	DM	M	40
1/9/78	1	DM	F	36
1/9/78	1	DS	F	135
1/20/78	1	DM	F	39
1/20/78	2	DM	M	43
1/20/78	2	DS	F	144
3/13/78	2	DM	F	51
3/13/78	2	DM	F	44
3/13/78	2	DS	F	146

Multiple tables



| | | |

Multiple tabs

Could you avoid new tab by adding a column to original spreadsheet?

Using formatting, comments, etc to convey information

Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52
	measurement device not calibrated		

Date collecte	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

Your turn: tidy up this messy dataset

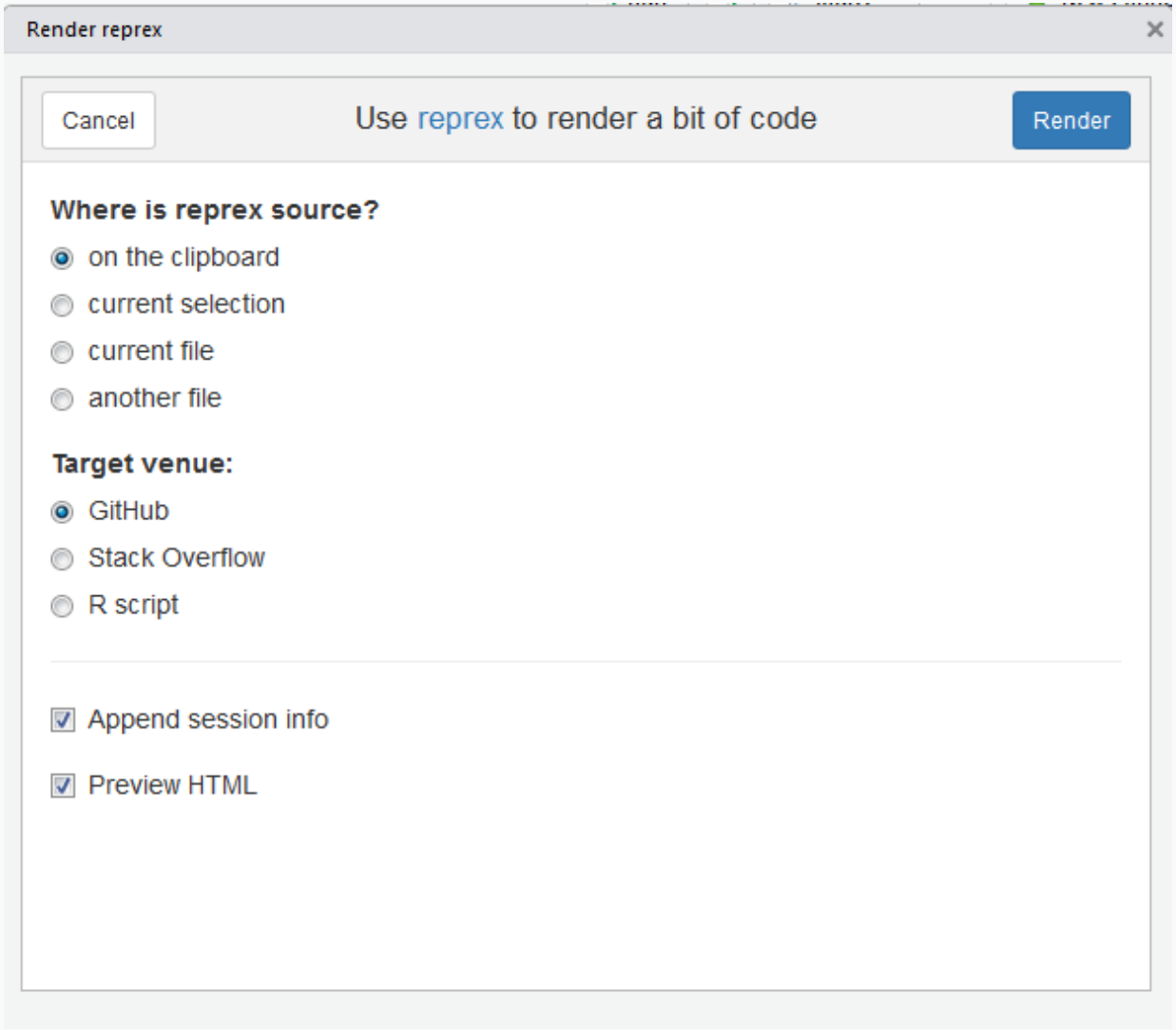
https://ndownloader.figshare.com/files/2252083

