

Ciencia Reproducible

Y buenas prácticas al momento de escribir código

Mg. Yanina Bellini Saibene - INTA Anguil

Dra. María Florencia D'Andrea - IRB - CNIA



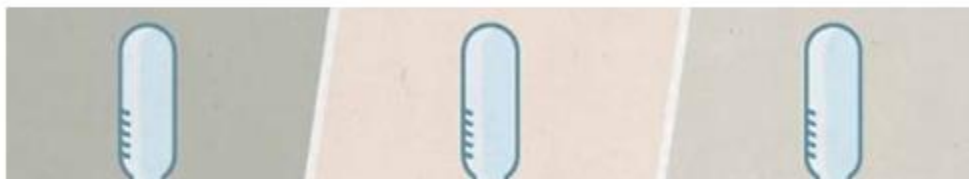
Instituto Nacional
de Tecnología Agropecuaria

SPECIAL | 18 OCTOBER 2018

Challenges in irreproducible research

Science moves forward by corroboration – we verify others' results. Science advances faster when we waste less time pursuing false leads. No research can ever be considered to be the final word, but the results that do not stand up to further study.

There is growing alarm about results that cannot be reproduced. Explanations include increased complexity of experiments and statistics, and the role of researchers. Journals, scientists, institutions all have a part in tackling reproducibility. Nature is taking substantive steps to improve the transparency in what we publish, and to promote awareness.



AAAS

Become a Member

Science

Contents ▾

News ▾

Careers ▾

Journals ▾

SHARE

PERSPECTIVE



Reproducible Research in Computational Science

Roger D. Peng

+ See all authors and affiliations

Science 02 Dec 2011:
Vol. 334, Issue 6060, pp. 1226-1227
DOI: 10.1126/science.1213847

Article

Figures & Data

Info & Metrics

eLetters

PDF

¿Qué es ciencia Reproducible?



Los resultados de una investigación son **reproducibles** si hay información disponible para que **investigadores independientes** lleguen a los mismos resultados usando idénticos procedimientos (King, 1995, 444).

> Computación

los datos y el código utilizados están disponibles

Reproducibilidad

```
graph TD; A[Reproducibilidad] --> B[Empírica]; A --> C[Computacional]; A --> D[Estadística];
```

Empírica

Detalles de reactivos
Líneas celulares
Identidad de las muestras
Configuración de instrumentos

Computacional

Detalles sobre el código
Software
Hardware e
implementation

Estadística

Detalles de tests estadísticos
Parametros del modelo
Valores umbrales

¿Por qué R me permite hacer ciencia reproducible?



Cuando escribís código, **explícitamente documentas los pasos de tu trabajo.**

```
ggsave(plot1, "C:/tp_2019/graficos/figura1.png", width = 1000, height = 500)
```

Esto promueve la reproducibilidad más que interacciones típicas con programas de interfaz gráfica de usuario

«Sin instrucciones claras, muchos investigadores luchan para evitar el caos en sus estructuras de archivos, y por eso son comprensiblemente reacios a exponer su flujo de trabajo para que otros lo vean. Esta puede ser una de las razones de la falta de respuesta o rechazo de muchos investigadores cuando se les solicita detalles sobre el método, incluidos los datos y código (Collberg and Proebsting 2016).»



Marwick B., Boettiger C. & Mullen L.
(2018) Packaging Data Analytical Work
Reproducibly Using R (and Friends), The
American Statistician, 72:1, 80-88

The R Series

Dynamic Documents with R and knitr

The R Series

Implementing Reproducible Research



Edited by
Victoria Stodden
Friedrich Leisch
Roger D. Peng

 **CRC Press**
Taylor & Francis Group
A CHAPMAN & HALL BOOK

The R Series

Reproducible Research with R and RStudio

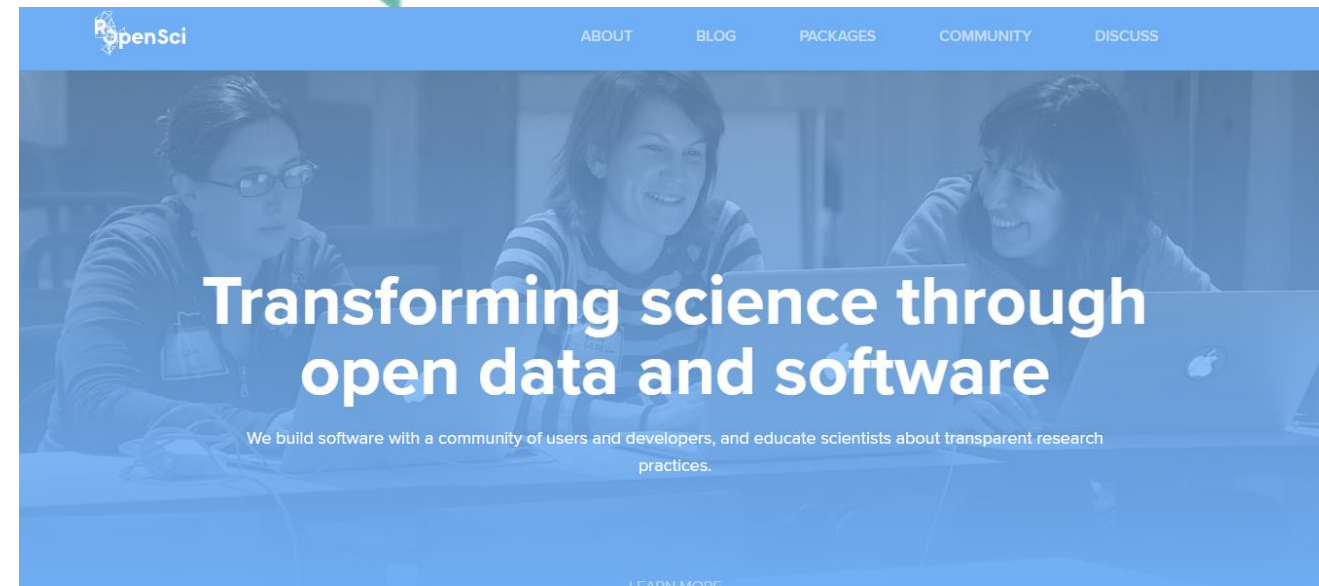
Root			
Research Project			
Data	Analysis	Presentation	
MainData.csv	MainAnalysis.R	Article	Other
GatherSource	ResultsFigures	Paper.Rnw	Slideshow
Makefile	Figure1.R	Main.bib	Slideshow.Rnw
MergeData.R	Figure2.R	Packages.bib	Website
Gather1.R	Figure3.R	figure	Website.Rmd
Gather2.R		Figure1.pdf	
Gather3.R		Figure2.pdf	
		Figure3.pdf	

Christopher Gandrud

 **CRC Press**
Taylor & Francis Group
A CHAPMAN & HALL BOOK



ROpenSci



**Transforming science through
open data and software**

We build software with a community of users and developers, and educate scientists about transparent research practices.

[LEARN MORE](#)

<https://cran.r-project.org/web/views/ReproducibleResearch.html>

<https://ropensci.org/>

1) Estilo de código



2) Uso de proyectos

3) Programación «literaria»

4) Control de Versiones

Nombres de variables

```
# Good  
day_one  
day_1  
  
# Bad  
DayOne  
dayone
```

- Minúscula
- Separaciones de palabras preferentemente usando _
- De ser posible no elegir nombre de funciones ya existentes como nombres de variables

avoid **dots** in names

confusing to others

ambiguous S3 dispatch

real runtime errors

style.tidyverse.org

adv-r.hadley.nz

@jimhester
@jimhester_

<https://style.tidyverse.org/>
<http://r-pkgs.had.co.nz/style.html>

Estilo del código



Usar guías de estilo para hacer el código fácilmente legible para otros (formatR / lintR / styleR)

Esta basado en una guía de estilo de R

Estilo del código

- Una línea de código debería tener **menos de 80 caracteres**
- Siempre debería haber un espacio luego una coma
- Los operadores `==`, `+`, `-`, `<-` deberían estar siempre rodeados de espacios

Good

```
x[, 1]
```

Bad

```
x[,1]
```

```
x[ ,1]
```

```
x[ , 1]
```

Los nombres de archivos deben poder ser leídos por humanos (o por ustedes dentro de 6 meses)



NO

myabstract.docx

Joe's Filenames Use Spaces and Punctuation.xlsx

figure 1.png

fig 2.png

JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt

YES

2014-06-08_abstract-for-sla.docx

joes-filenames-are-getting-better.xlsx

fig01_scatterplot-talk-length-vs-interest.png

fig02_histogram-talk-attendance.png

1986-01-28_raw-data-from-challenger-o-rings.txt

Los nombres de archivos deben poder ser leídos por humanos ¡y por la máquina!



Se bueno con vos mismo y evita:

- **Espacios** en los nombres de los archivos
- **Acentos** y ñ
- Diferencias en archivos por uso de **mayúsculas** en los nombres

Es fácil extraer información con R leyendo el nombre del archivo

-> separar con «-» y «_»

Todos los archivos deben poder ser leídos por humanos y por la máquina!



```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
```

```
> flist <- list.files(pattern = "Plasmid") %>% head
```

```
> stringr::str_split_fixed(flist, "[_\\.]", 5)
```