Data analysis

Always use code

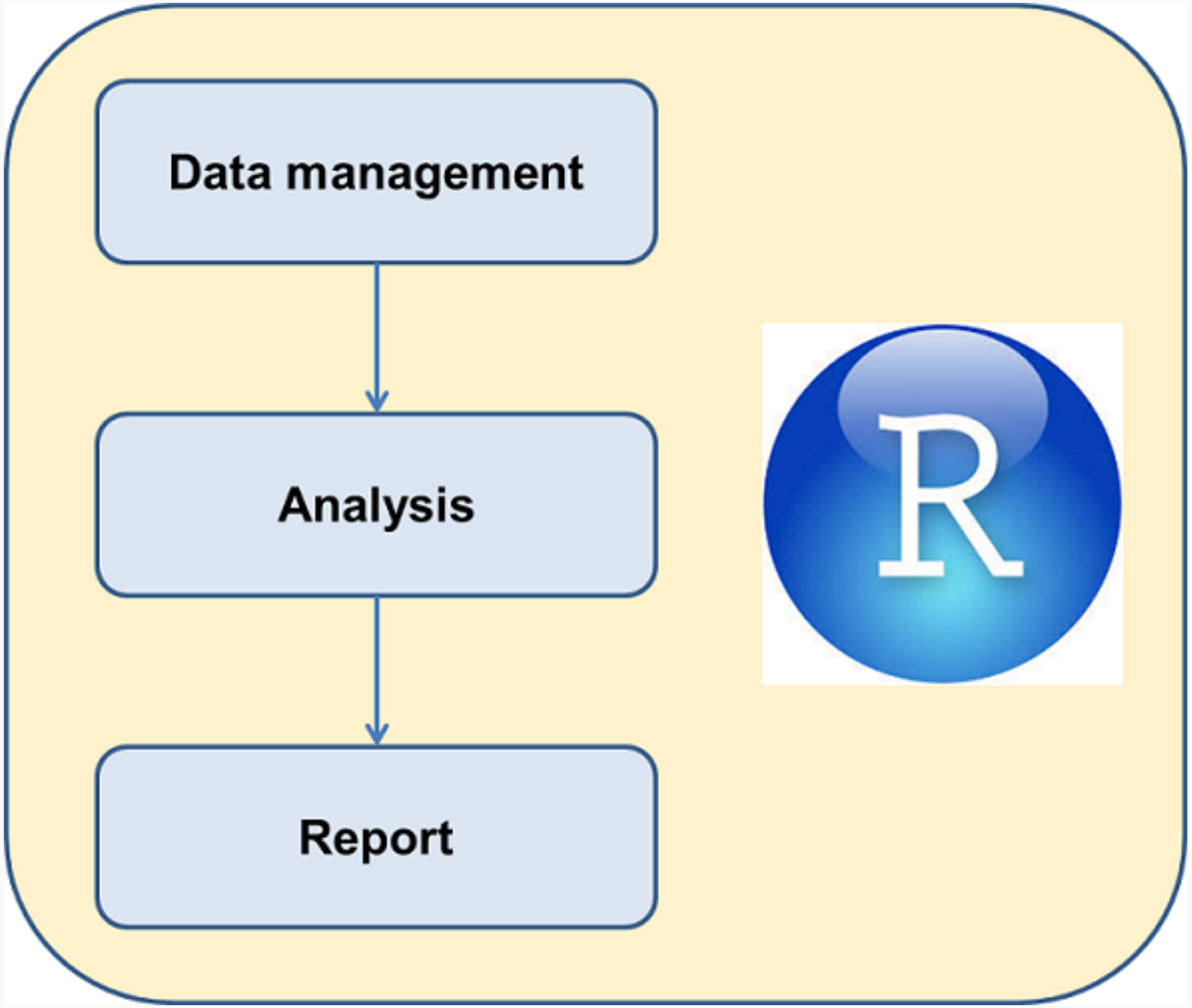
- Reproducible
- Reusable

Stuck with error / Found bug? Use `reprex`



https://reprex.tidyverse.org/

Dynamic reports



Rmarkdown documents

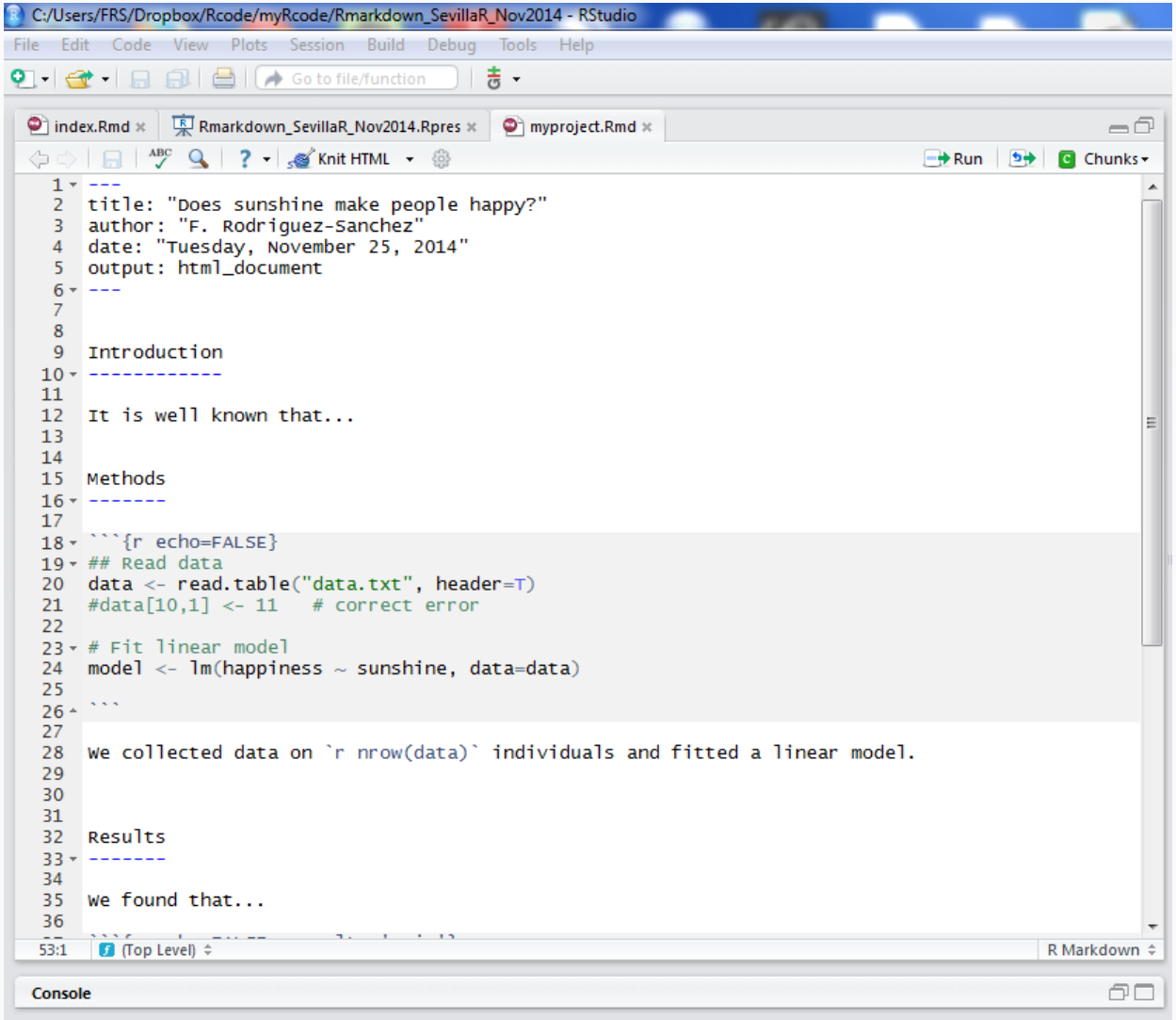
- Fully reproducible (trace all results inc. tables and plots)
- Dynamic (regenerate with 1 click)
- Suitable for
  - documents (HTML, Word, PDF)
  - presentations (HTML, PDF, PowerPoint)
  - books
  - websites
  - ...

Let's see Rmarkdown in action

In Rstudio, create new Rmarkdown document and click on `Knit HTML`.

Example: Does sunshine influence happiness?

See myproject.Rmd (http://bit.ly/rmdsun)



HTML output includes text, plot and formatted table

Does sunshine make people happy?

F. Rodriguez-Sanchez

Tuesday, November 25, 2014

Introduction

It is well known that individual well-being can be influenced by climatic conditions. However, ...

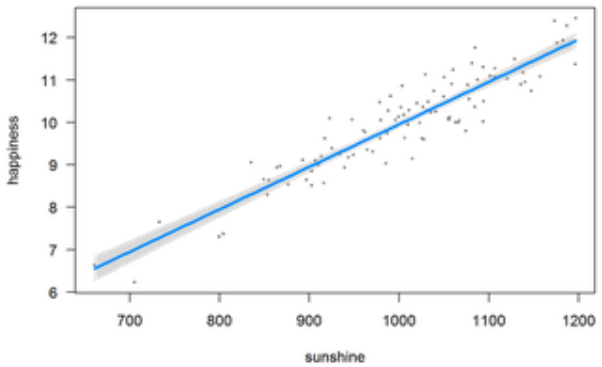
Methods

We collected data on 100 individuals and fitted a linear model.

Results

We found that...

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0651657	0.4264970	-0.1527928	0.8788758
sunshine	0.0100228	0.0004232	23.6833264	0.0000000



Discussion

These results confirm that sunshine is good for happiness (slope = 0.0100228).

Acknowledgements

Y. Xie, J. MacFarlane, Rstudio...

Spotted error in the data? No problem!

Make changes in Rmarkdown document, click `knit` and report will **update automatically!**

Why Rmarkdown?

Convince me to start using R Markdown

R Markdown rmarkdown

Darren_Dahly

1. Start using R Markdown to generate reports of your data analyses.
2. If the data changes, rerun the report with a click of the mouse.
3. Take 3 days off of work.
4. On the 4th day, tell your collaborators that the re-analysis is complete.
5. Be hailed as a hero.

https://community.rstudio.com/t/convince-me-to-start-using-r-markdown/1636/12

Other formats: PDF, Word, PowerPoint...

Does sunshine make people happy?

F. Rodriguez-Sanchez

Tuesday, November 25, 2014

Introduction

It is well known that individual well-being can be influenced by climatic conditions. However, ...

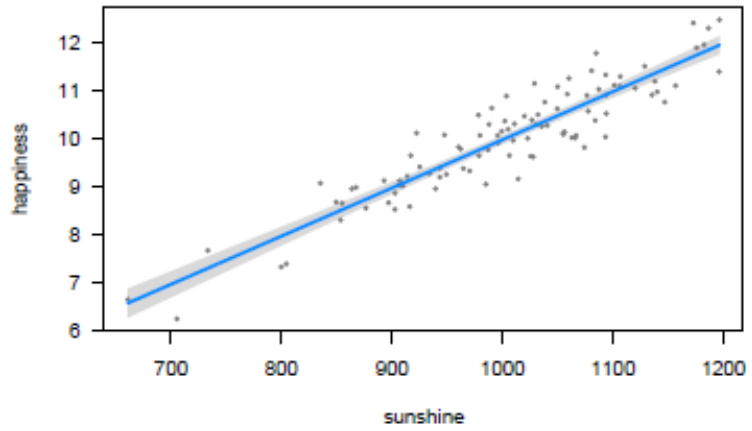
Methods

We collected data on 100 individuals and fitted a linear model.

Results

We found that...

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0651657	0.4264970	-0.1527928	0.8788758
sunshine	0.0100228	0.0004232	23.6833264	0.0000000



Does sunshine make people happy?

F. Rodriguez-Sanchez

Tuesday, November 25, 2014

Introduction

It is well known that individual well-being can be influenced by climatic conditions. However, ...

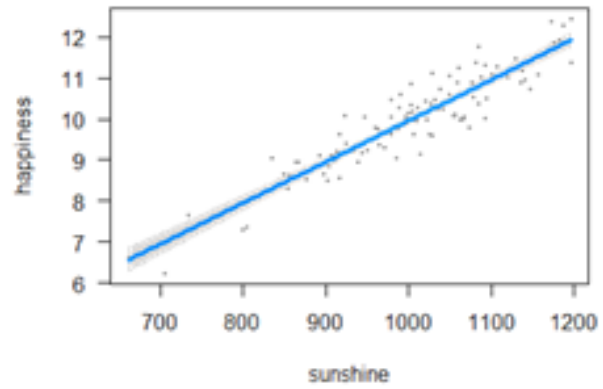
Methods

We collected data on 100 individuals and fitted a linear model.

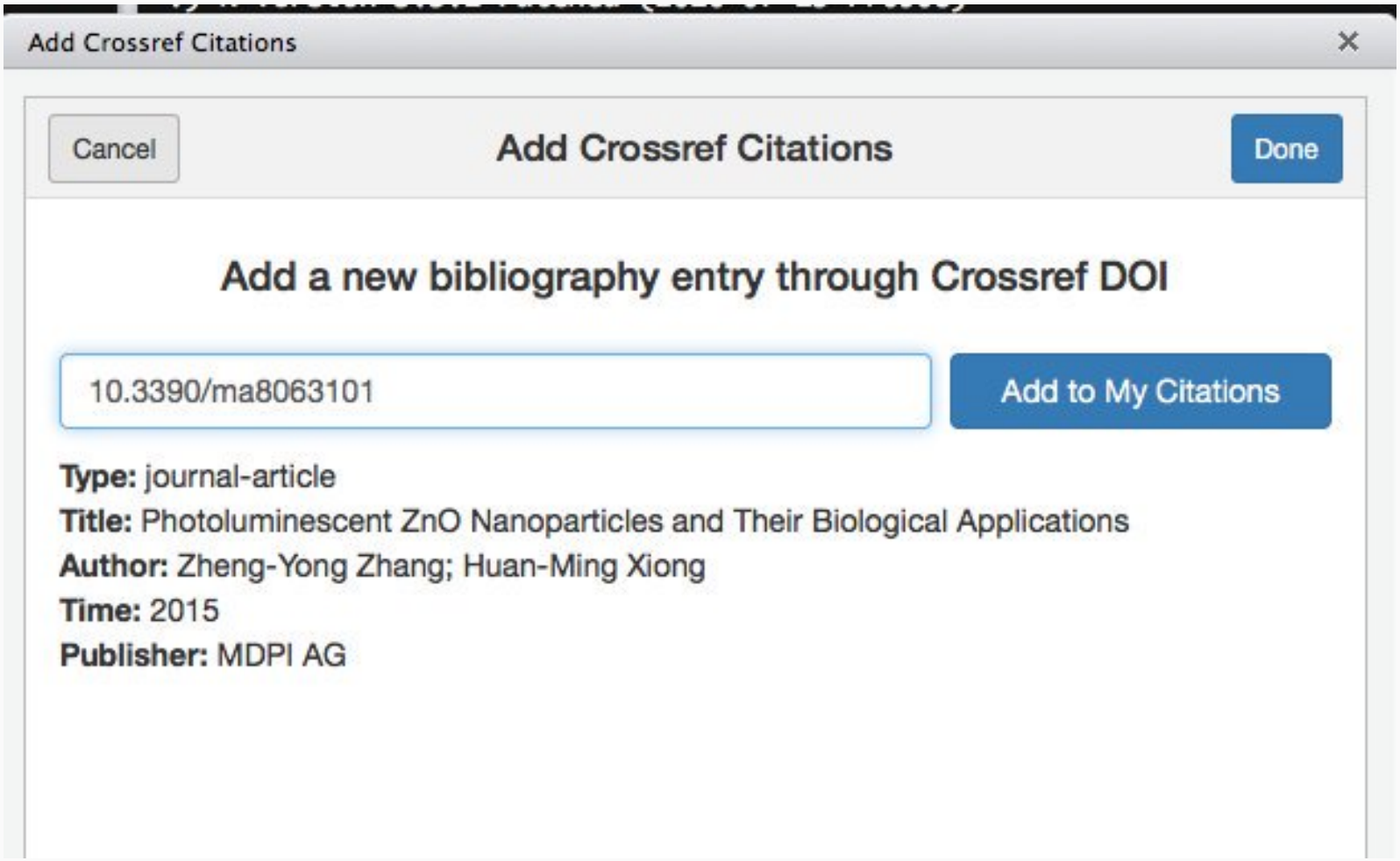
Results

We found that...

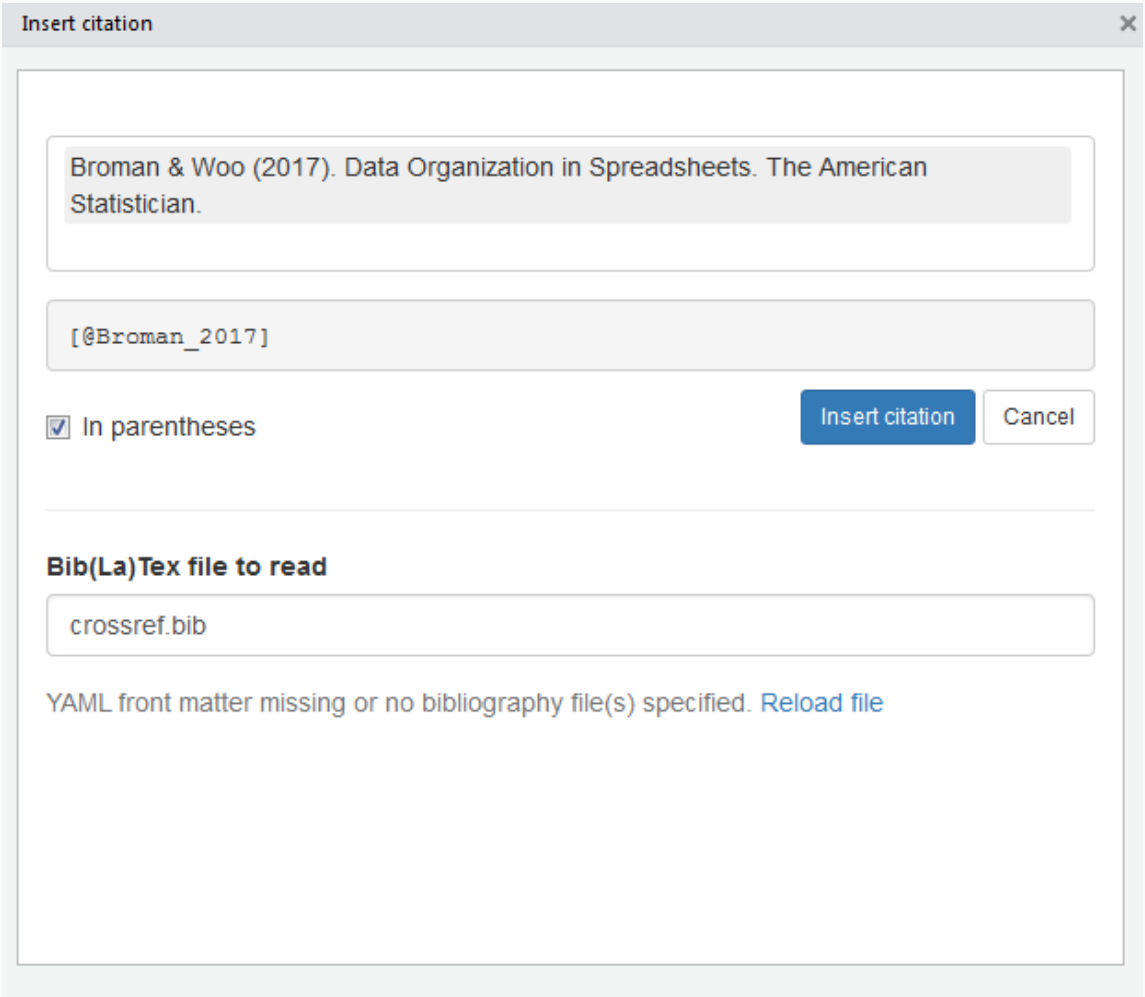
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0651657	0.4264970	-0.1527928	0.8788758
sunshine	0.0100228	0.0004232	23.6833264	0.0000000



Adding citations by DOI: rcrossref

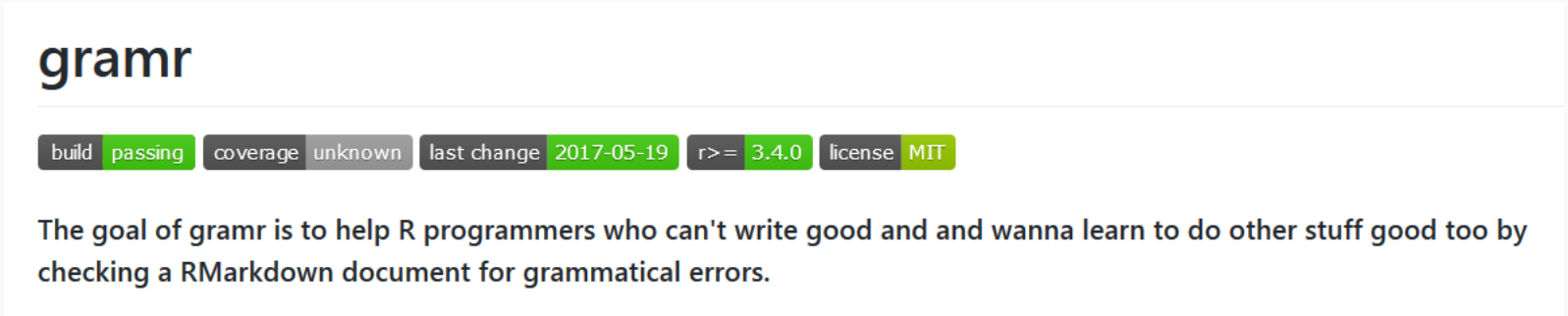


Adding citations from BibTeX file: citr



https://github.com/crsh/citr/

Revise writing style: gramr



https://github.com/ropenscilabs/gramr

Manuscript templates

- rticles: Elsevier, Springer, PeerJ, PNAS, PLoS...
- rmdTemplates
- Check out GitHub for Rmarkdown templates...

Write your next PNAS/PLoS/PeerJ/whatever in Rmarkdown!

My cool paper written in Rmarkdown

F. Rodriguez-Sanchez[a,1,2] and And Friends[a,b]

[a]Some Institute of Technology, Department, Street, City, State, Zip; [b]Another University Department, Street, City, State, Zip

This manuscript was compiled on September 19, 2018

Please provide an abstract of no more than 250 words in a single paragraph. Abstracts should explain to the general reader the major contributions of the article. References in the abstract must be cited in full within the abstract itself and cited in the text.

one | two | optional | optional | optional

This PNAS journal template is provided to help you write your work in the correct journal format. Instructions for use are provided below.

Note: please start your introduction without including the word "Introduction" as a section heading (except for math articles in the Physical Sciences section); this heading is implied in the first paragraphs.

Guide to using this template

Please note that whilst this template provides a preview of the typeset manuscript for submission, to help in this preparation, it will not necessarily be the final publication layout. For more detailed information please see the PNAS Information for Authors.

Author Affiliations. Include department, institution, and complete address, with the ZIP/postal code, for each author. Use lower case letters to match authors with institutions, as shown in the example. Authors with an ORCID ID may supply this information at submission.

Submitting Manuscripts. All authors must submit their articles at PNAScentral. If you are using Overleaf to write your article, you can use the "Submit to PNAS" option in the top bar of the editor window.

Format. Many authors find it useful to organize their manuscripts with the following order of sections; Title, Author Affiliation, Keywords, Abstract, Significance Statement, Results, Discussion, Materials and methods, Acknowledgments, and References. Other orders and headings are permitted.