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	"2013-06-26"	"BRAFWTNEGASSAY"	"Plasmid-Cellline-100-1MutantFraction"	"A01"	"csv"
[2,]	"2013-06-26"	"BRAFWTNEGASSAY"	"Plasmid-Cellline-100-1MutantFraction"	"A02"	"csv"
[3,]	"2013-06-26"	"BRAFWTNEGASSAY"	"Plasmid-Cellline-100-1MutantFraction"	"A03"	"csv"
[4,]	"2013-06-26"	"BRAFWTNEGASSAY"	"Plasmid-Cellline-100-1MutantFraction"	"B01"	"csv"
[5,]	"2013-06-26"	"BRAFWTNEGASSAY"	"Plasmid-Cellline-100-1MutantFraction"	"B02"	"csv"
[6,]	"2013-06-26"	"BRAFWTNEGASSAY"	"Plasmid-Cellline-100-1MutantFraction"	"B03"	"csv"

date

assay

sample set

well

1) Estilo de código

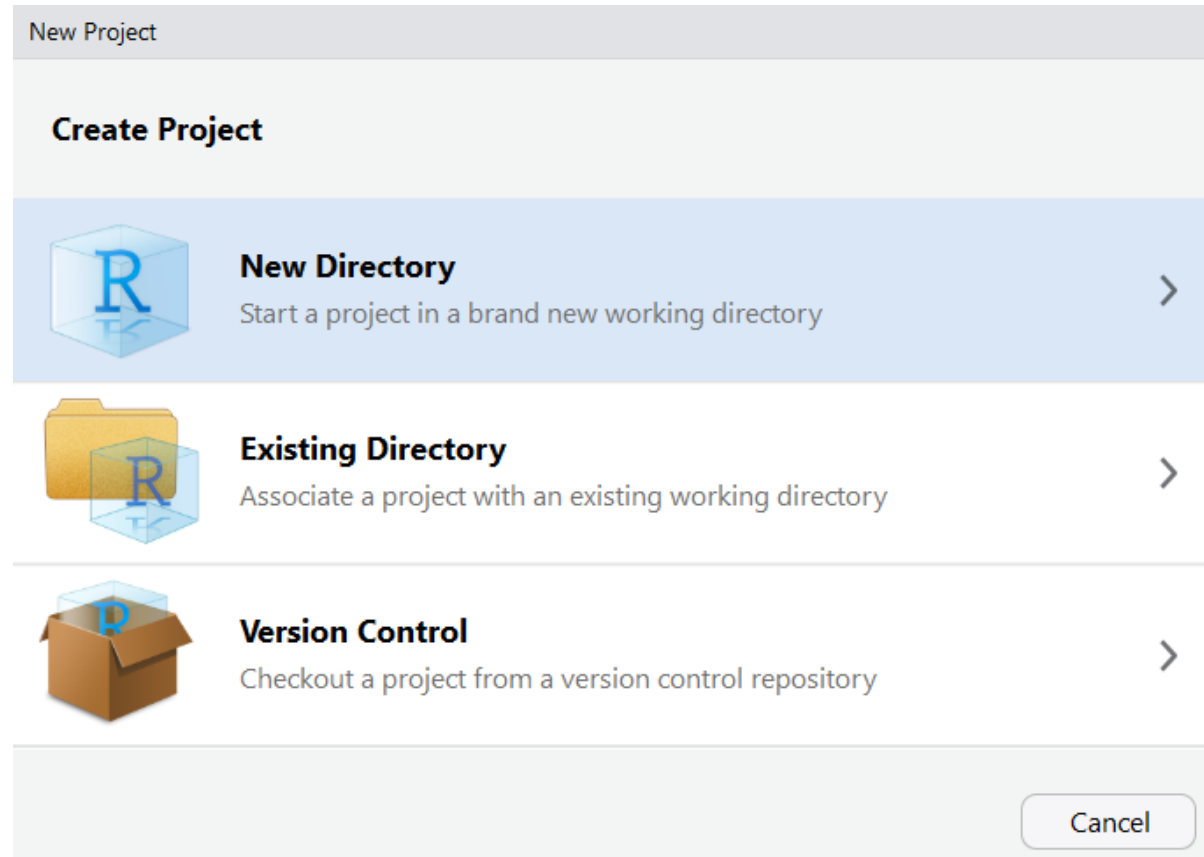
2) **Uso de proyectos**



3) Programación «literaria»

4) Control de Versiones

Uso de proyectos



<https://support.rstudio.com/hc/en-us/articles/200526207-Using-Projects>

Trabajar con proyectos

- Hace que sea más sencillo compartir código con otra persona
- Permiten cargar fácilmente código con el envío de un manuscrito
- Hacen que sea más fácil retomar el proyecto después de un descanso.

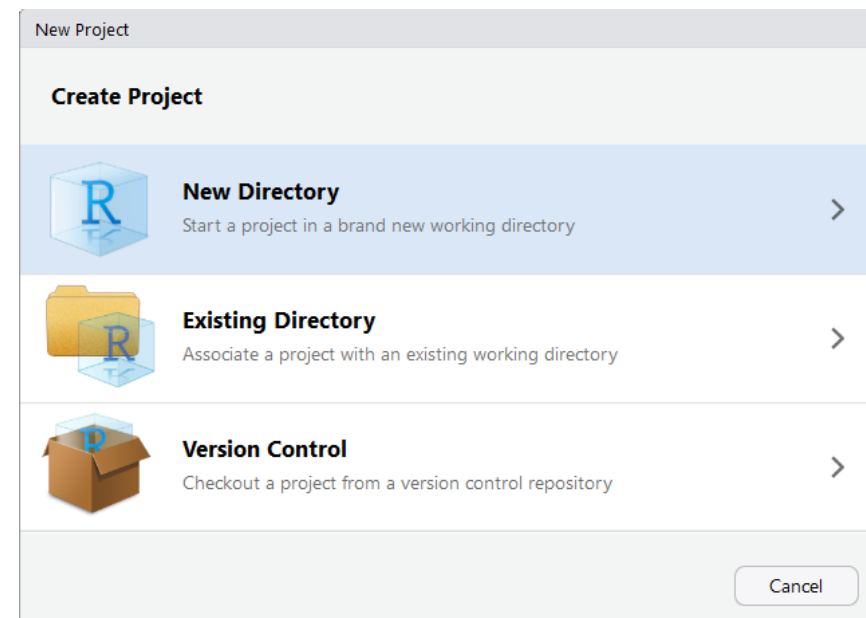


Vince Buffalo
@vsbuffalo

Managing your projects in a reproducible fashion doesn't just make your science reproducible, it makes your life easier.

12:26 AM - Apr 15, 2013

♥ 38 💬 35 people are talking about this






¿Alguna vez actualizaste un paquete
y algún código en otro proyecto dejó de funcionar?

New Project

[Back](#) **Create New Project**



Directory name:

Create project as subdirectory of:
 [Browse...](#)

☐ Create a git repository

☒ Use packrat with this project

☐ Open in new session

[Create Project](#) [Cancel](#)



sistema de
administración de
dependencias para R

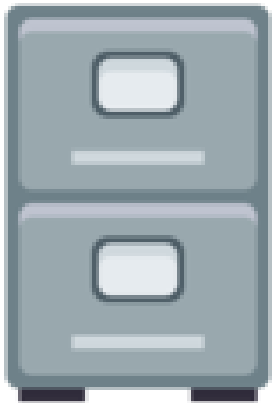
Registra las versiones de
los paquetes

Asegura que esas versiones
exactas sean las que se
instalen donde quiera que
abra el proyecto

<http://rstudio.github.io/packrat/>

Uso de archivos de texto plano

.CSV
.txt



Los archivos de **texto plano** son mejores para guardar los datos de nuestras investigaciones.

Otros formatos de archivo cambian regularmente y puede que no sean compatibles con futuras versiones de estos programas.

1) Estilo de código

2) Uso de proyectos

3) Programación «literaria»



4) Control de Versiones

RMarkdown



#A veces comentar un archivo no es suficiente



Programación literaria

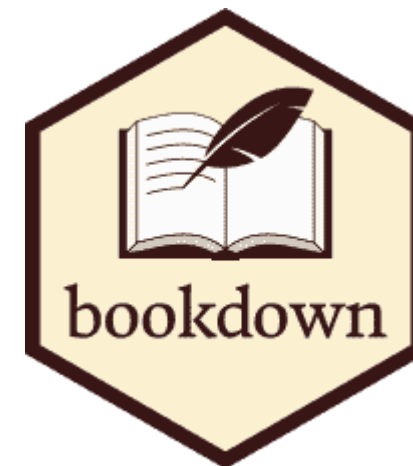
RMarkdown

«El paradigma de **programación literaria**, tal como fue concebido por Knuth, representa un alejamiento de la escritura de programas en la forma y el orden impuestos por la computadora, y en su lugar permite a los programadores desarrollar programas en el orden exigido por la lógica y el flujo de sus pensamientos»



Escribir tu paper en RMarkdown

Paquete **rticle**: Plantillas de artículos de revistas para R Markdown



Algunas revistas con plantilla

[Elsevier](#) journal submissions

[PeerJ](#) articles

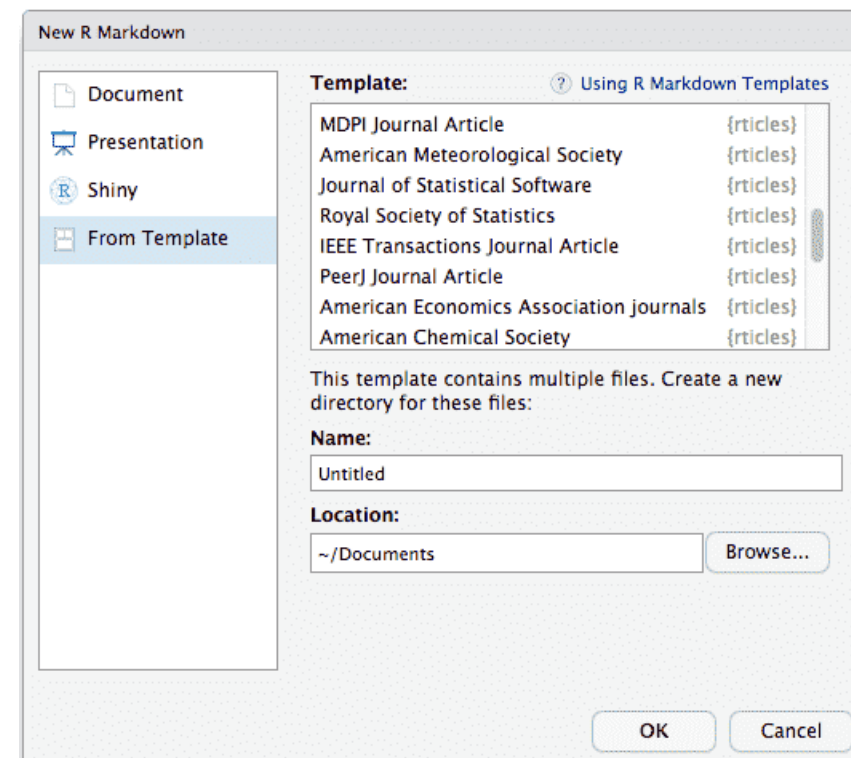
[Royal Society Open Science](#) journal submissions

[Sage](#) journal submissions

[Springer](#) journal submissions

[The R Journal](#) articles

[Taylor & Francis](#) articles



<https://github.com/rstudio/rticles>

<https://bookdown.org/baydap/bookdownplus/academic.html#articles>

<https://bookdown.org/yihui/rmarkdown/journals.html>

1) Estilo de código

2) Uso de proyectos

3) Programación «literaria»