Manuscript Length. PNAS generally uses a two-column format averaging 67 characters, including spaces, per line. The maximum length of a Direct Submission research article is six pages and a PNAS PLUS research article is ten pages including all text, spaces, and the number of characters displaced by figures, tables, and equations. When submitting tables, figures, and/or equations in addition to text, keep the text for your manuscript under 39,000 characters (including spaces) for Direct Submissions and 72,000 characters (including spaces) for PNAS PLUS.

References. References should be cited in numerical order as they appear in text; this will be done automatically via bibtex, e.g. (1) and (2, 3). All references, including for the SI, should be included in the main manuscript file. References appearing in both sections should not be duplicated. SI references included in tables should be included with the main reference section.

Fig. 1. Placeholder image of a frog with a longer sample caption to show justification setting.

Data Archival. PNAS must be able to archive the data essential to a published article. Where such archiving is not possible, deposition of data in public databases, such as GenBank, ArrayExpress, Protein Data Bank, Unidata, and others outlined in the Information for Authors, is acceptable.

Language-Editing Services. Prior to submission, authors who believe their manuscripts would benefit from professional editing are encouraged to use a language-editing service (see list at www.pnas.org/site/authors/language-editing.xhtml). PNAS does not take responsibility for or endorse these services, and their use has no bearing on acceptance of a manuscript for publication.

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate-educated scientist outside their field of speciality. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership. The Significance Statement appears in the paper itself and is required for all research papers.

Please provide details of author contributions here.

Please declare any conflict of interest here.

DRAFT

Some real examples

Impactos de la herbivoría por ungulados en las comunidades de plantas leñosas del Parque Natural Los Alcornocales

Francisco Rodríguez-Sánchez[1,4], Virginia Luque-Oliva[1,2], Vicente Jurado[3]

1. Departamento de Ecología Integrativa, Estación Biológica de Doñana, Consejo Superior de Investigaciones Científicas, Avda. Américo Vespucio 26, 41092 Sevilla, España.

1. Universidad Pablo de Olavide, Ctra. de Utrera s/n, Dos Hermanas, Sevilla, España.

1. Facultad de Geografía e Historia, C/ Doña María de Padilla s/n, 41004 Sevilla, España.

1. Autor para correspondencia [frodriguez.work@gmail.com]

Resumen

Las poblaciones de ciervos y otros ungulados se han incrementado enormemente en muchas áreas mediterráneas en las últimas décadas. Los impactos del incremento de la presión de herbivoría en las comunidades de plantas y la eficacia de distintas medidas de conservación están aún poco claras. En este trabajo se evalúan los efectos de la herbivoría en comunidades arbustivas del Campo de Gibraltar y Parque Natural de Los Alcornocales. Se ha comparado la diversidad de plantas, la cobertura, la altura y la intensidad de daños dentro y fuera de un total de 20 cercados de exclusión de herbívoros establecidos hace 10-15 años para favorecer la regeneración del bosque. Aunque la composición de la comunidad de plantas (riqueza de especies y diversidad) fue similar en las zonas cercadas y no cercadas, la altura de los arbustos y su cobertura fueron mayores dentro de los cercados de protección. En contraste, el daño de herbivoría fue

https://github.com/Pakillo/exclosures-Almoraima/

Can write full thesis in Rmarkdown!

See thesis.Rmd.

See thesis.pdf.

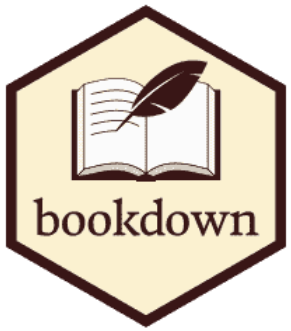
Bookdown

BOOKDOWN

Write HTML, PDF, ePub, and Kindle books with R Markdown

The **bookdown** package is an open-source R package that facilitates writing books and long-form articles/reports with R Markdown. Features include:

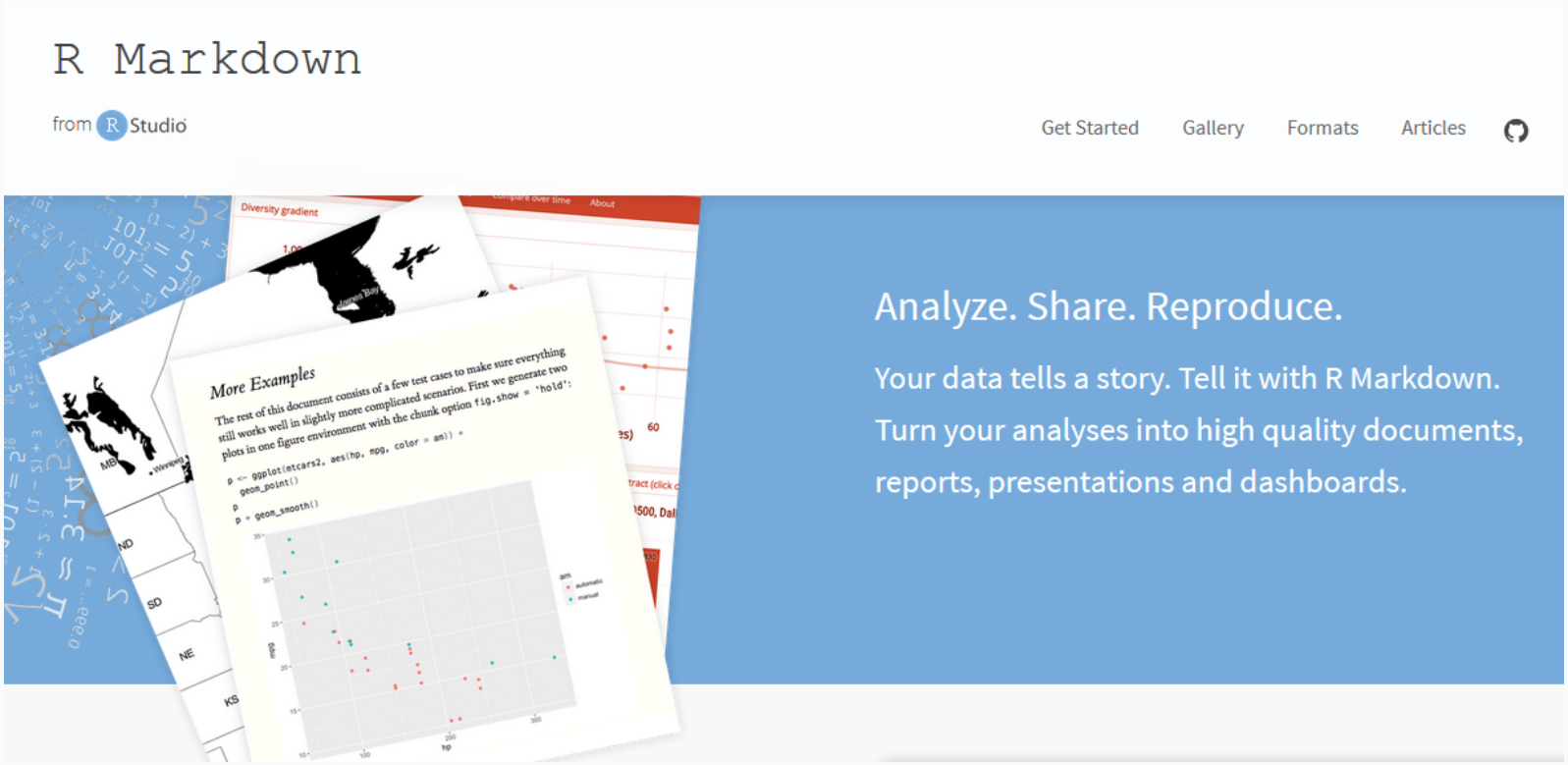
- Generate printer-ready books and ebooks from R Markdown documents.
- A markup language easier to learn than LaTeX, and to write elements such as section headers, lists, quotes, figures, tables, and citations.
- Multiple choices of output formats: PDF, LaTeX, HTML, EPUB, and Word.
- Possibility of including dynamic graphics and interactive applications (HTML widgets and Shiny apps).
- Support a wide range of languages: R, C/C++, Python, Fortran, Julia, Shell scripts, and SQL, etc.
- LaTeX equations, theorems, and proofs work for all output formats.
- Can be published to GitHub, bookdown.org, and any web servers.
- Integrated with the RStudio IDE.
- One-click publishing to https://bookdown.org.