4) Control de Versiones





git

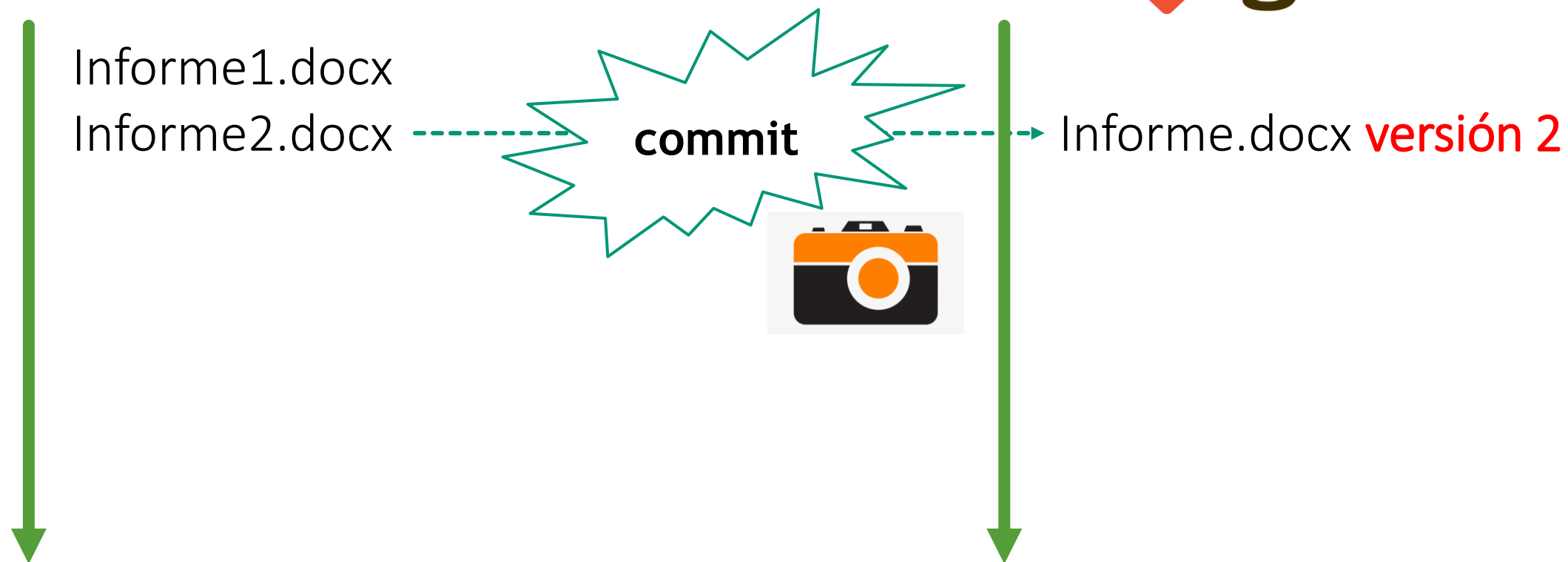


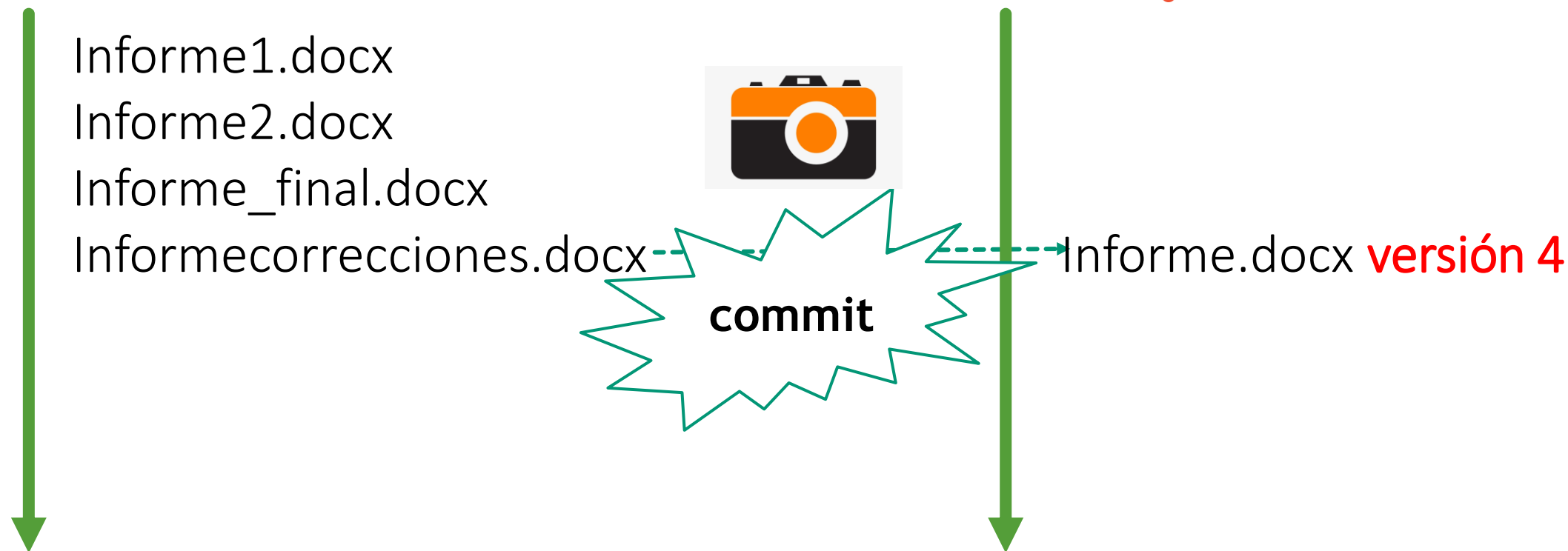
Git es un sistema de control de versiones

Su propósito original fue ayudar a grupos de desarrolladores a trabajar en colaboración en grandes proyectos de software.

ES MUY POCO INTUITIVO





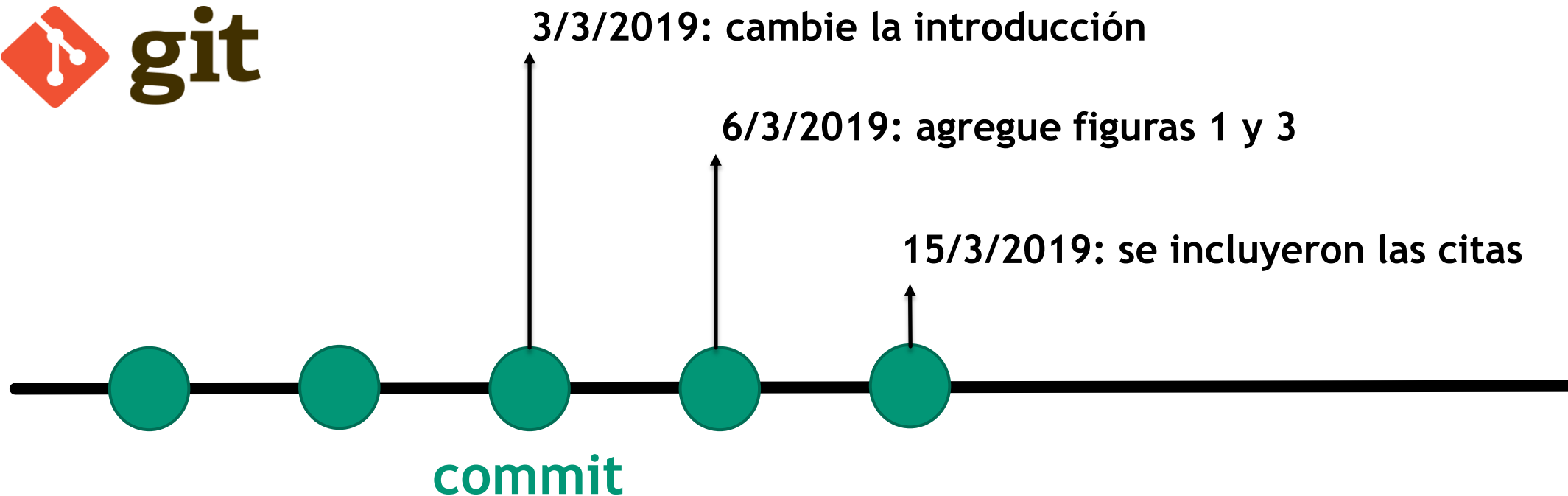




↓
Informe1.docx
Informe2.docx
Informe_final.docx
Informecorrecciones.docx
Informe_final2.docx
Informemasfinal.docx
Informe_ultimodeverdad.docx
Informe_listoparaentregar.docx