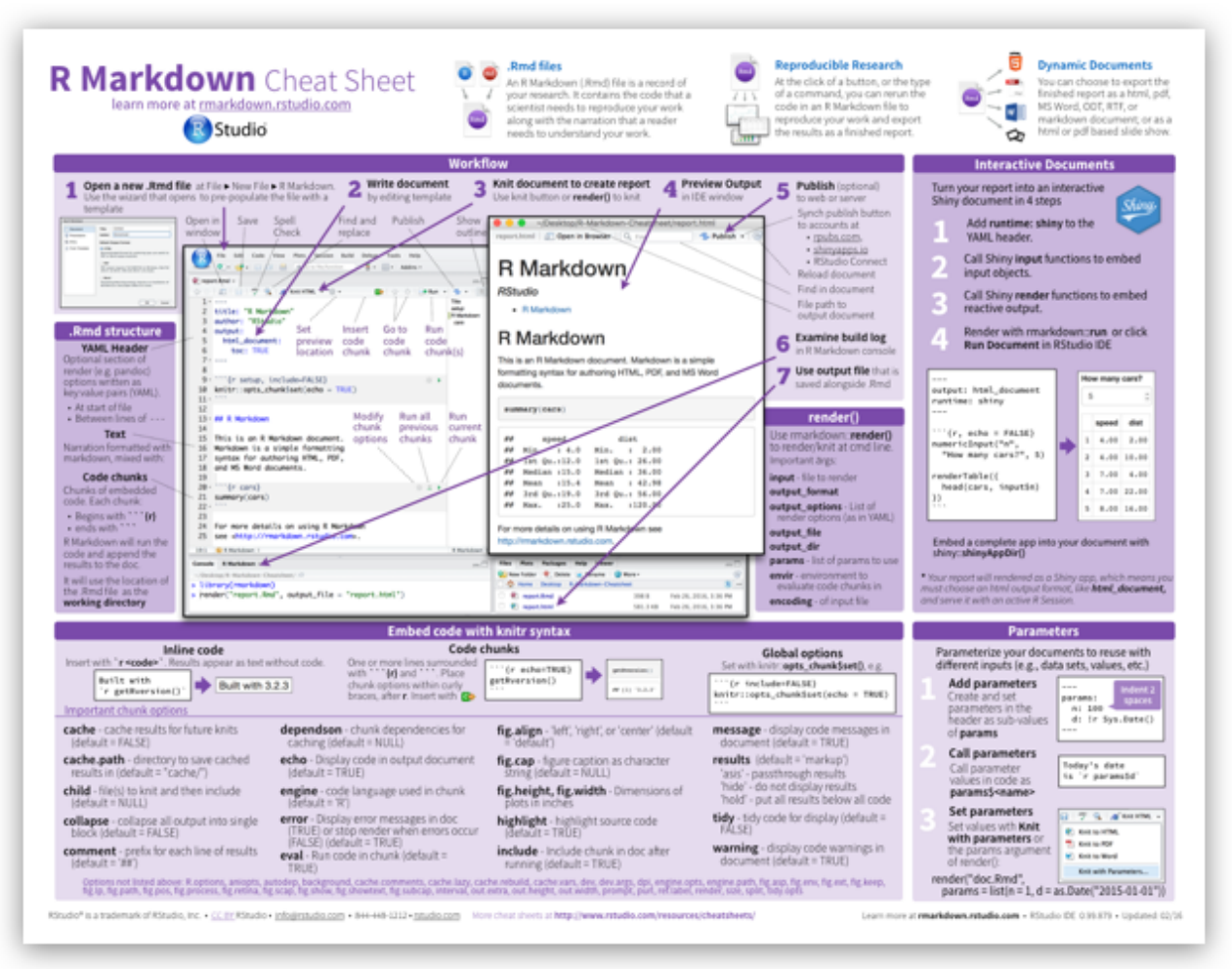
https://bookdown.org/

Rmarkdown website

http://rmarkdown.rstudio.com/index.html

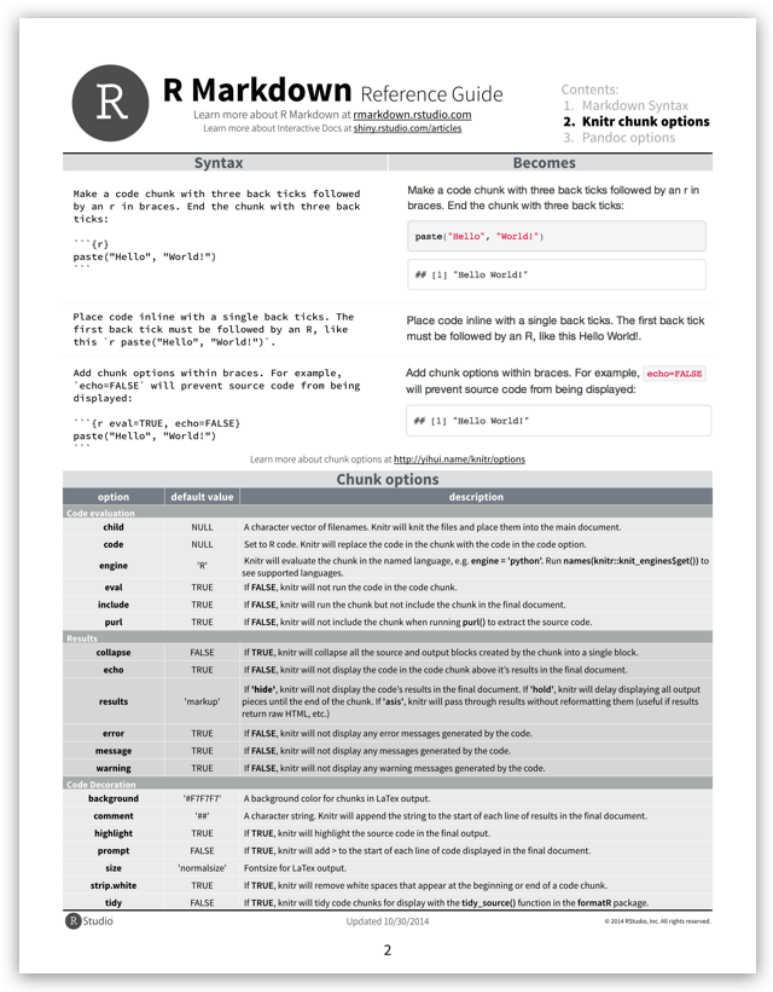


Rmarkdown cheat sheet



https://www.rstudio.org/links/r_markdown_cheat_sheet

Rmarkdown reference guide



R Markdown Reference Guide

Learn more about R Markdown at rmarkdown.rstudio.com

Learn more about Interactive Docs at shiny.rstudio.com/articles

Contents:

1. Markdown Syntax
2. **Knitr chunk options**
3. Pandoc options

Syntax	Becomes

Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks:
``{r}
paste("Hello", "World!")
``
	Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks:
	`paste("Hello", "World!")`
	`## [1] "Hello World!"`

Place code inline with a single back ticks. The first back tick must be followed by an R, like this `r paste("Hello", "World!")`.	Place code inline with a single back ticks. The first back tick must be followed by an R, like this Hello World!.

Add chunk options within braces. For example, `echo=FALSE` will prevent source code from being displayed:
``{r eval=TRUE, echo=FALSE}
paste("Hello", "World!")
``
	Add chunk options within braces. For example, `echo=FALSE` will prevent source code from being displayed:
	`## [1] "Hello World!"`

Learn more about chunk options at http://yihui.name/knitr/options

Chunk options

option	default value	description
Code evaluation		
child	NULL	A character vector of filenames. Knitr will knit the files and place them into the main document.
code	NULL	Set to R code. Knitr will replace the code in the chunk with the code in the code option.
engine	'R'	Knitr will evaluate the chunk in the named language, e.g. **engine = 'python'**. Run **names(knitr::knit_engines$get())** to see supported languages.
eval	TRUE	If **FALSE**, knitr will not run the code in the code chunk.
include	TRUE	If **FALSE**, knitr will run the chunk but not include the chunk in the final document.
purl	TRUE	If **FALSE**, knitr will not include the chunk when running **purl()** to extract the source code.
Results		
collapse	FALSE	If **TRUE**, knitr will collapse all the source and output blocks created by the chunk into a single block.
echo	TRUE	If **FALSE**, knitr will not display the code in the code chunk above it's results in the final document.
results	'markup'	If **'hide'**, knitr will not display the code's results in the final document. If **'hold'**, knitr will delay displaying all output pieces until the end of the chunk. If **'asis'**, knitr will pass through results without reformatting them (useful if results return raw HTML, etc.)
error	TRUE	If **FALSE**, knitr will not display any error messages generated by the code.
message	TRUE	If **FALSE**, knitr will not display any messages generated by the code.
warning	TRUE	If **FALSE**, knitr will not display any warning messages generated by the code.
Code Decoration		
background	'#F7F7F7'	A background color for chunks in LaTex output.
comment	'##'	A character string. Knitr will append the string to the start of each line of results in the final document.
highlight	TRUE	If **TRUE**, knitr will highlight the source code in the final output.
prompt	FALSE	If **TRUE**, knitr will add > to the start of each line of code displayed in the final document.
size	'normalsize'	Fontsize for LaTex output.
strip.white	TRUE	If **TRUE**, knitr will remove white spaces that appear at the beginning or end of a code chunk.
tidy	FALSE	If **TRUE**, knitr will tidy code chunks for display with the **tidy_source()** function in the **formatR** package.

RStudio Updated 10/30/2014 © 2014 RStudio, Inc. All rights reserved.