↓ Informe.docx **versión 8**

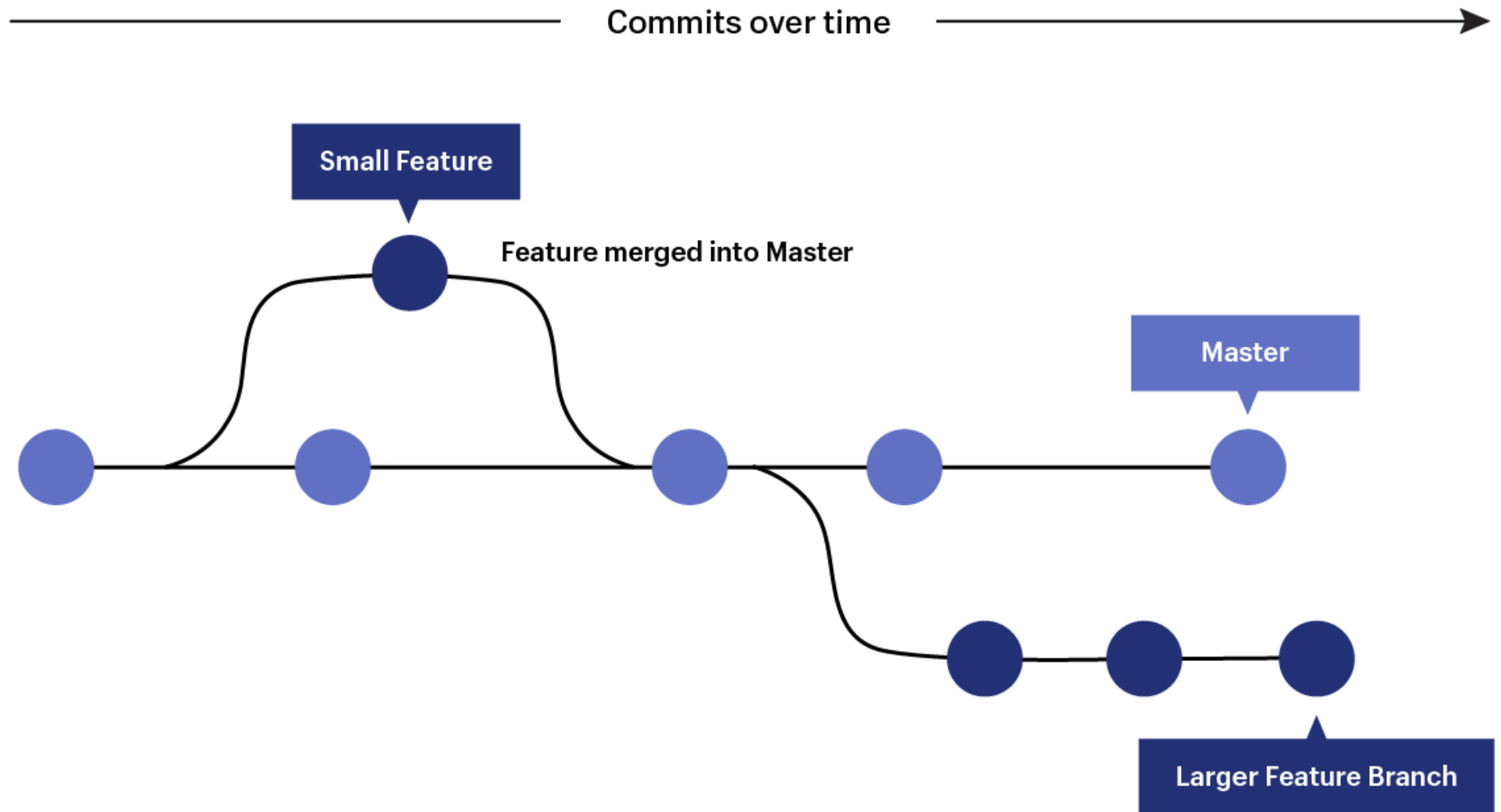
Se guarda como una **versión** diferente
no como un archivo diferente



Informe.docx

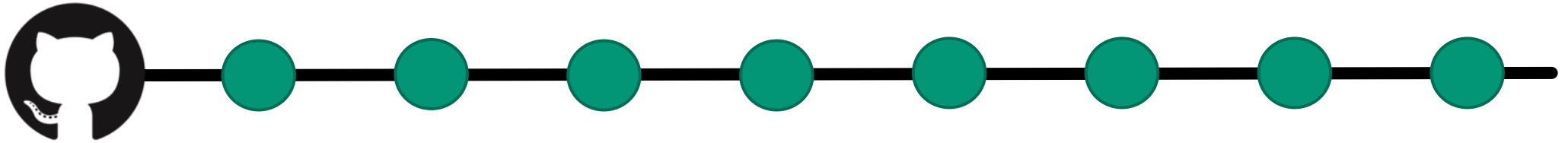
Permite

- El seguimiento de cambios en los documentos
- Revertir cambios



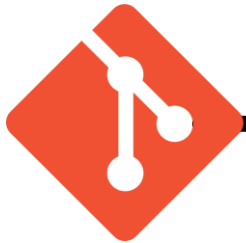
- Branches o ramas: Experimentar con diferentes versiones de un documento manteniendo la versión original
- Combinar dos versiones de un documento