2

https://www.rstudio.org/links/r_markdown_reference_guide

Rmarkdown book

R Markdown: The Definitive Guide

by Yihui Xie, J. J. Allaire, Garrett Grolemund

2018-09-11

Star 239

The first official book authored by the core R Markdown developers that provides a comprehensive and accurate reference to the R Markdown ecosystem. With R Markdown, you can easily create reproducible data analysis reports, presentations, dashboards, interactive applications, books, dissertations, websites, and journal articles, while enjoying the simplicity of Markdown and the great power of R and other languages. *Read more* →

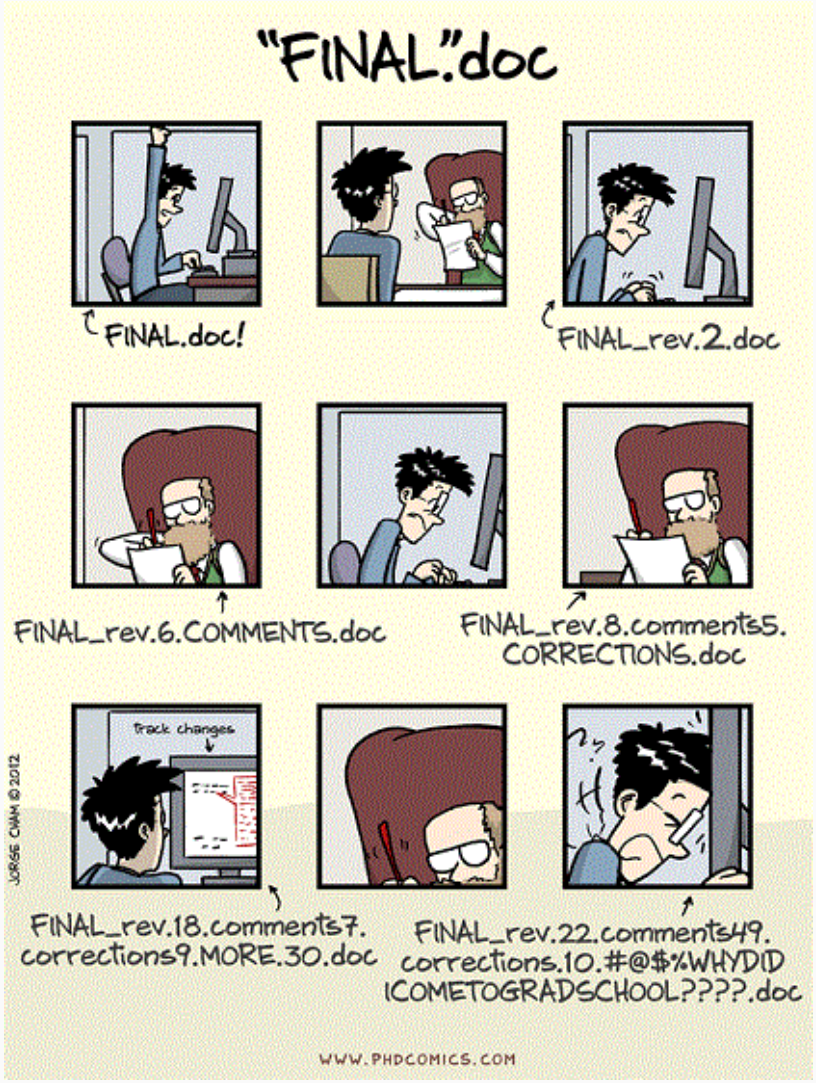
https://bookdown.org/yihui/rmarkdown/

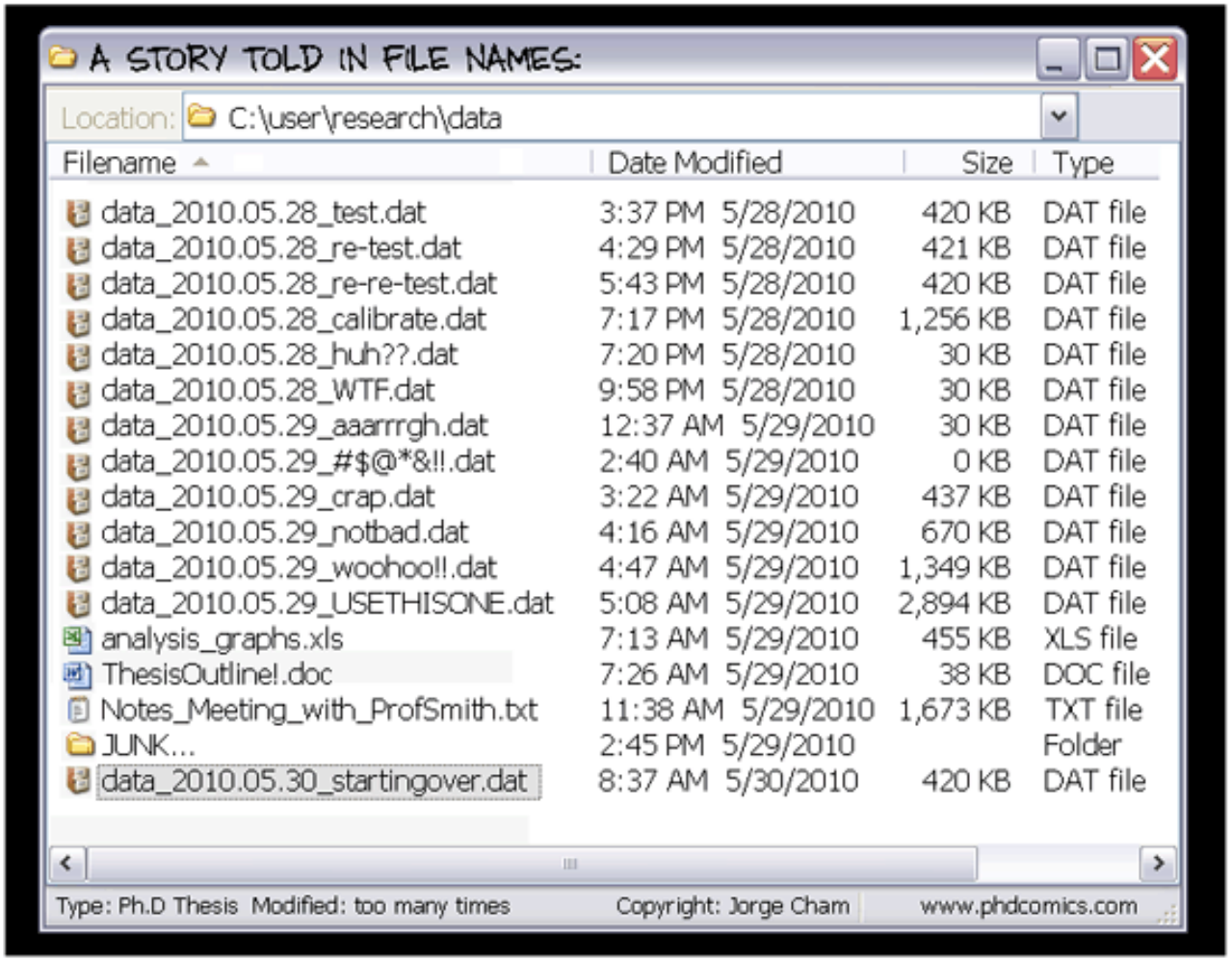
Managing software dependencies

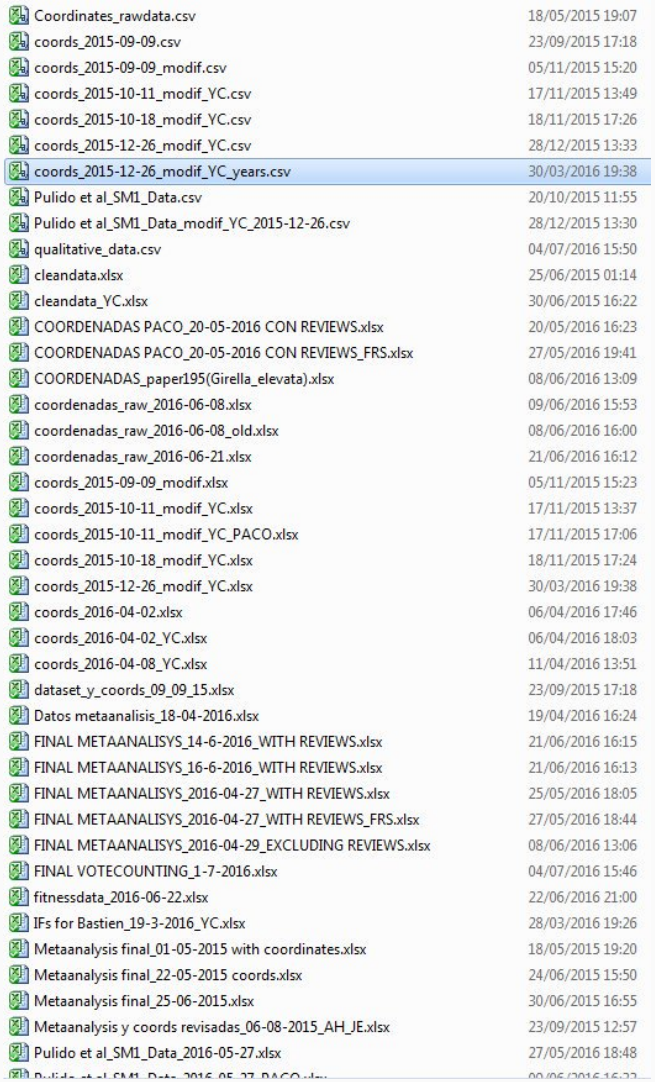
Managing package dependencies in R

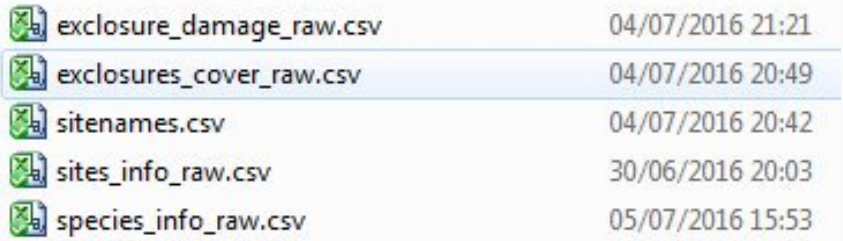
- **sessionInfo** (or session_info)
- **checkpoint**
- packrat
- switchr
- docker