Remoto
github.com

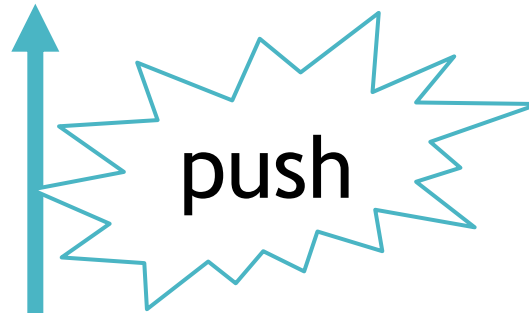


Local

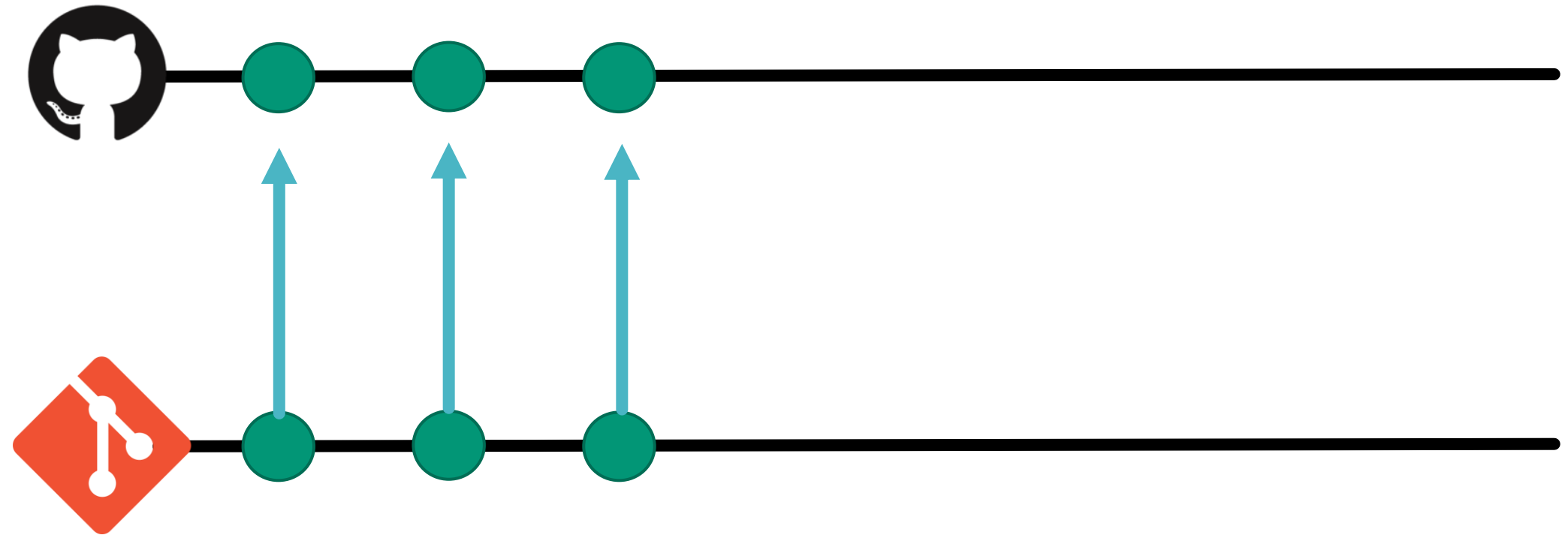
Remoto
github.com



Local

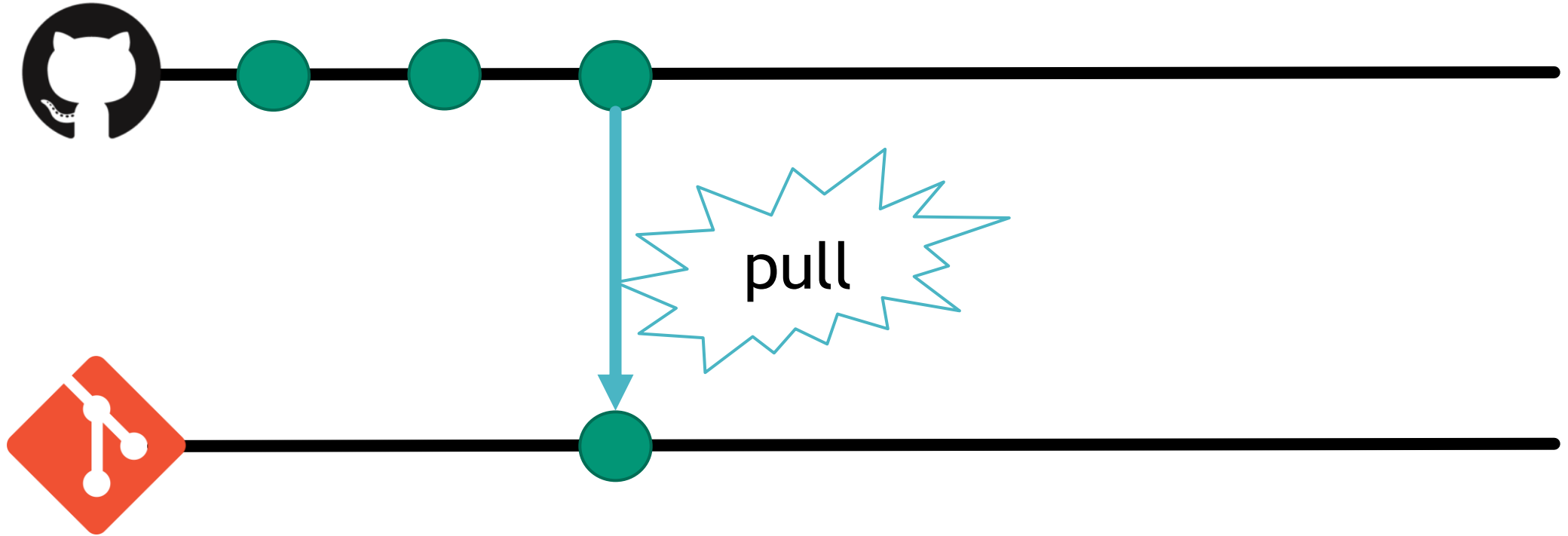


Remoto
github.com



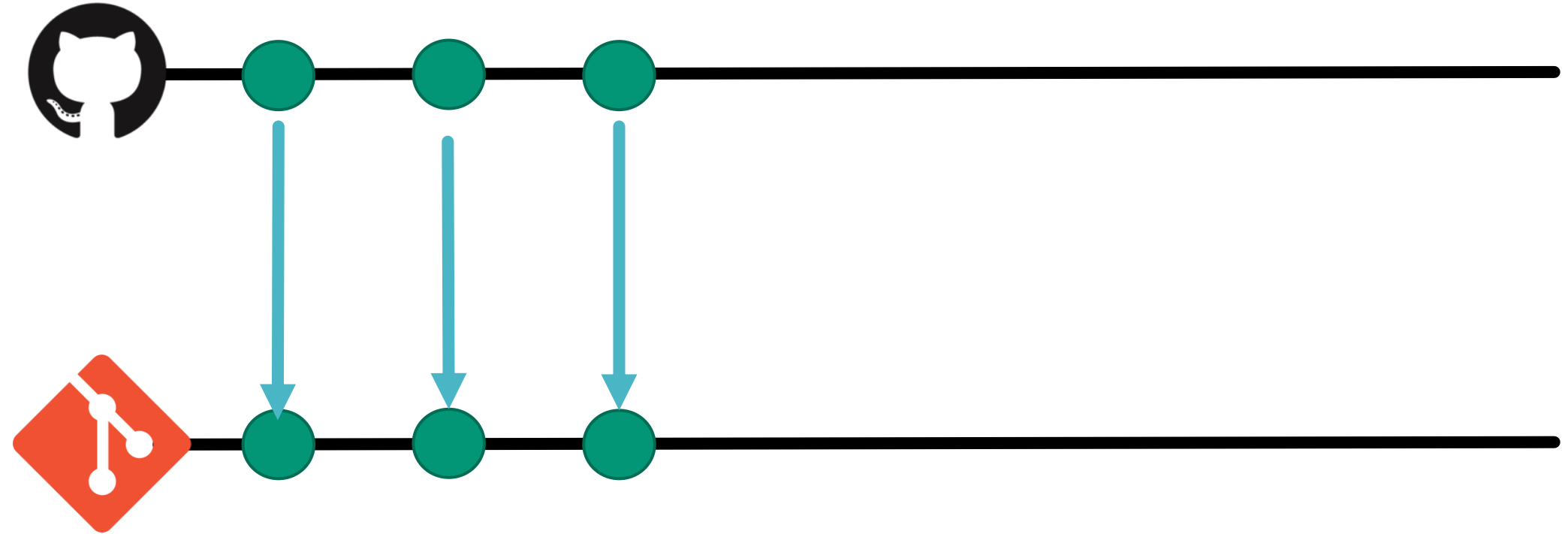
Local

Remoto
github.com

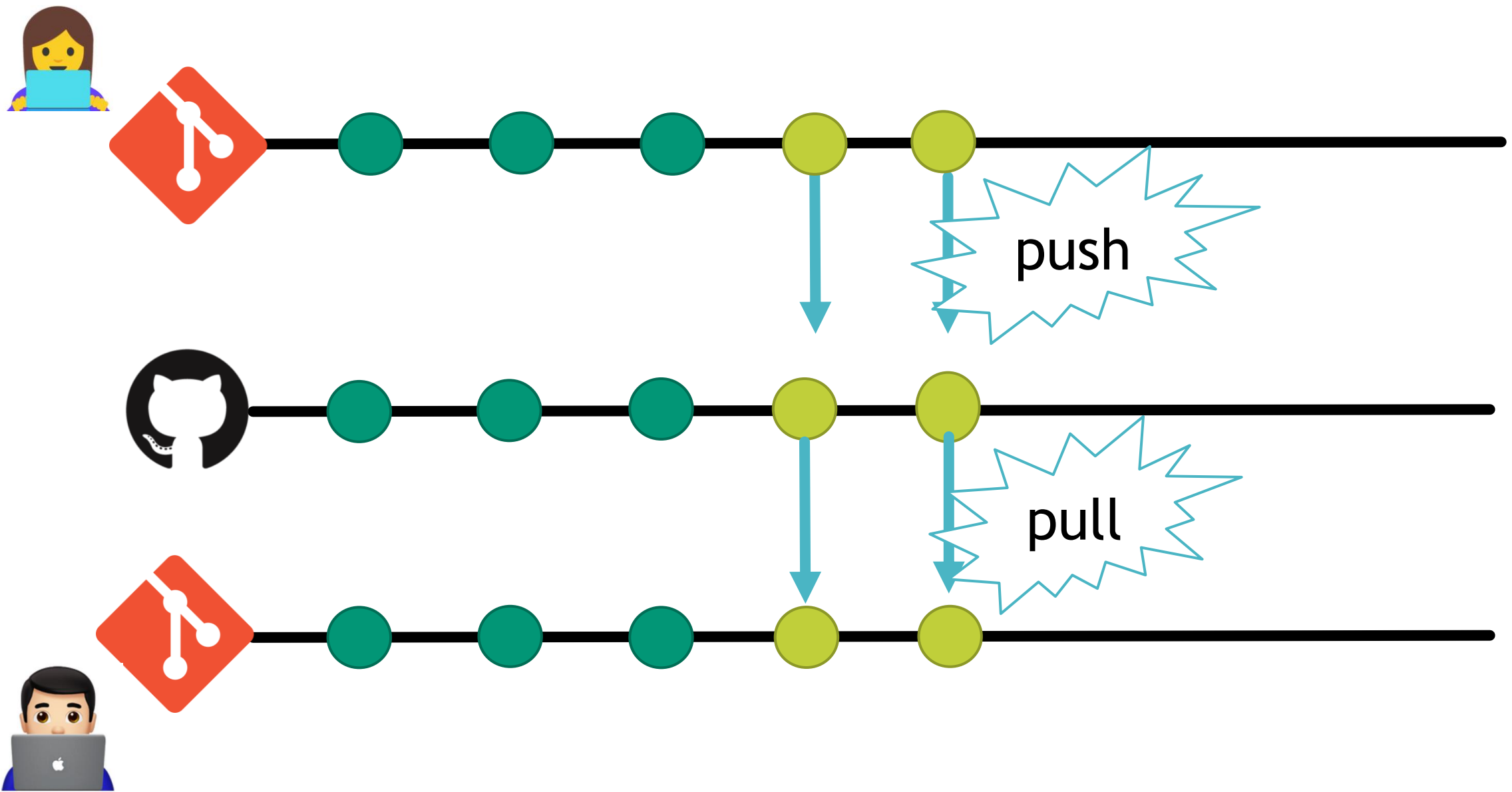


Local

Remoto
github.com



Local





Git se puede integrar a los proyectos directamente desde R-Studio

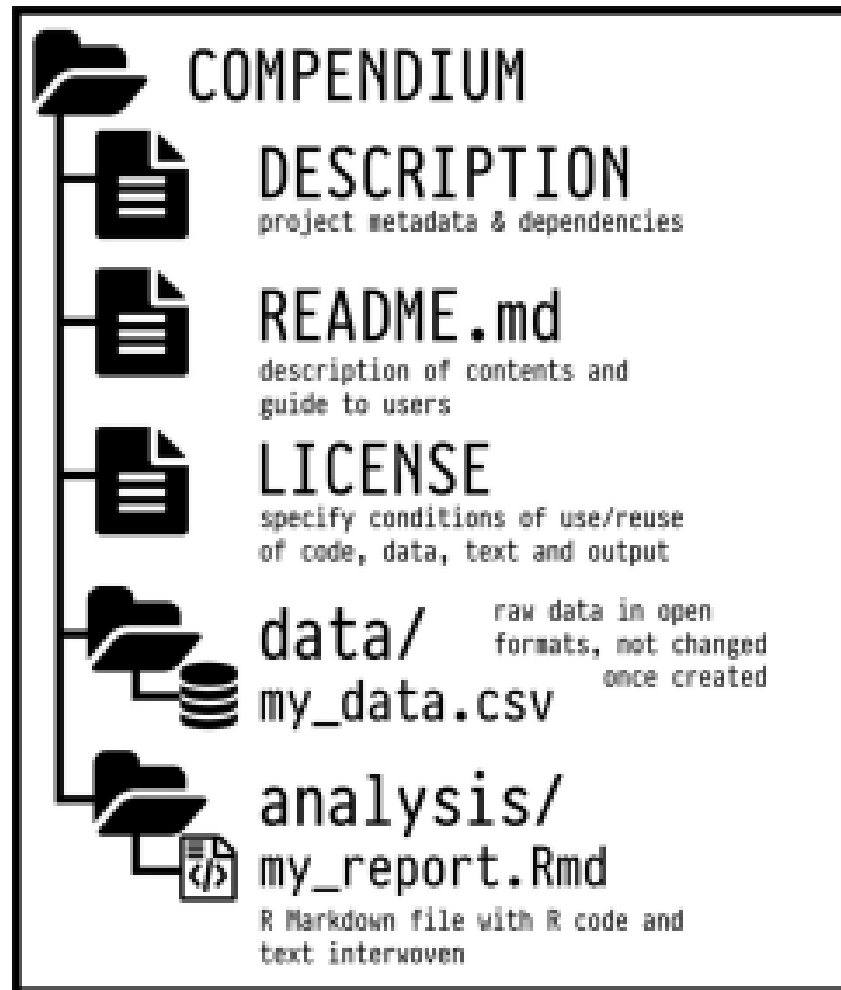
Happy Git and GitHub for the useR
Jenny Bryan

<https://happygitwithr.com/>

**¿Cómo compartir mi código y
datos al momento de publicar?**

Compendio

forma de **organizar los materiales digitales** de un proyecto para permitir que otros reproduzcan y extiendan la investigación



Se obtiene un DOI y se
Se puede citar.



- Organización de archivos de acuerdo estándares de la comunidad
- **Mantiene datos y el método separados**
 - *Datos como "solo lectura»*
 - *El «paso a paso» de la metodología en el código*
- Especifica el entorno computacional que se utilizó para el análisis

El paquete *rrtools*

<https://github.com/benmarwick/rrtools>

Te ayuda a generar un
compendio con tus
datos y códigos

```
analysis/  
├── paper/  
│   ├── paper.Rmd      # this is the main document to edit  
│   └── references.bib  # this contains the reference list information  
  
├── figures/           # location of the figures produced by the Rmd  
  
├── data/  
│   ├── raw_data/      # data obtained from elsewhere  
│   └── derived_data/  # data generated during the analysis  
  
└── templates  
    ├── journal-of-archaeological-science.csl  
    │   # this sets the style of citations & reference list  
    ├── template.docx  # used to style the output of the paper.Rmd  
    └── template.Rmd
```

<https://ropensci.org/commcalls/2019-07-30/>

Una publicación reproducible

- **Aumenta el impacto de nuestra investigación**
datos libres = mayor número de citas (Pienta et al. 2010)
- **Minimiza la probabilidad de cometer errores**
si bien la reproducibilidad no es garantía de que el trabajo es realizado forma correcta, permite detectar errores con mayor facilidad.
- **Nos permite comunicar nuestros resultados de forma clara y sencilla**
tanto nuestros colaboradores directos como a nosotros mismos

ReproHack

<https://github.com/reprohack>



Hackaton Experiencia práctica en reproducibilidad con materiales reales publicados.