Version control


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM


A STORY TOLD IN FILE NAMES:
Location: C:\user\research\data
Filename
Date Modified
Size
Type
data_2010.05.28_test.dat 3:37 PM 5/28/2010 420 KB DAT file
data_2010.05.28_re-test.dat 4:29 PM 5/28/2010 421 KB DAT file
data_2010.05.28_re-re-test.dat 5:43 PM 5/28/2010 420 KB DAT file
data_2010.05.28_calibrate.dat 7:17 PM 5/28/2010 1,256 KB DAT file
data_2010.05.28_huh??.dat 7:20 PM 5/28/2010 30 KB DAT file
data_2010.05.28_WTF.dat 9:58 PM 5/28/2010 30 KB DAT file
data_2010.05.29_aaarrrgh.dat 12:37 AM 5/29/2010 30 KB DAT file
data_2010.05.29_#$@*&!!.dat 2:40 AM 5/29/2010 0 KB DAT file
data_2010.05.29_crap.dat 3:22 AM 5/29/2010 437 KB DAT file
data_2010.05.29_notbad.dat 4:16 AM 5/29/2010 670 KB DAT file
data_2010.05.29_woohoo!!.dat 4:47 AM 5/29/2010 1,349 KB DAT file
data_2010.05.29_USETHISONE.dat 5:08 AM 5/29/2010 2,894 KB DAT file
analysis_graphs.xls 7:13 AM 5/29/2010 455 KB XLS file
ThesisOutline!.doc 7:26 AM 5/29/2010 38 KB DOC file
Notes_Meeting_with_ProfSmith.txt 11:38 AM 5/29/2010 1,673 KB TXT file
JUNK... 2:45 PM 5/29/2010 Folder
data_2010.05.30_startingover.dat 8:37 AM 5/30/2010 420 KB DAT file
Type: Ph.D Thesis Modified: too many times
Copyright: Jorge Cham
www.phdcomics.com


Coordinates_rawdata.csv 18/05/2015 19:07
coords_2015-09-09.csv 23/09/2015 17:18
coords_2015-09-09_modif.csv 05/11/2015 15:20
coords_2015-10-11_modif_YC.csv 17/11/2015 13:49
coords_2015-10-18_modif_YC.csv 18/11/2015 17:26
coords_2015-12-26_modif_YC.csv 28/12/2015 13:33
coords_2015-12-26_modif_YC_years.csv 30/03/2016 19:38
Pulido et al_SM1_Data.csv 20/10/2015 11:55
Pulido et al_SM1_Data_modif_YC_2015-12-26.csv 28/12/2015 13:30
qualitative_data.csv 04/07/2016 15:50
cleandata.xlsx 25/06/2015 01:14
cleandata_YC.xlsx 30/06/2015 16:22
COORDENADAS PACO_20-05-2016 CON REVIEWS.xlsx 20/05/2016 16:23
COORDENADAS PACO_20-05-2016 CON REVIEWS_FRS.xlsx 27/05/2016 19:41
COORDENADAS_paper195(Girella_elevata).xlsx 08/06/2016 13:09
coordenadas_raw_2016-06-08.xlsx 09/06/2016 15:53
coordenadas_raw_2016-06-08_old.xlsx 08/06/2016 16:00
coordenadas_raw_2016-06-21.xlsx 21/06/2016 16:12
coords_2015-09-09_modif.xlsx 05/11/2015 15:23
coords_2015-10-11_modif_YC.xlsx 17/11/2015 13:37
coords_2015-10-11_modif_YC_PACO.xlsx 17/11/2015 17:06
coords_2015-10-18_modif_YC.xlsx 18/11/2015 17:24
coords_2015-12-26_modif_YC.xlsx 30/03/2016 19:38
coords_2016-04-02.xlsx 06/04/2016 17:46
coords_2016-04-02_YC.xlsx 06/04/2016 18:03
coords_2016-04-08_YC.xlsx 11/04/2016 13:51
dataset_y_coords_09_09_15.xlsx 23/09/2015 17:18
Datos metaanalisis_18-04-2016.xlsx 19/04/2016 16:24
FINAL METAANALISYS_14-6-2016_WITH REVIEWS.xlsx 21/06/2016 16:15
FINAL METAANALISYS_16-6-2016_WITH REVIEWS.xlsx 21/06/2016 16:13
FINAL METAANALISYS_2016-04-27_WITH REVIEWS.xlsx 25/05/2016 18:05
FINAL METAANALISYS_2016-04-27_WITH REVIEWS_FRS.xlsx 27/05/2016 18:44
FINAL METAANALISYS_2016-04-29_EXCLUDING REVIEWS.xlsx 08/06/2016 13:06
FINAL VOTECOUNTING_1-7-2016.xlsx 04/07/2016 15:46
fitnessdata_2016-06-22.xlsx 22/06/2016 21:00
IFs for Bastien_19-3-2016_YC.xlsx 28/03/2016 19:26
Metaanalysis final_01-05-2015 with coordinates.xlsx 18/05/2015 19:20
Metaanalysis final_22-05-2015 coords.xlsx 24/06/2015 15:50
Metaanalysis final_25-06-2015.xlsx 30/06/2015 16:55
Metaanalysis y coords revisadas_06-08-2015_AH_JE.xlsx 23/09/2015 12:57
Pulido et al_SM1_Data_2016-05-27.xlsx 27/05/2016 18:48

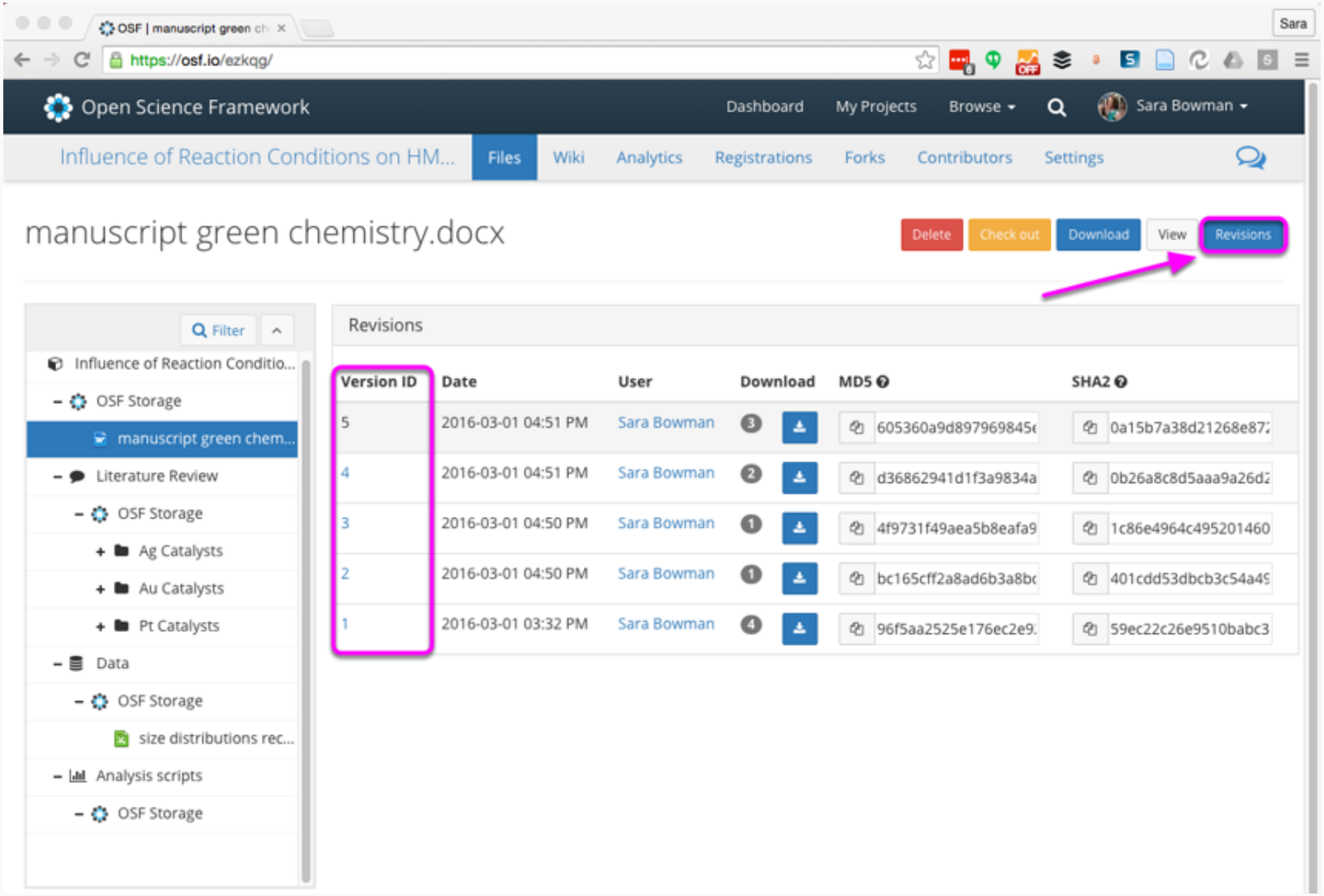

exclosure_damage_raw.csv 04/07/2016 21:21
exclosures_cover_raw.csv 04/07/2016 20:49
sitenames.csv 04/07/2016 20:42
sites_info_raw.csv 30/06/2016 20:03
species_info_raw.csv 05/07/2016 15:53