- Aumenta el potencial de reutilización y comprensión de las publicaciones.
- Permite identificar dónde están las debilidades más apremiantes en nuestros enfoques.



Dra. Anna Krystalli
University of Sheffield

<https://twitter.com/annakrystalli>

RaukR 2019 • Course content

Course materials are listed below. (The GitHub repo for RaukR 2019 can be accessed by teachers [here](#).)

Link to Jennys Bryan, RStudio, lectures

<http://rstd.io/raukr>

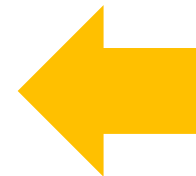
Week 1

10-Jun-2019 (Mon)

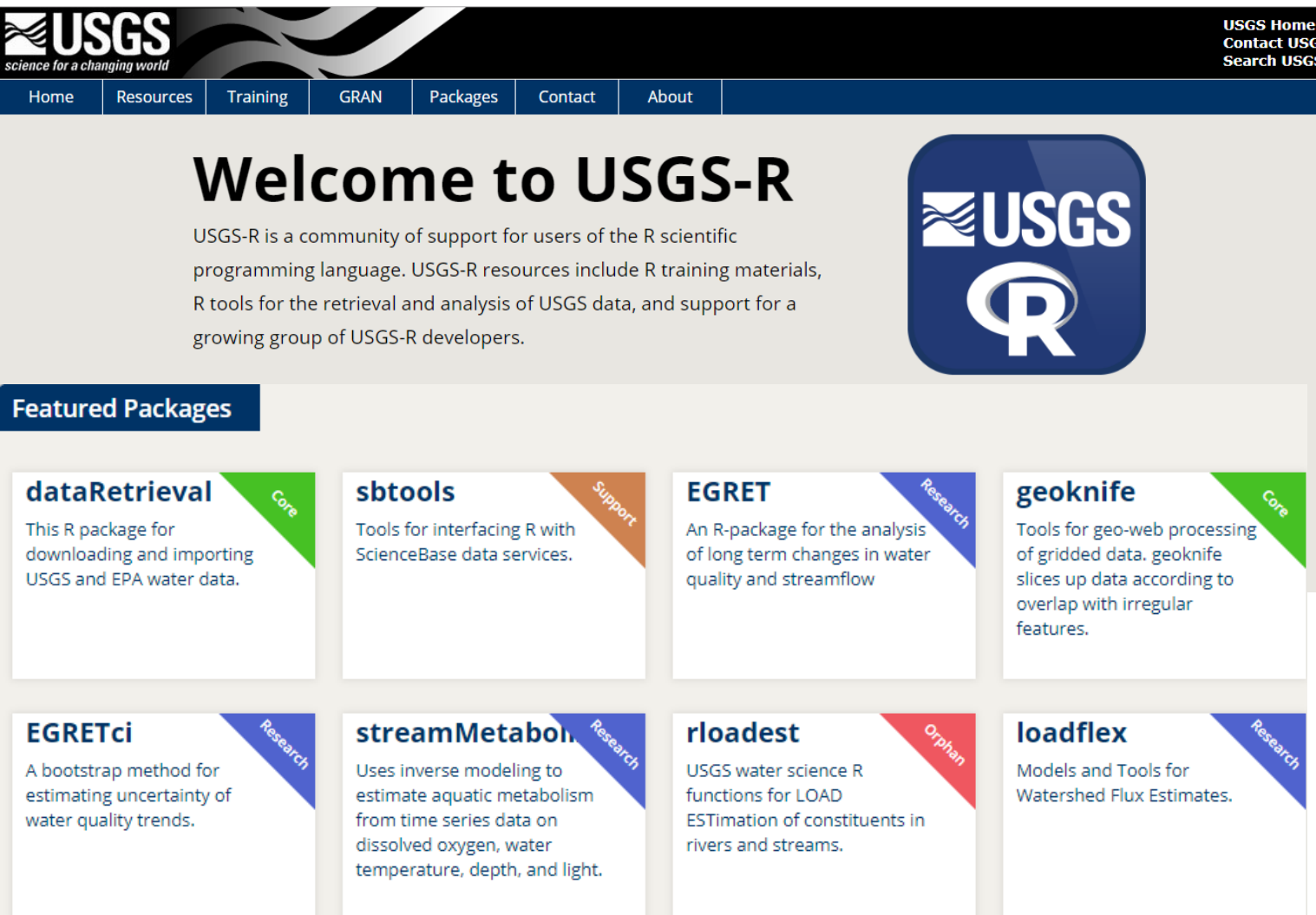
- **Introduction to RaukR 2019** (*Author: Marcin Kierczak*)
 - [Presentation](#)
- **Reproducible research** (*Author: Roy Francis, Reviewer: Marcin Kierczak*)
 - [Presentation](#)
 - [Lab](#)
- **Best Coding Practises** (*Author: Marcin Kierczak, Reviewer: Roy Francis*)
 - [Presentation](#)
 - [Lab](#)

RaukR. Summer School, 2019.

<https://nbisweden.github.io/RaukR-2019/>



Uso de R en Instituciones



The screenshot shows the USGS-R website homepage. At the top is the USGS logo with the tagline "science for a changing world". To the right of the logo are links for "USGS Home", "Contact USGS", and "Search USGS". Below this is a navigation bar with links for "Home", "Resources", "Training", "GRAN", "Packages", "Contact", and "About". The main heading is "Welcome to USGS-R". Below this heading is a paragraph: "USGS-R is a community of support for users of the R scientific programming language. USGS-R resources include R training materials, R tools for the retrieval and analysis of USGS data, and support for a growing group of USGS-R developers." To the right of this text is a large USGS-R logo. Below the main heading is a section titled "Featured Packages" which contains eight cards, each representing a different R package. Each card has a title, a description, and a category label in a colored triangle.

USGS
science for a changing world

USGS Home
Contact USGS
Search USGS

Home Resources Training GRAN Packages Contact About

Welcome to USGS-R

USGS-R is a community of support for users of the R scientific programming language. USGS-R resources include R training materials, R tools for the retrieval and analysis of USGS data, and support for a growing group of USGS-R developers.

Featured Packages

Package Name	Description	Category
dataRetrieval	This R package for downloading and importing USGS and EPA water data.	Core
sbtools	Tools for interfacing R with ScienceBase data services.	Support
EGRET	An R-package for the analysis of long term changes in water quality and streamflow	Research
geoknife	Tools for geo-web processing of gridded data. geoknife slices up data according to overlap with irregular features.	Core
EGRETci	A bootstrap method for estimating uncertainty of water quality trends.	Research
streamMetabolism	Uses inverse modeling to estimate aquatic metabolism from time series data on dissolved oxygen, water temperature, depth, and light.	Research
rlodeast	USGS water science R functions for LOAD ESTimation of constituents in rivers and streams.	Orphan
loadflex	Models and Tools for Watershed Flux Estimates.	Research



<https://owi.usgs.gov/R/>



How to create BBC style graphics

- Load all the libraries you need
- Install the bbplot package
- How does the bbplot package work?
- Save out your finished chart
- Make a line chart
- Make a multiple line chart
- Make a bar chart
- Make a stacked bar chart
- Make a grouped bar chart
- Make a dumbbell chart
- Make a histogram
- Make changes to the legend
- Make changes to the axes
- Add annotations
- Work with small multiples
- Do something else entirely

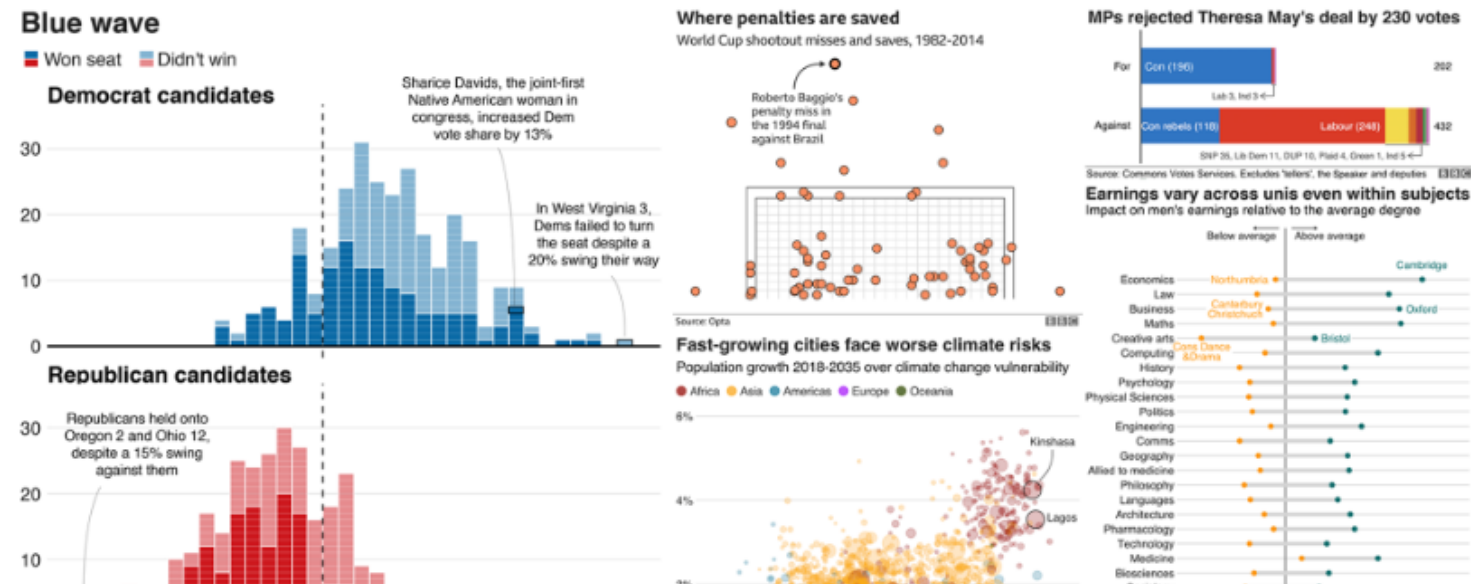
BBC Visual and Data Journalism cookbook for R graphics

Last updated: 2019-01-24

How to create BBC style graphics

At the BBC data team, we have developed an R package and an R cookbook to make the process of creating publication-ready graphics in our in-house style using R's ggplot2 library a more reproducible process, as well as making it easier for people new to R to create graphics.

The cookbook below should hopefully help anyone who wants to make graphics like these:



https://bbc.github.io/rcookbook/#how_to_create_bbc_style_graphics

Comunidades

Links en página web https://flor14.github.io/r_inta/

R-Ladies

¡Sumate a nuestra comunidad!



/RLadiesBA



R Ladies
Buenos Aires



@RLadiesBA



/RLadies-
Buenos-Aires



/RLadies-BA



129

R-Ladies groups on meetup.com

40

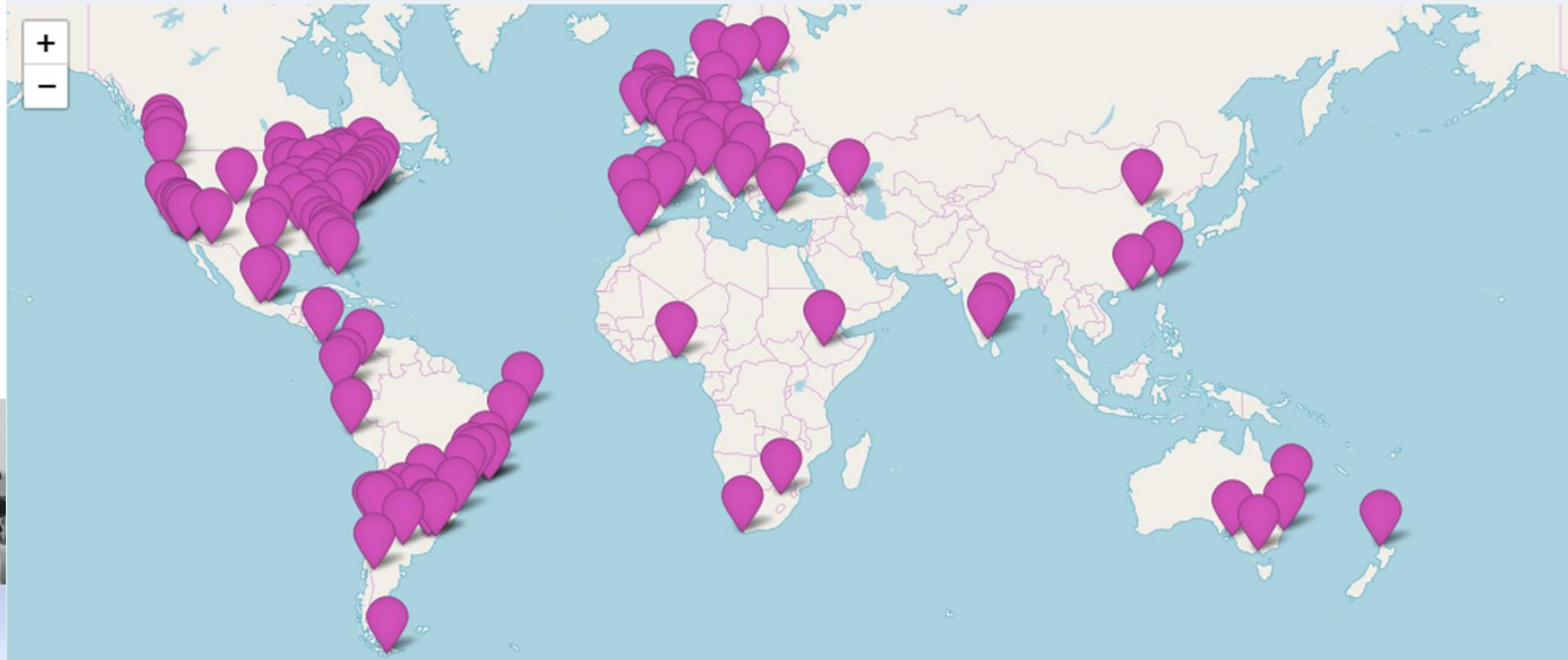
R-Ladies Countries

129

R-Ladies Cities

30904

R-Ladies members on meetup.com



Ggplot2
Ciencia de Datos
Taller de
Casos



<https://gqueiroz.shinyapps.io/rshinylady/>

R en Buenos Aires

<https://renbares.github.io/>



R-Spatial ES

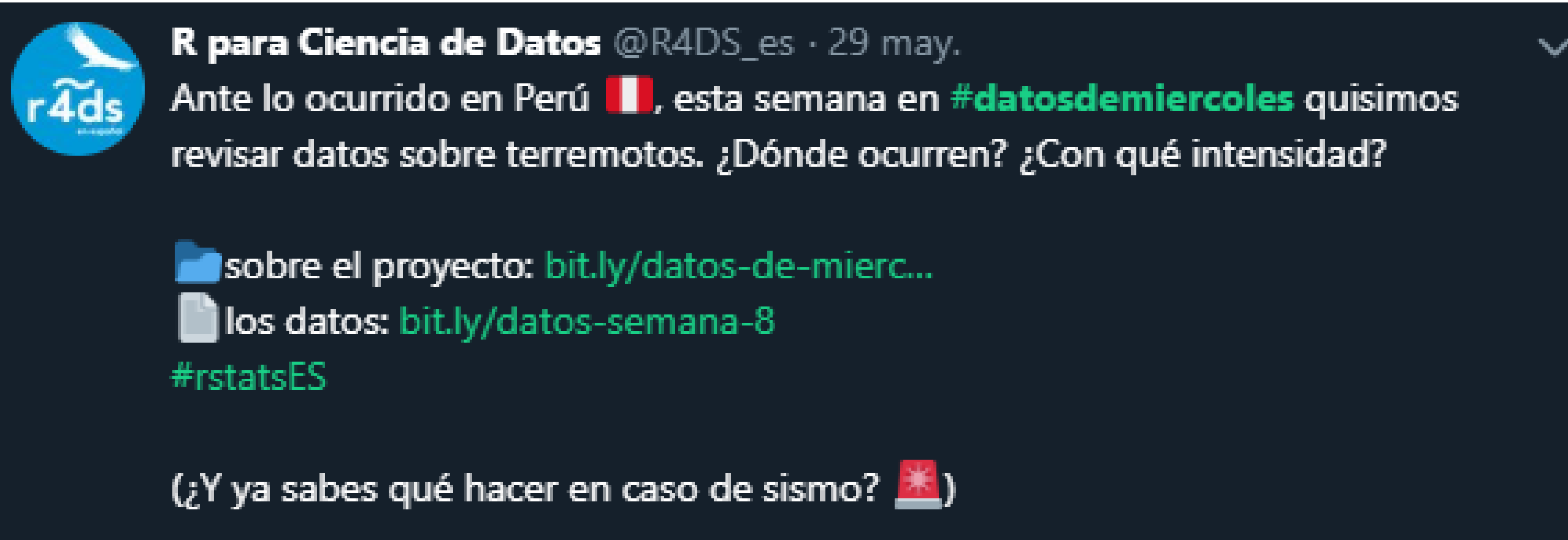
- Datos espaciales con R

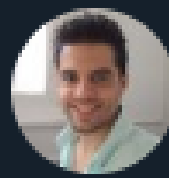
R4DS_ES

- Traducción del libro R4DS al español
- #datosdemiercoles



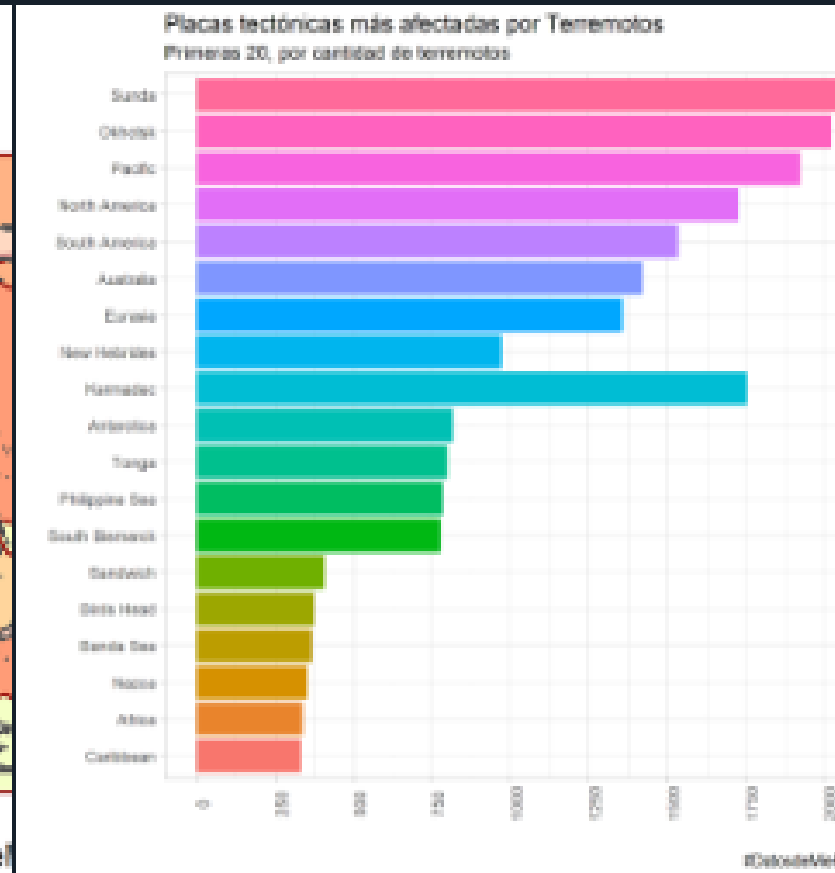
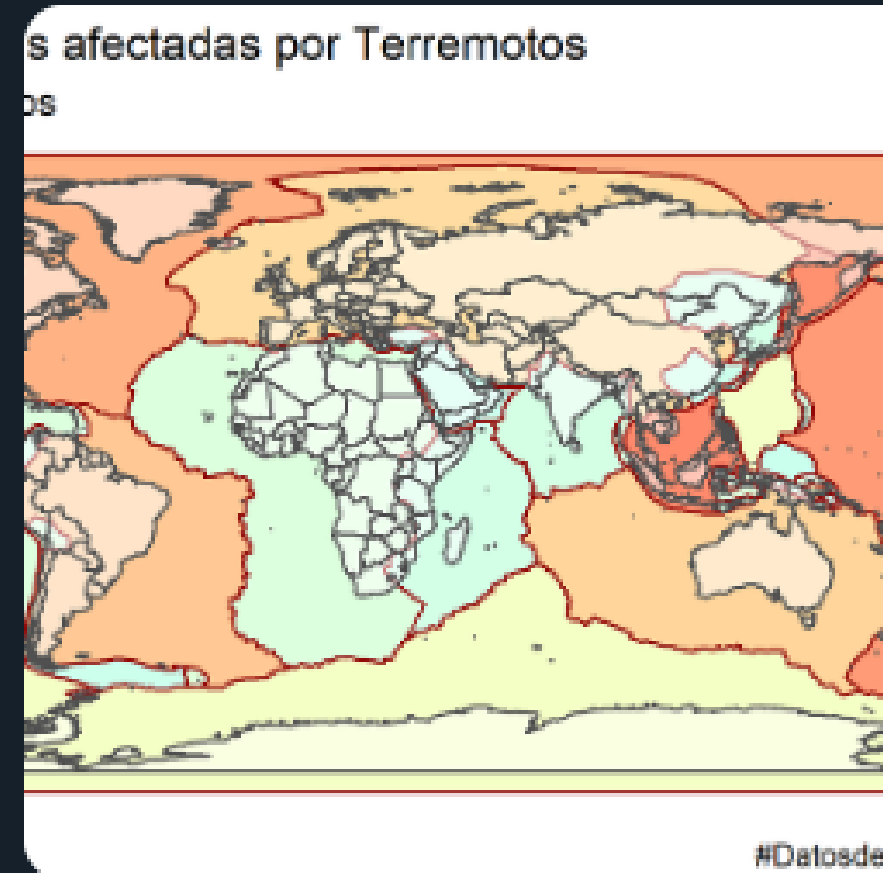
#datosdemiercoles