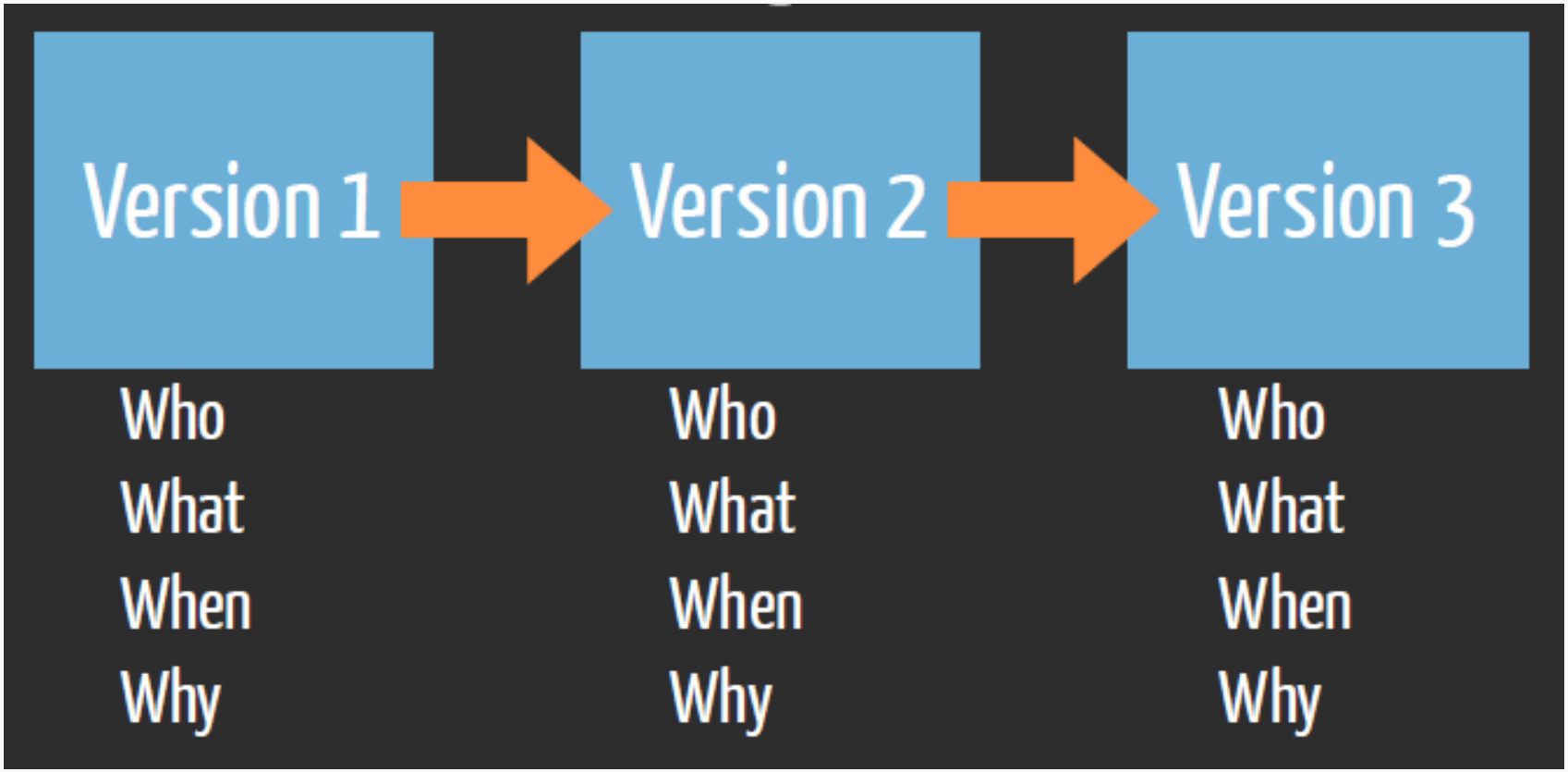
Dropbox

Dropbox keeps record of deleted/edited files for 30 days

Open Science Framework

Automatic version control, no time limit.





R. Fitzjohn (https://github.com/richfitz/reproducibility-2014)

Git & GitHub

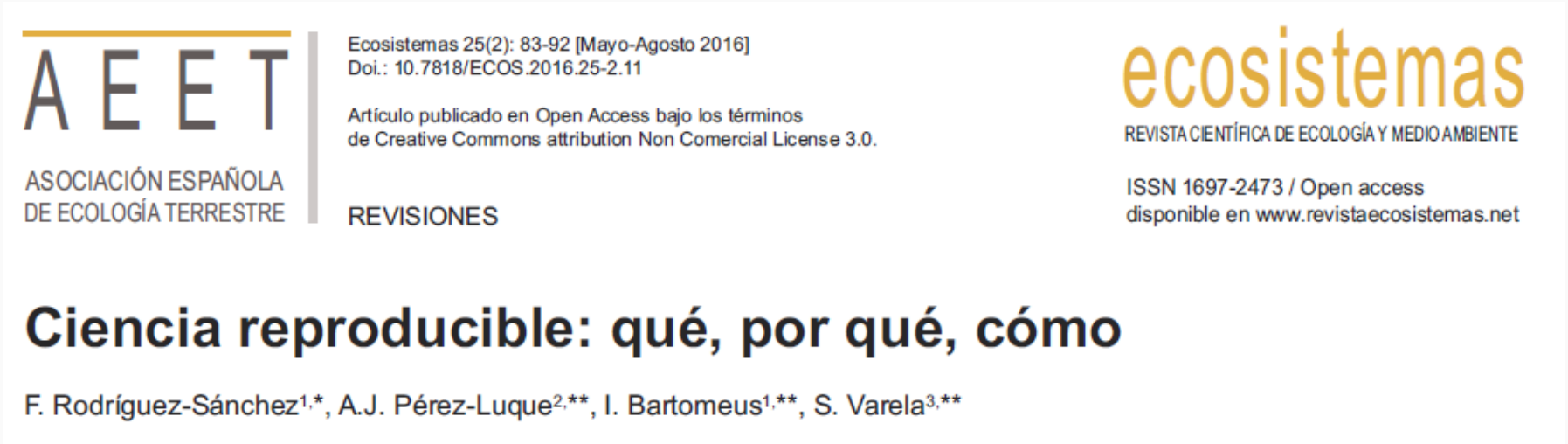
- Sign up for GitHub
- Install Git
- Introduce yourself
- Create repo on GitHub
- Clone repo in Rstudio
- Make changes, push, pull
- Collaboration

Collaborative writing

Many alternatives

- Rmarkdown + GitHub
- Word + Dropbox
- Google Docs
- Overleaf
- Authorea
- ...

To read more



AEET
ASOCIACIÓN ESPAÑOLA DE ECOLOGÍA TERRESTRE

Ecosistemas 25(2): 83-92 [Mayo-Agosto 2016]
Doi.: 10.7818/ECOS.2016.25-2.11

Artículo publicado en Open Access bajo los términos
de Creative Commons attribution Non Comercial License 3.0.

REVISIONES

ecosistemas
REVISTA CIENTÍFICA DE ECOLOGÍA Y MEDIO AMBIENTE

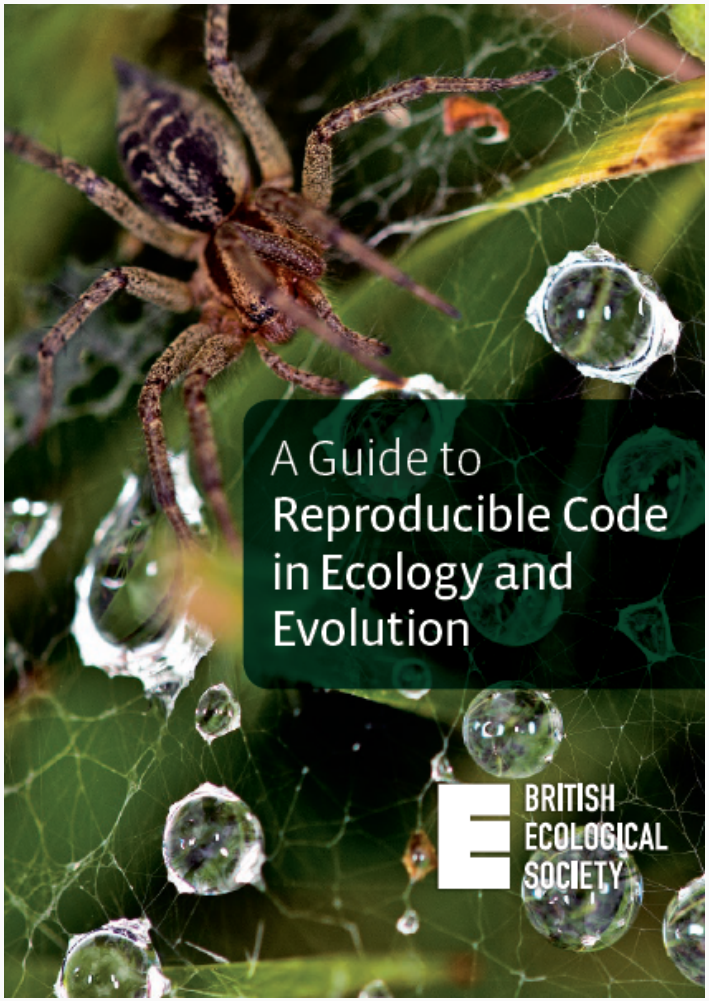
ISSN 1697-2473 / Open access
disponible en www.revistaecosistemas.net

Ciencia reproducible: qué, por qué, cómo

F. Rodríguez-Sánchez[1,*], A.J. Pérez-Luque[2,**], I. Bartomeus[1,**], S. Varela[3,**]

http://www.revistaecosistemas.net/index.php/ecosistemas/article/viewFile/1178/973

To read more



https://www.britishecologicalsociety.org/publications/guides-to/

Happy writing!

Slides and source code available at

https://github.com/Pakillo/ReproducibleScience