Julio Spairani @jspairani · 2 jun.

2019-05-28 #DatosdeMieRcoles #rstats_ES Desafío Terremotos! inspirado en el post de placas de @violetrzn , Quise ver que placas eran las que más terremotos tienen asociados. Va barchart complementario para ver los nombres. 😊
para contexto: youtube.com/watch?v=T2WqVj...



2

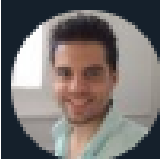


1



15





Julio Spairani @jspairani · 2 jun.

algunos recursos: 📖

- para datos de placas use estos que ya traian las placas como polígonos y no como líneas: github.com/fraxen/tectoni...
- para ver puntos en poligonos: spatialEco, referido de aca: stackoverflow.com/questions/3647...
- paleta de colores inspirada en:



Summer Sunset at the lake Color Palette

color-hex.com



violeta ❤️ @violetrzn · 2 jun.

En respuesta a [@jspairani](#)

copado! me gustó



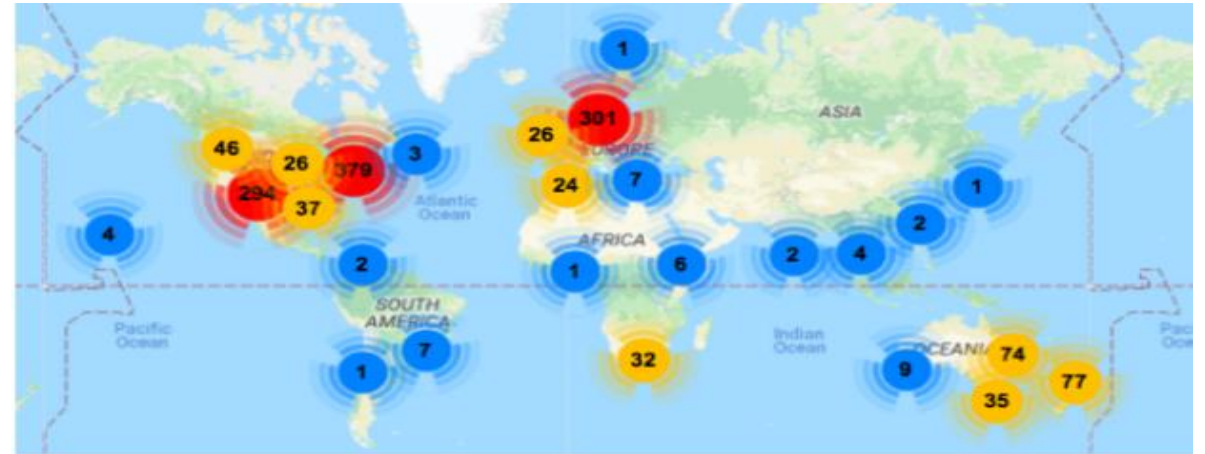


 **THE
CARPENTRIES**

We teach foundational coding and data science skills to researchers worldwide.

Somos una **comunidad global** que enseña habilidades básicas de computación y ciencia de datos a investigadores en

- el mundo académico,
- la industria
- y el gobierno.



<https://twitter.com/thecarpentries>

<https://carpentries.org>

Organización sin fines de lucro

- > Hay gran **demanda de entrenamiento** en habilidades básicas de programación
- > Los libros de texto de ingeniería de software no son apropiados para **enseñar a programar** a la mayoría de los científicos.

Software Carpentry



Los materiales de aprendizaje se encuentran abiertos y disponibles

<http://swcarpentry.github.io/r-novice-gapminder/>



software carpentry

Teaching basic lab skills
for research computing

R para Análisis Científicos Reproducibles

El objetivo de esta lección es enseñar a las programadoras principiantes a escribir códigos modulares y adoptar buenas prácticas en el uso de R para el análisis de datos. R nos provee un conjunto de paquetes desarrollados por terceros que se usan comúnmente en diversas disciplinas científicas para el análisis estadístico. Encontramos que muchos científicos que asisten a los talleres de Software Carpentry utilizan R y quieren aprender más. Nuestros materiales son relevantes ya que proporcionan a los asistentes una base sólida en los fundamentos de R y enseñan las mejores prácticas del cómputo científico: desglose del análisis en módulos, automatización tareas y encapsulamiento.

Ten en cuenta que este taller se enfoca en los fundamentos del lenguaje de programación R y no en el análisis estadístico.

A lo largo de este taller se utilizan una variedad de paquetes desarrollados por terceros, los cuales no son necesariamente los mejores ni se encuentran explicadas todas sus funcionalidades, pero son paquetes que consideramos útiles y han sido elegidos principalmente por su facilidad de uso.

Data Carpentry

<https://datacarpentry.org/lessons/>



Curriculum materials

- [Ecology curriculum](#)
- [Genomics curriculum](#)
- [Social Sciences curriculum](#)
- [Geospatial data curriculum](#)

The background features abstract, overlapping green geometric shapes, primarily triangles and polygons, in various shades of green, creating a modern and dynamic visual effect. The shapes are layered, with some appearing more prominent than others, and they extend from the right side of the frame towards the center.

Eventos

Congreso Argentino de Agroinformática



Universidad Nacional de Salta / SALTA / Argentina / 16 al 20 de Septiembre de 2019 / www.48jaiio.sadio.org.ar/cai



Temas:

- Software y sistemas de información agropecuarios.
- Modelización de sistemas de producción.
- Integración y trazabilidad de cadenas agro-industriales.
- Geomática, Sistemas de Información Geográficos, IDEs, Teledetección y Observación Terrestre.
- Robótica agro-industrial.
- Agricultura y ganadería de precisión.
- Redes de sensores en cultivos, tambos, feed-lots y plantas de procesamiento.
- Sistemas embebidos y desarrollos electrónicos en la agro-industria.
- Monitoreo y control medio ambiental.
- Ontologías, Big Data, Open Data y DataMining e inteligencia artificial aplicadas al agro.
- Bioinformática y registros biológicos.
- Servicios Web Agroindustriales y Web 2.0.
- Nuevos desarrollos y experimentos en AgroTICs.
- Aplicaciones móviles.
- Internet de las cosas aplicadas al agro.
- Experiencias educativas en TICs aplicadas al agro.

Llamado a presentación de trabajos #CAI2019
Cierre de recepción de trabajos:
26 de abril de 2019

Más detalles:
<http://48jaiio.sadio.org.ar/simposios/cai>



<http://latin-r.com/>

2da Conferencia Latinoamericana sobre Uso de R en Investigación + Desarrollo



2018 – Buenos Aires / 2019 – Santiago de Chile



Conferencia Latinoamericana sobre
el Uso de R en Investigación + Desarrollo

25 - 27 DE SEPTIEMBRE | 2019
SANTIAGO DE CHILE

Keynote Speakers



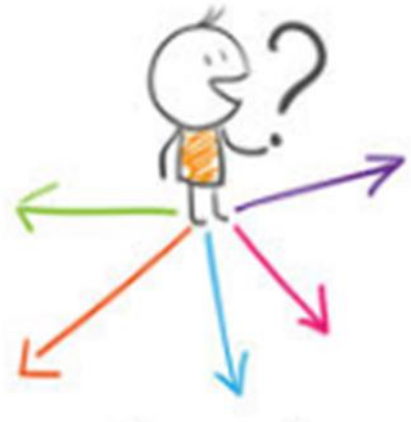
Mine Çetinkaya-Rundel



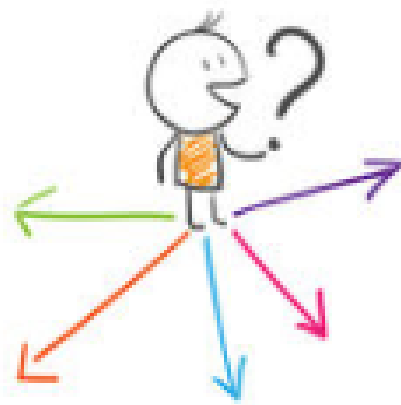
Hadley Wickham



Entonces



- ▶ *¿Cuán necesario es adoptar R en nuestro flujo de trabajo?*
- ▶ *¿Existen usos de R no relacionados con la estadística?*
- ▶ *¿Necesito formación especial para comenzar a emplearlo?*



¿Preguntas?

¡Muchas gracias!